Regularizing Score-based Models with Score Fokker-Planck Equations

Chieh-Hsin Lai
Sony Group Corporation
Tokyo, Japan 108-007
`Chieh-hsin.lai@sony.com`

Yuhta Takida
Sony Group Corporation
Tokyo, Japan 108-007
`yuta.takida@sony.com`

Naoki Murata
Sony Group Corporation
Tokyo, Japan 108-007
`naoki.murata@sony.com`

Toshimitsu Uesaka
Sony Group Corporation
Tokyo, Japan 108-007
`toshimitsu.uesaka@sony.com`

Yuki Mitsufuji
Sony Group Corporation
Tokyo, Japan 108-007
`yuhki.mitsufuji@sony.com`

Stefano Ermon
Stanford University
Stanford, CA 94305
`ermon@cs.stanford.edu`

Abstract

Score-based generative models learn a family of noise-conditional score functions corresponding to the data density perturbed with increasingly large amounts of noise. These perturbed data densities are tied together by the *Fokker-Planck equation* (FPE), a PDE governing the spatial-temporal evolution of a density undergoing a diffusion process. In this work, we derive a corresponding equation characterizing the noise-conditional scores of the perturbed data densities (i.e., their gradients), termed the *score FPE*. Surprisingly, despite impressive empirical performance, we observe that scores learned via denoising score matching (DSM) do not satisfy the underlying score FPE. We mathematically analyze two implications of satisfying the score FPE and a potential explanation for why the score FPE is not satisfied in practice. At last, we propose to regularize the DSM objective to enforce satisfaction of the score FPE, and show its effectiveness on synthetic data and MNIST.

1 Score-based generative models

[16] unifies denoising score matching [14] and diffusion probabilistic models [13, 4] via a stochastic process $\boldsymbol{x}(t)$ of continuous time $t \in [0, T]$ driven by the forward SDE

$$d\boldsymbol{x}(t) = \boldsymbol{f}(\boldsymbol{x}(t), t)dt + g(t)d\boldsymbol{w}_t, \tag{1}$$

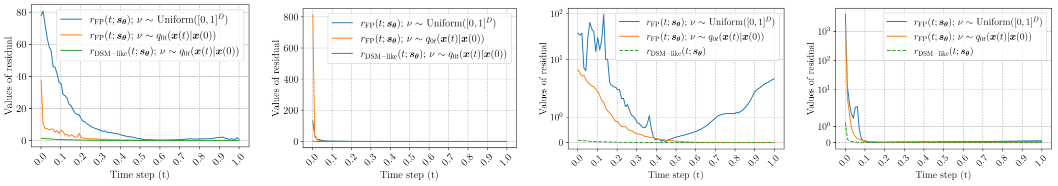
where $\boldsymbol{f}(\cdot, t) \colon \mathbb{R}^D \to \mathbb{R}^D$, $g(\cdot) \colon \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{w}_t$ is a standard Wiener process. Under some moderate conditions [1], one can obtain a reverse time SDE from T to 0

$$d\boldsymbol{x}(t) = [\boldsymbol{f}(\boldsymbol{x}(t), t) - g^2(t)\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}(t))]dt + g(t)d\bar{\boldsymbol{w}}_t, \tag{2}$$

where $\bar{\boldsymbol{w}}_t$ is a standard Wiener process in reverse time. Let $q_t(\boldsymbol{x})$ denote the ground truth marginal density of $\boldsymbol{x}(t)$ following Eq. 1. We can train a neural network $\boldsymbol{s}_{\boldsymbol{\theta}} = \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ to approximate $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ by minimizing the denoising score matching (DSM) loss [18, 16]:

$$\mathcal{J}_{\text{DSM}}(\boldsymbol{\theta}; \lambda(\cdot)) := \frac{1}{2}\int_0^T \lambda(t)\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}(t), t) - \nabla_{\boldsymbol{x}(t)} \log q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))\right\|_2^2\right]dt, \tag{3}$$

NeurIPS 2022 Workshop on Score-Based Methods.



(a) VE SDE; MNIST (b) VP SDE; MNIST (c) VE SDE; CIFAR-10 (d) VP SDE; CIFAR-10

Figure 1: Comparison of the numerical scale of $r_{\text{DSM}}(t; \boldsymbol{s_\theta})$ and $r_{\text{FP}}(t; \boldsymbol{s_\theta})$ of pre-trained scores $\boldsymbol{s_\theta}$ on MNIST and CIFAR-10. Pre-trained models do not numerically satisfy score FPE in contrast to their denoising score matching-like errors. We attempt to explain this phenomena in Sec. 3.1 and 3.3.

where $q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$ is the transition kernel from $\boldsymbol{x}(0)$ to $\boldsymbol{x}(t)$. After $\boldsymbol{s_\theta}(\boldsymbol{x}, t) \approx \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ is learned, we replace $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ in Eq. 2 with $\boldsymbol{s_\theta}$ and get a parametrized reverse-time SDE for stochastic process $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(t)$

$$d\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(t) = [\boldsymbol{f}(\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t) - g^2(t)\boldsymbol{s_\theta}(\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t)]dt + g(t)\bar{\boldsymbol{w}}_t, \tag{4}$$

Let $p_{t,\boldsymbol{\theta}}^{\text{SDE}}$ denote the marginal distribution of $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(t)$. We can design $\boldsymbol{f}$ and g in Eq. 2 so that $q_T(\boldsymbol{x})$ approximates a simple prior π, and hence, can generate samples $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(0) \sim p_{0,\boldsymbol{\theta}}^{\text{SDE}}$ by numerically solving Eq. 4 backward with an initial sample from the prior $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(T) \sim \pi$. Intuitively, $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(0)$ should be close to a sample from the data distribution.

2 The Fokker-Planck equation for a score vector field

It is well known that the evolution of the ground truth density $q_t(\boldsymbol{x})$ associated to Eq. 1 is governed by the Fokker-Planck equation (FPE) [10] (details in Appx. E). As there is a one-to-one mapping between densities and their scores, we can derive an equivalent system of PDEs that the ground truth scores $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ must satisfy. We call it as a *score Fokker-Planck equation*, for short *score FPE*.

Corollary 1 (score FPE)**.** *Assume that the ground truth density $q_t(\boldsymbol{x})$ is sufficiently smooth for $(\boldsymbol{x}, t) \in \mathbb{R}^D \times [0, T]$. Then its score $\boldsymbol{s}(\boldsymbol{x}, t) := \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ satisfies the following system of PDEs*

$$\partial_t \boldsymbol{s} - \nabla_{\boldsymbol{x}} \left[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}\rangle - \text{div}_{\boldsymbol{x}}(\boldsymbol{f})\right] = 0, \quad (\boldsymbol{x}, t) \in \mathbb{R}^D \times [0, T]. \tag{5}$$

This result shows that the time-conditional scores $\boldsymbol{s_\theta}(\boldsymbol{x}, t)$ learned by score-based models (via Eq. 3) are highly redundant. In principle, given a ground truth score at an initial time t_0, we can theoretically recover scores for all times $t \geq t_0$ by solving the score FPE. We explain its intuition by considering a special case when $\boldsymbol{f} \equiv \boldsymbol{0}$ and $g \equiv 1$. That is, $\boldsymbol{x}(t)$ is obtained by adding Gaussian noise. It is well-known that the densities q_t and q_{t_0} are related in a convolutional way as $q_t = q_{t_0} * \mathcal{N}(0, t)$, and q_t can be analytically obtained from q_{t_0} [8] (e.g., by applying a Fourier transform and dividing). Hence, all scores can in principle be obtained analytically from the score at a single time-step, without any further learning. In Appx. B we empirically support this idea.

Theoretically, with sufficient data and model capacity, (denoising) score matching ensures the optimal solution to Eq. 3 should satisfy Eq. 5 as it approximates the ground truth score well. However, we observe that pre-trained $\boldsymbol{s_\theta}$ learned via Eq. 3 do not numerically satisfy the score FPE. We hereby introduce an error term $\boldsymbol{\epsilon}_{\boldsymbol{s_\theta}} = \boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}(\boldsymbol{x}, t)$ in order to quantify how $\boldsymbol{s_\theta}$ deviates from the score FPE

$$\boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}(\boldsymbol{x}, t) := \partial_t \boldsymbol{s_\theta} - \nabla_{\boldsymbol{x}} \left[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s_\theta}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s_\theta}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s_\theta}\rangle - \text{div}_{\boldsymbol{x}}(\boldsymbol{f})\right]. \tag{6}$$

We further define the following averaged residuals of DSM and the score FPE for $t \in [0, 1]$:

$$r_{\text{DSM-like}}(t; \boldsymbol{s_\theta}) := \frac{1}{D}\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{\boldsymbol{x}(t)|\boldsymbol{x}(0)}\left[\left\|\boldsymbol{s_\theta}(\boldsymbol{x}(t), t) - \nabla_{\boldsymbol{x}(t)} \log q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))\right\|_2\right]$$

$$r_{\text{FP}}(t; \boldsymbol{s_\theta}) := \frac{1}{D}\mathbb{E}_{\boldsymbol{x}\sim\nu}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}(\boldsymbol{x}, t)\right\|_2\right], \quad \nu \sim \text{Uniform}\Big([0, 1]^D\Big) \text{ or } \nu \sim q_t(\boldsymbol{x}(t)|\boldsymbol{x}(0)).$$

Fig. 1 plots these residuals for score models pre-trained via DSM on MNIST and CIFAR-10. Despite achieving low $r_{\text{DSM-like}}$ score-matching loss across all t (green curve), pre-trained score models fail to satisfy the score FPE equation especially for small t (blue, orange curves). In Sec. 3, we theoretically analyze these findings and provide new insights into the score FPE. In Sec. 4, we propose a novel score matching objective with score FPE as a regularizer (Eq. 8) and examine its effectiveness on synthetic dataset and MNIST. We refer to Appx. C for implementation details and Appx. E for proofs.

3 Theoretical implications and interpretations of score FPE

In this section, we first study two implications of satisfying the score FPE. More precisely, we show in Sec. 3.1 that controlling $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}$ can implicitly enforce *conservativity* of $\boldsymbol{s}_{\boldsymbol{\theta}}$. Moreover, if the score FPE is satisfied, we prove in Sec. 3.2 the equivalence of $\boldsymbol{s}_{\boldsymbol{\theta}}$, ground truth score $\boldsymbol{s}$ and $\nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{\text{SDE}}$ holds under some conditions, where $p_{t,\boldsymbol{\theta}}^{\text{SDE}}$ is defined in Sec. 1 as the marginal density of parametrized diffusion process. In Sec. 3.3, we investigate the connection between *higher-order score matching* [9, 7] and score FPE.

3.1 Conservativity

The ground truth score $\boldsymbol{s}(\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ is a conservative vector field. That is, it can be expressed as a gradient of some real-valued function. However, scores learned in practice do not satisfy this property [12]. Below we prove that we can implicitly enforce conservativity by minimizing the time-averaged residual of the score FPE.

Proposition 1. *If there is a $t_{\boldsymbol{\theta}} \in [0, T]$ so that $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{x}} \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^D$, then there is a real-valued function $\Psi_{\boldsymbol{\theta}} \colon \mathbb{R}^D \times [0, T] \to \mathbb{R}$ given by $\Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x}) + \int_{t_{\boldsymbol{\theta}}}^{t} \Big[\frac{1}{2} g^2(\tau) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(\tau) \left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] d\tau$ so that for all $(\boldsymbol{x}, t) \in \mathbb{R}^D \times [0, T]$*

$$\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\right\|_2 \leq \int_{\min\{t_{\boldsymbol{\theta}}, t\}}^{\max\{t_{\boldsymbol{\theta}}, t\}} \left\|\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, \tau)\right\|_2 d\tau. \tag{7}$$

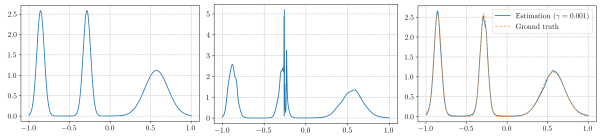
Consider a model $\boldsymbol{s}_{\boldsymbol{\theta}}$, and assume that for a large enough timestep t_θ, it captures exactly the perturbed density ($\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{x}} \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x})$) which is close to the prior (normal distribution) because $t_\theta \approx T$. Intuitively, $\boldsymbol{s}_{\boldsymbol{\theta}}$ is "nearly" conservative as the score of the prior is known to be conservative (because its score has a closed form as the gradient of log-density). Indeed, Prop .1 supports the intuition by saying that the the estimated score should nearly be conservative if it approximately satisfies the score FPE. Actually, Fig. 1 shows that $\int_{t_{\boldsymbol{\theta}}}^{t} \left\|\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}_0}}(\boldsymbol{x}, \tau)\right\|_2 d\tau$ is numerically small. Therefore, $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ is close to the gradient of a scalar function $\Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$; namely, it is conservative.

3.2 Equivalence between $\boldsymbol{s}_{\boldsymbol{\theta}}$, $\boldsymbol{s}$ and $\nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{\textbf{SDE}}$

We now investigate another implication of satisfying the score FPE which connects the score $\boldsymbol{s}_{\boldsymbol{\theta}}$ with the ground truth $\boldsymbol{s}$ and $\nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{\text{SDE}}$. The following proposition states that all aforementioned scores are identical if we train to reach a zero residual of score FPE for all $(\boldsymbol{x}, t)$ (under some technical assumptions ensuring the system of PDEs has a unique solution).

Proposition 2. *Suppose we know that in some suitable function space, $\mathbf{0}$ is the unique strong solution to the PDEs $\partial_t \boldsymbol{v} - \nabla_{\boldsymbol{x}} \left[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{v}) + \frac{1}{2} g^2(t) \big(\left\|\boldsymbol{v}\right\|_2^2 + 2\langle \boldsymbol{v}, \boldsymbol{s} \rangle\big) - \langle \boldsymbol{f}, \boldsymbol{v} \rangle\right] = 0$ with zero initial condition $\boldsymbol{v}(\boldsymbol{x}, 0) \equiv 0$ and zero boundary condition. If there is some $\boldsymbol{\theta}_0$ so that $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}_0}}(\boldsymbol{x}, t) = 0$ for all $(\boldsymbol{x}, t)$, then $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{s}$. Moreover, suppose that the PDEs $\partial_t \boldsymbol{v} + \nabla_{\boldsymbol{x}} \left[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{v}) + \frac{1}{2} g^2(t) \left\|\boldsymbol{v}\right\|_2^2 + \langle \boldsymbol{f}, \boldsymbol{v} \rangle\right] = 0$ with zero initial and boundary condition has $\mathbf{0}$ as the unique strong solution, then $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}_0}} \equiv 0$ implies $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}_0}^{\text{SDE}}$.*

We hypothesize that satisfying the score FPE has a *smoothing effect* when $\boldsymbol{f}(\boldsymbol{x}, t)$ is linear in $\boldsymbol{x}$. Suppose the assumptions of Prop. 2 hold and hence, $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}_0}^{\text{SDE}}$. As linearly transforming normal distributions (by $\boldsymbol{f}$) remains normal, [7] proves that $p_{t,\boldsymbol{\theta}_0}^{\text{SDE}}$ turns out to be a Gaussian distribution for any t. In practice, the assumptions are not likely to be met exactly, i.e. the residual will not be exactly zero $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}_0}} \equiv 0$. In this case, we hypothesize that learning $\boldsymbol{\theta}$ to reduce $\left\|\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}\right\|$ can reduce the gap $\left\|\boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{\text{SDE}}\right\|$, and may further modify the direction $\boldsymbol{s}_{\boldsymbol{\theta}}$ toward high density region of Gaussian (*smoothing effect*).



(a) Data density (b) VE (c) VE $+ 0.001$FP

Figure 2: Comparison of (a) data density, (b) estimated density by probability flow ODE with $\boldsymbol{s_\theta}$ trained with $\gamma = 0.0$, and (c) with $\gamma = 0.001$. Score FPE-regularizer improves density estimation.

Table 1: NLL (in bpd) for different FP weights γ's and weighting functions $\lambda(\cdot)$'s

SDE type	$\lambda(\cdot)$	$\gamma = 0.0$	$\gamma = 0.01$	$\gamma = 0.1$	$\gamma = 1.0$	$\gamma = 10.0$
VE + γFP	[16]	3.86	3.63	3.66	**3.28**	3.37
	[15]	3.63	3.94	3.53	**3.20**	3.23
VP + γFP	[16]	2.95	3.06	3.09	**2.91**	3.34
	[15]	3.11	3.14	**3.04**	3.28	3.28
RVE + γFP	[16]	3.45	3.68	3.77	3.57	**3.13**
	[15]	3.62	3.78	3.49	**3.16**	3.36

3.3 Higher-order score matching

Higher-order derivatives of score can yield additional information about the data distribution. We prove a property stating that error bounds of higher-order score matching can further control the residual $\left\|\int_0^t \boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}(\boldsymbol{x}, \tau) d\tau\right\|_2$ for all $t \in [0, T]$. This may explain why the scores learned via Eq. 3 are not sufficient to satisfy the score FPE as their higher-order scores may deviate from the ground truth.

Proposition 3. *Denote $C(t) := \frac{1}{2}\int_0^t g^2(\tau) d\tau$. Assume the following error estimates hold for higher-order score matching that for all $(\boldsymbol{x}, t) \in \mathbb{R}^D \times [0, T]$*

$$\|\boldsymbol{s} - \boldsymbol{s_\theta}\|_2 \leq \delta_0, \quad \|\nabla_{\boldsymbol{x}}(\boldsymbol{s} - \boldsymbol{s_\theta})\|_F \leq \delta_1, \quad \|\nabla_{\boldsymbol{x}} \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s} - \boldsymbol{s_\theta})\|_2 \leq \delta_2.$$

Then we have that for all $(\boldsymbol{x}, t) \in \mathbb{R}^D \times [0, T]$

$$\left\|\int_0^t \boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}(\boldsymbol{x}, \tau) d\tau\right\|_2 \leq 2\delta_0 + (\delta_2 + 2\delta_1\delta_0)C(t) + \delta_1 \int_0^t \left(g^2(\tau) \|\boldsymbol{s}(\boldsymbol{x}, \tau)\|_2 + \|\boldsymbol{f}(\boldsymbol{x}, \tau)\|_2\right) d\tau$$
$$+ \delta_0 \int_0^t \left(g^2(\tau) \|\nabla_{\boldsymbol{x}} \boldsymbol{s}(\boldsymbol{x}, \tau)\|_F + \|\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, \tau)\|_F\right) d\tau.$$

4 Experimental results

We have demonstrated that score models learned via $\mathcal{J}_{\mathrm{DSM}}$ (Eq. 3) do not satisfy the score FPE, a property that ground truth scores should satisfy *a priori*. Therefore, we devise a novel loss function, consisting of $\mathcal{J}_{\mathrm{DSM}}$ and a *score FPE-regularizer* $\mathcal{R}_{\mathrm{FP}}(\boldsymbol{\theta}) := \frac{1}{D}\mathbb{E}_{t \sim \mathcal{U}[0,T]}\mathbb{E}_{\boldsymbol{x} \sim \nu} \|\boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}(\boldsymbol{x}, t)\|_2$ as

$$\mathcal{J}_{\mathrm{FP}}(\boldsymbol{\theta}; \lambda(\cdot), \gamma) := \mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}; \lambda(\cdot)) + \gamma \mathcal{R}_{\mathrm{FP}}(\boldsymbol{\theta}), \tag{8}$$

where $\gamma \geq 0$ is a hyper-parameter. Since $\boldsymbol{\epsilon}_{\boldsymbol{s_\theta}}$ in $\mathcal{R}_{\mathrm{FP}}$ is generally expensive to calculate, we propose to exploit the finite difference method [3] for $\partial_t \boldsymbol{s_\theta}$ and Hutchinson's trace estimator [5] for $\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s_\theta})$ to reduce the computational efforts (details in Appx. D). The effectiveness of $\mathcal{J}_{\mathrm{FP}}$ is examined on synthetic dataset (Gaussian mixture models) and MNIST.

Synthetic dataset We consider a Gaussian mixture model as the training data distribution. Fig. 2 illustrates (a) ground truth density, and the density produced by probability flow ODE [16] of scores trained with (b) $\lambda = 0.0$ (i.e, conventional score matching training) and (c) $\lambda = 0.001$. The score trained with score FPE-regularizer can approximate the data density well, improving over vanilla score-matching. We hypothesize score FPE-regularizer may improve density estimation with the probability flow ODE, as it enforces a known self-consistency property of the ground truth score.

MNIST We evaluate the proposed $\mathcal{J}_{\mathrm{FP}}(\boldsymbol{\theta}; \lambda(\cdot), \gamma)$ on MNIST with different γ's. Table 1 reports negative log-likelihood (NLL) in bits/dim (bpd) across three instantiations of the forward SDE (see Appx. A) and two choices of weighting functions $\lambda(\cdot)$'s ([16] and [15]). We observe a general improvement in NLL with $\gamma = 1.0$ (see Appx. F for a demonstration of generated samples).

5 Conclusion

We introduce the score FPE and theoretically study its relation with score matching, conservativity and density induced by parametric reverse diffusion. Moreover, we propose to penalize on residual of score FPE and show its effectiveness on simple dataset. However, it is unclear how the dynamics of score FPE affects, for instance, training of a larger scale dataset or variational lower bound.

References

[1] Brian DO Anderson. “Reverse-time diffusion equation models”. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326.

[2] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. “Score-based generative modeling with critically-damped langevin diffusion”. In: *arXiv preprint arXiv:2112.07068* (2021).

[3] Bengt Fornberg. “Generation of finite difference formulas on arbitrarily spaced grids”. In: *Mathematics of computation* 51.184 (1988), pp. 699–706.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[5] Michael F Hutchinson. “A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines”. In: *Communications in Statistics-Simulation and Computation* 18.3 (1989), pp. 1059–1076.

[6] Dongjun Kim et al. “Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 11201–11228.

[7] Cheng Lu et al. “Maximum Likelihood Training for Score-based Diffusion ODEs by High Order Denoising Score Matching”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 14429–14460.

[8] Elias Masry and John A Rice. “Gaussian deconvolution via differentiation”. In: *Canadian Journal of Statistics* 20.1 (1992), pp. 9–21.

[9] Chenlin Meng et al. “Estimating high order gradients of the data distribution by denoising”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25359–25369.

[10] Bernt Øksendal. “Stochastic differential equations”. In: *Stochastic differential equations*. Springer, 2003, pp. 65–84.

[11] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”. In: *Journal of Computational physics* 378 (2019), pp. 686–707.

[12] Tim Salimans and Jonathan Ho. “Should EBMs model the energy or the score?” In: *Energy Based Models Workshop-ICLR 2021*. 2021.

[13] Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.

[14] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[15] Yang Song et al. “Maximum likelihood training of score-based diffusion models”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1415–1428.

[16] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *arXiv preprint arXiv:2011.13456* (2020).

[17] Yang Song et al. “Sliced score matching: A scalable approach to density and score estimation”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 574–584.

[18] Pascal Vincent. “A connection between score matching and denoising autoencoders”. In: *Neural computation* 23.7 (2011), pp. 1661–1674.

A Instantiation of SDE and score FPE

[16] categorizes the forward SDE into three types based on the behavior of the variance during evolution. Here we focus on two of them, which are V̲ariance E̲xplosion (VE) SDE and V̲ariance P̲reserving (VP) SDE.

VE SDE It has a zero drift term $\boldsymbol{f} = 0$ and diffusion term $g(t) = \sqrt{\frac{d\sigma^2(t)}{dt}}$ with some function $\sigma(t)$. Hence, the forward SDE (Eq. 1) becomes

$$d\boldsymbol{x}(t) = \sqrt{\frac{d\sigma^2(t)}{dt}} d\boldsymbol{w}_t. \tag{9}$$

A typical instance of VE SDE is Score Matching of Langevin dynamics (SMLD) [14], where $\sigma(t) := \sigma_{\text{min}} \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t$ for $t \in (0, 1]$. In our implementation, we follow the conventional setup of $(\sigma_{\text{min}}, \sigma_{\text{max}}) := (0.01, 50)$.

In [6], they proposed a variant of VE SDE attempting to resolve the unbounded score problem [2], which is called R̲eciprocal V̲E̲ (RVE). Let $\epsilon > 0$ be a fixed constant. RVE SDE also has zero drift term but with a different parametrization for diffusion

$$g(t) := \begin{cases} \sigma_{\text{max}} \left(\frac{\sigma_{\text{min}}}{\sigma_{\text{max}}}\right)^{\frac{\epsilon}{t}} \frac{\sqrt{2\epsilon \log(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}})}}{t}, & \text{if } t > 0 \\ 0, & \text{if } t = 0 \end{cases}$$

VP SDE Let β be a non-negative function of t. VP SDE has a linear drift term $\boldsymbol{f}(\boldsymbol{x}, t) = -\frac{1}{2}\beta(t)\boldsymbol{x}$ and diffusion term $g(t) = \sqrt{\beta(t)}$. Thus, the forward SDE is

$$d\boldsymbol{x}(t) = -\frac{1}{2}\beta(t)\boldsymbol{x}(t)dt + \sqrt{\beta(t)}d\boldsymbol{w}_t.$$

A classic example of VP SDE is Denoising Diffusion Probabilistic Modeling (DDPM) [13, 4], where $\beta(t) := \beta_{\text{min}} + t(\beta_{\text{max}} - \beta_{\text{min}})$ for $t \in [0, 1]$. We adopt the common setup of $(\beta_{\text{min}}, \beta_{\text{max}}) := (0.1, 20)$ in our implementation.

We summarize the aforementioned instantiations of SDE and their associated score FPE in Table 2.

Table 2: Summary of the forward SDEs and their score FPEs

	VE SDE	RVE SDE	VP SDE
$\boldsymbol{f}(\boldsymbol{x}, t)$	$\boldsymbol{0}$		$-\frac{1}{2}\beta(t)\boldsymbol{x}$
$g(t)$	$\sigma_{\text{min}} \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t \sqrt{2 \log \left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)}$	$\begin{cases} \sigma_{\text{max}} \left(\frac{\sigma_{\text{min}}}{\sigma_{\text{max}}}\right)^{\frac{\epsilon}{t}} \frac{\sqrt{2\epsilon \log(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}})}}{t}, & t > 0 \\ 0, & t = 0 \end{cases}$	$\sqrt{\beta(t)}$
SDE	$d\boldsymbol{x}(t) = g(t)d\boldsymbol{w}_t$		$d\boldsymbol{x}(t) = -\frac{1}{2}\beta(t)\boldsymbol{x}(t)dt + \sqrt{\beta(t)}d\boldsymbol{w}_t$
score FPE	$\partial_t \boldsymbol{s} = \nabla_{\boldsymbol{x}} \left[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \frac{1}{2}g^2(t) \lVert\boldsymbol{s}\rVert_2^2\right]$		$\partial_t \boldsymbol{s} = \frac{1}{2}\beta(t)\nabla_{\boldsymbol{x}} \left[\text{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \lVert\boldsymbol{s}\rVert_2^2 + \langle \boldsymbol{x}, \boldsymbol{s} \rangle\right]$

B How scores satisfy score FPE?

We experimentally demonstrate how score functions should satisfy the score FPE in two different aspects. We consider the data distribution as a Gaussian mixture model (GMM) of the density $\frac{1}{5}\mathcal{N}\big((-5, -5), \boldsymbol{I}\big) + \frac{4}{5}\mathcal{N}\big((5, 5), \boldsymbol{I}\big)$ on $\mathbb{R}^2$ whose samples are illustrated in Fig. 3a. The diffusion process is taken as VE SDE (Eq. 9). The ground truth score of GMM, denoted as $\boldsymbol{s}^{\text{GMM}}$, can be expressed explicitly with a closed formula throughout the diffusion (as the diffusion process is linear in $\boldsymbol{x}$).

First of all, we examine if $\boldsymbol{s}^{\text{GMM}}$ satisfies the score FPE by computing $r_{\text{FP}}(t; \boldsymbol{s}^{\text{GMM}})$ and plot it in Fig. 5a (blue curve). We can see that the score FPE residual of the ground truth is almost zero, which empirically supports Corollary 1.

Second, as we explain in Sec. 2, we can solve score FPE for the score at any time if we are merely given a score at a single time moment. Namely, once we find a solution $\tilde{\boldsymbol{s}}$ to the following initial value problem of system of PDEs, we know a score at all time.

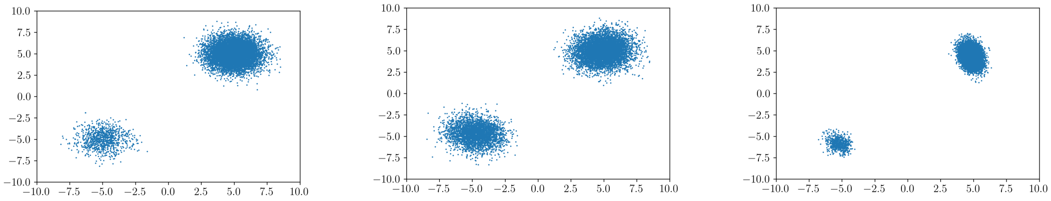
$$\begin{cases} \partial_t \tilde{\boldsymbol{s}} &= \nabla_{\boldsymbol{x}}\left[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\tilde{\boldsymbol{s}}) + \frac{1}{2}g^2(t)\left\|\tilde{\boldsymbol{s}}\right\|_2^2\right], \quad (\boldsymbol{x},t) \in \mathbb{R}^D \times (0,T] \\ \tilde{\boldsymbol{s}}(\boldsymbol{x},0) &= \boldsymbol{s}^{\mathrm{GMM}}(\boldsymbol{x},0), \quad \boldsymbol{x} \in \mathbb{R}^D \end{cases} \tag{10}$$

Solutions of Eq. 10 can be parametrized as neural network $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$ [11]. We then can solve the PDEs by learning parameters $\boldsymbol{\theta}$ to reduce both the residuals of the initial condition $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}(\boldsymbol{x},0) - \boldsymbol{s}^{\mathrm{GMM}}(\boldsymbol{x},0)$ and evolution $\boldsymbol{\epsilon}_{\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}} := \partial_t \tilde{\boldsymbol{s}}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}}\left[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(t)\left\|\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}\right\|_2^2\right]$. That is, $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$ is learned with the *score FPE-guided* objective function:

$$\min_{\boldsymbol{\theta}} \left\{ \mathbb{E}_{t\sim\mathcal{U}[0,T]}\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))}\left\|\boldsymbol{\epsilon}_{\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}}(\boldsymbol{x},t)\right\|_2 + \mathbb{E}_{\boldsymbol{x}(0)}\left\|\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}(\boldsymbol{x},0) - \boldsymbol{s}^{\mathrm{GMM}}(\boldsymbol{x},0)\right\|_2 \right\}. \tag{11}$$

We demonstrate generated samples by the learnt $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$ in Fig. 3b and plot its score FPE residual $r_{\mathrm{FP}}(t;\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}})$ in Fig. 5a (orange curve). Interestingly, it also generates quite satisfactory samples and Fig. 4b show it estimates the ground truth score (Fig. 4a) well. This supports our argument.

On the other hand, we compare with a score $\boldsymbol{s}_{\theta}^{\mathrm{GMM}}$ learned from the denoising score matching (Eq. 3). We observe that from Fig. 5b that $\boldsymbol{s}_{\theta}^{\mathrm{GMM}}$ does not satisfy score FPE even though it works decently on generation (Fig. 3c) and score estimation (Fig. 4c).

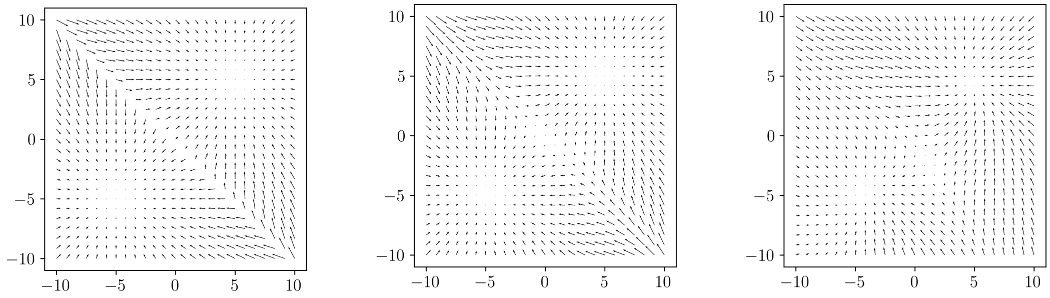


(a) Ground truth data

(b) Samples generated by $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$

(c) Samples generated by $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$

Figure 3: Comparison of instances generated using the score functions learned by our score FPE-guided objective fucntion (Eq. 11) and the conventional denoising score matching (Eq. 3), which are denoted as $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$, respectively. Both scores can synthesize reasonable quality samples.

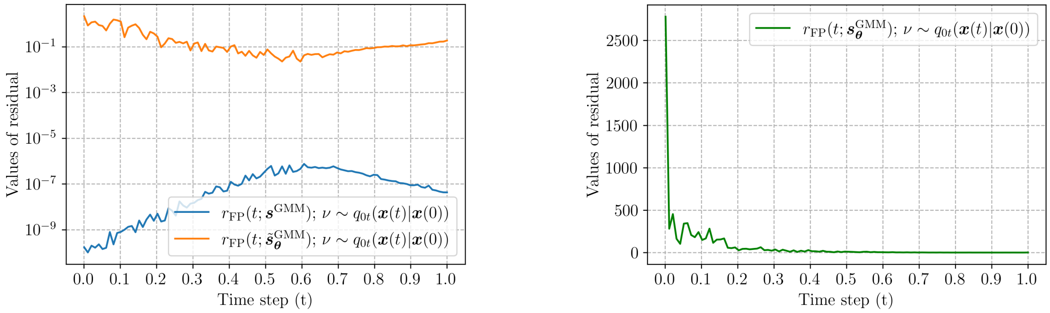


(a) Ground truth score at $t = 0$

(b) Estimated Score $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}(\cdot, 0)$

(c) Estimated Score $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{GMM}}(\cdot, 0)$

Figure 4: The fluid flow graph of ground truth score and estimated scores at $t = 0$ by $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$. The score $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$, which is learned from the score FPE-guided objective, can also approximate the ground truth well.



(a) FP residuals of ground truth score and the score learned from Eq. 11

(b) FP residuals of the score learned from Eq. 3

Figure 5: Comparison of the score FPE residuals of $\boldsymbol{s}^{\mathrm{GMM}}$, $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$. A further evidence of the claim in Sec. 2 that $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{GMM}}$, which is learned from denoising score matching, does not satisfy score FPE.

C Explanation of Implementation

In Fig. 1, we train a score on MNIST for 200 epochs with a learning rate $1e-3$ and batch size 32 by using an identical neural network structure to the repository [1] but modify the forward SDE as VE SDE or VP SDE (see Appx. A). The network structure of synthetic dataset in Appx. B is similar to the aforementioned one but we simply replace all convolutional layers with fully connected layers. We train for $2,000$ epochs with a learning rate $1e-3$ and batch size 500.

For the case of CIFAR10, we use the pre-trained score models provided by [16] [2] instead of training them from scratch. The VE SDE and VP SDE are taken as NCSN++ cont. and DDPM++ cont., respectively.

The neural network setup in Fig. 2 is the same as toy model structures provided in the repository of [7] [3]. We found out setting the weight of score FPE to $\lambda = 0.001$ can generally work well for the toy dataset.

D Techniques to reduce computation costs for score FPE

The computation of $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t)$ in $\mathcal{R}_{\mathrm{FP}}(\boldsymbol{\theta})$ is expensive and hard to scale-up to higher dimensional data. We thus propose two techniques which can help reduce the computation costs.

We recall that $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t)$ is defined as $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(t) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f}) \Big]$. We explain the details of how the computation of $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ and $\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$ can be respectively relieved by the finite difference method and Hutchinson's trace estimator.

D.1 $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ term

Lemma 1. *[3] Let $\alpha \colon [0, 1] \to \mathbb{R}^D$ be a vector-valued function which is continuously differentiable up to third order derivatives. Denote h_s and h_d as hyper-parameters of step sizes. Then we have the following estimate of $\alpha'(t)$:*

$$\frac{h_s^2 \alpha(t + h_d) + (h_d^2 - h_s^2)\alpha(t) - h_d^2 \alpha(t - h_s)}{h_s h_d (h_s + h_d)} + \mathcal{O}\Big(\frac{h_d h_s^2 + h_s h_d^2}{h_s + h_d}\Big).$$

In particular, if $h_s = h_d =: h$, then the estimate becomes

$$\frac{\alpha(t + h) - \alpha(t - h)}{2h} + \mathcal{O}(h^2).$$

[1] https://colab.research.google.com/drive/120kYYBOVa1iOTD85RjlEkFjaWDxSFUx3?usp=sharing

[2] https://github.com/yang-song/score_sde_pytorch

[3] https://github.com/LuChengTHU/mle_score_ode

In our implementation for MNIST, we consider $\alpha(\cdot) := \boldsymbol{s}_{\boldsymbol{\theta}}(\cdot, \boldsymbol{x})$ and set $(h_s, h_d) = (0.001, 0.0005)$ for the approximation of $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$.

D.2 $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$ term

Hutchinson's trace estimator [5] stochastically estimates the trace tr($\boldsymbol{A}$) of any square matrix $\boldsymbol{A}$. Its idea is choose a distribution $p_{\boldsymbol{v}}$ so that $\mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\boldsymbol{v}] = \boldsymbol{0}$ and $\mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\boldsymbol{v}\boldsymbol{v}^T] = \boldsymbol{I}$. Hence, $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}\mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\boldsymbol{v}\boldsymbol{v}^T]) = \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\text{tr}(\boldsymbol{A}\boldsymbol{v}\boldsymbol{v}^T)] = \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\text{tr}(\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}^T)] = \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}^T]$. By i.i.d. sampling $\{\boldsymbol{v}_j\}_{j=1}^M$ from $p_{\boldsymbol{v}}$, we can use the following unbiased estimator

$$\frac{1}{M}\sum_{j=1}^{M} \boldsymbol{v}_j \boldsymbol{A} \boldsymbol{v}_j^T$$

to estimate tr($\boldsymbol{A}$). We notice that $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)) = \text{tr}\big(\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}\big)$. Thus, we can apply Hutchinson's trick and replace $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$ term with the estimation

$$\frac{1}{M}\sum_{j=1}^{M} \boldsymbol{v}_j \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \boldsymbol{v}_j^T.$$

In the implementation, $p_{\boldsymbol{v}}$ is usually taken as a standard normal distribution or a Rademacher distribution whose random vector has components equal to $+1$ or -1 with equal probability. We follow the convention in [17] which sets $M = 1$ and shows its effectiveness.

D.3 Projection of $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t)$

We further propose another potential trick to reduce the computation cost of differentiation.

$$\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t) = \underbrace{\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}}_{\text{(I)}} - \underbrace{\nabla_{\boldsymbol{x}}\left[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \text{div}_{\boldsymbol{x}}(\boldsymbol{f})\right]}_{\text{(II)}} \tag{12}$$

Using automatic differentiation to compute the gradient in $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t)$ (the (II) part in Eq. 12) is generally cumbersome for high dimensional data. We propose to use random projection to relieve the computation of gradient (multi-dimension) to directional derivative (one-dimensional). Thus, we can further apply the finite difference trick introduced in Appx. D.1 to reduce the computation efforts. We first recall a fundamental property before rigorously formulating the technique.

Lemma 2. *Let $M := M(\boldsymbol{x}, t) \colon \mathbb{R}^D \times [0, T] \to \mathbb{R}$ be a continuously differentiable function of $\boldsymbol{x}$. For any $\boldsymbol{v} \in \mathbb{R}^D$,*

$$D_{\boldsymbol{v}} M(\boldsymbol{x}, t) = \langle \nabla_{\boldsymbol{x}} M(\boldsymbol{x}, t), \boldsymbol{v}\rangle,$$

where $D_{\boldsymbol{v}} M(\boldsymbol{x}, t)$ means the directional derivative of M in $\boldsymbol{x}$ along the direction $\boldsymbol{v}$ which is defined as:

$$D_{\boldsymbol{v}} M(\boldsymbol{x}, t) := \lim_{h\to 0}\frac{M(\boldsymbol{x} + h\boldsymbol{v}, t) - M(\boldsymbol{x}, t)}{h} = \frac{d}{dh}M(\boldsymbol{x} + h\boldsymbol{v}, t)\Big|_{h=0}. \tag{13}$$

For simplicity, let us denote $M(\boldsymbol{x}, t) := \frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \text{div}_{\boldsymbol{x}}(\boldsymbol{f})$ and let $\boldsymbol{v} \in \mathbb{R}^D$ be arbitrary vector. We project $\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t)$ along direction $\boldsymbol{v}$ and apply the Lemma 2,

$$\langle \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t), \boldsymbol{v}\rangle = \langle \partial_t \boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}} M(\boldsymbol{x}, t), \boldsymbol{v}\rangle = \langle \partial_t \boldsymbol{s}_{\boldsymbol{\theta}}, \boldsymbol{v}\rangle - \langle \frac{d}{dh}M(\boldsymbol{x} + h\boldsymbol{v}, t)\Big|_{h=0}, \boldsymbol{v}\rangle.$$

Notice that both $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ and $\frac{d}{dh}M(\boldsymbol{x} + h\boldsymbol{v}, t)\Big|_{h=0}$ are one-dimensional differentiation, which can be estimated via Lemma 1 and hence, we can avoid automatic differentiation. We hereafter propose an *estimated score FPE-regularizer* which may replace $\mathcal{R}_{\text{FP}}$ with

$$\hat{\mathcal{R}}_{\text{FP}}(\boldsymbol{\theta}) := \frac{1}{D}\mathbb{E}_{t\sim\mathcal{U}[0,T]}\mathbb{E}_{\boldsymbol{x}\sim\nu}\mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}\left|\langle \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, t), \boldsymbol{v}\rangle\right|, \tag{14}$$

where $p_{\boldsymbol{v}}$ is a distribution of random vector $\boldsymbol{v} \in \mathbb{R}^D$. We observe that the performance may degrade by using $\hat{\mathcal{R}}_{\text{FP}}$, which may due to the inaccurate approximation to the exact score FPE. Therefore, a further study is required to have lower computation costs while preventing the deterioration in the performance.

E Proofs and discussion

E.1 Proof of Corollary 1

We prove the result with a more general forward SDE

$$d\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x},t)dt + \boldsymbol{G}(\boldsymbol{x},t)d\boldsymbol{w}_t, \tag{15}$$

where $\boldsymbol{F}(\cdot,t)\colon \mathbb{R}^D \to \mathbb{R}^D$ and $\boldsymbol{G}(\cdot,t)\colon \mathbb{R}^D \to \mathbb{R}^{D\times D}$.

We know that the density $q_t(\boldsymbol{x})$ satisfies the Fokker-Planck equation [10]

$$\partial_t q_t(\boldsymbol{x}) = -\sum_{j=1}^{D} \partial_{x_j}\big(\tilde{\boldsymbol{F}}_j(\boldsymbol{x},t)q_t(\boldsymbol{x})\big), \tag{16}$$

where $\tilde{\boldsymbol{F}}(\boldsymbol{x},t) := \boldsymbol{F}(\boldsymbol{x},t) - \frac{1}{2}\nabla\cdot[\boldsymbol{G}(\boldsymbol{x},t)\boldsymbol{G}(\boldsymbol{x},t)^T] - \frac{1}{2}\boldsymbol{G}(\boldsymbol{x},t)\boldsymbol{G}(\boldsymbol{x},t)^T\nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x})$. We further denote $\boldsymbol{A}(\boldsymbol{x},t) := \boldsymbol{F}(\boldsymbol{x},t) - \frac{1}{2}\nabla\cdot[\boldsymbol{G}(\boldsymbol{x},t)\boldsymbol{G}(\boldsymbol{x},t)^T]$ and $\boldsymbol{B}(\boldsymbol{x},t) := -\frac{1}{2}\boldsymbol{G}(\boldsymbol{x},t)\boldsymbol{G}(\boldsymbol{x},t)^T$.

Now $\tilde{\boldsymbol{F}}(\boldsymbol{x},t) = \boldsymbol{A}(\boldsymbol{x},t) + \boldsymbol{B}(\boldsymbol{x},t)\boldsymbol{s}(\boldsymbol{x},t)$, and we have

$$\begin{align}
\partial_t \log q_t(\boldsymbol{x}) &= \frac{1}{q_t(\boldsymbol{x})}\partial_t q_t(\boldsymbol{x}) \tag{17}\\
&= -\frac{1}{q_t(\boldsymbol{x})}\sum_{j=1}^{D}\partial_{x_j}\big(\tilde{\boldsymbol{F}}_j(\boldsymbol{x},t)q_t(\boldsymbol{x})\big) \tag{18}\\
&= -\frac{1}{q_t(\boldsymbol{x})}\sum_{j=1}^{D}\big(\partial_{x_j}\tilde{\boldsymbol{F}}_j(\boldsymbol{x},t)q_t(\boldsymbol{x}) + \tilde{\boldsymbol{F}}_j(\boldsymbol{x},t)\partial_{x_j}q_t(\boldsymbol{x})\big) \tag{19}\\
&= -\sum_{j=1}^{D}\big(\partial_{x_j}\tilde{\boldsymbol{F}}_j(\boldsymbol{x},t) + \tilde{\boldsymbol{F}}_j(\boldsymbol{x},t)\partial_{x_j}\log q_t(\boldsymbol{x})\big) \tag{20}\\
&= -\big(\mathrm{div}_{\boldsymbol{x}}(\tilde{\boldsymbol{F}}) + \langle\tilde{\boldsymbol{F}},\boldsymbol{s}\rangle\big) \tag{21}\\
&= -\Big[\mathrm{div}_{\boldsymbol{x}}\big(\boldsymbol{B}\boldsymbol{s}\big) + \langle\boldsymbol{B}\boldsymbol{s},\boldsymbol{s}\rangle + \langle\boldsymbol{A},\boldsymbol{s}\rangle + \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{A})\Big] \tag{22}\\
&= \frac{1}{2}\mathrm{div}_{\boldsymbol{x}}\big(\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{s}\big) + \frac{1}{2}\left\|\boldsymbol{G}^T\boldsymbol{s}\right\|_2^2 - \langle\boldsymbol{A},\boldsymbol{s}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{A}). \tag{23}
\end{align}$$

Since $\log q_t(\boldsymbol{x})$ is sufficiently smooth, we can swap the order of differentiations and get $\partial_t\boldsymbol{s} = \partial_t\nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}\partial_t\log q_t(\boldsymbol{x})$. Hence, the statement is proved.

■

Remark 1. *In Eq. 1 where $\boldsymbol{G}$ does not depend on $\boldsymbol{x}$, namely $\boldsymbol{G}(\boldsymbol{x},t) \equiv g(t)\boldsymbol{I}$, then $\tilde{\boldsymbol{F}}(\boldsymbol{x},t) = \boldsymbol{f}(\boldsymbol{x},t) - \frac{1}{2}g^2(t)\nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x})$ and*

$$\begin{align}
\partial_t\log q_t(\boldsymbol{x}) &= \frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}\right\|_2^2 - \langle\boldsymbol{f},\boldsymbol{s}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f}) \tag{24}\\
\partial_t\boldsymbol{s} &= \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}\right\|_2^2 - \langle\boldsymbol{f},\boldsymbol{s}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big]. \tag{25}
\end{align}$$

E.2 Proof of Proposition 1

Integrating the following equation w.r.t. time from $\tau = t_{\boldsymbol{\theta}}$ to $\tau = t$ with $t \in [0,T]$ fixed,

$$\partial_t\boldsymbol{s}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle\boldsymbol{f},\boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] + \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x},t),$$

leads to

$$\begin{aligned}
\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{x}}\Big\{\int_{t_{\boldsymbol{\theta}}}^{t}\Big[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle\boldsymbol{f},\boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big]d\tau\Big\}\\
+ \int_{t_{\boldsymbol{\theta}}}^{t}\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x},t)d\tau,
\end{aligned}$$

where the swap of integration and differentiation is valid if the integrand is sufficiently smooth.

With the assumption, we obtain that for all $t \in [0, T]$

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Big\{ \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x}) + \int_{t_{\boldsymbol{\theta}}}^{t} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(t) \left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] d\tau \Big\}$$
$$= \int_{t_{\boldsymbol{\theta}}}^{t} \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, \tau) d\tau.$$

By taking the norm of the above equation, one can obtain

$$\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\right\|_2 = \left\| \int_{t_{\boldsymbol{\theta}}}^{t} \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, \tau) d\tau \right\|_2 .$$

From which we obtain

$$\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\right\|_2 = \left\| \int_{t_{\boldsymbol{\theta}}}^{t} \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, \tau) d\tau \right\|_2 \leq \left| \int_{t_{\boldsymbol{\theta}}}^{t} \left\|\boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, \tau)\right\|_2 d\tau \right| .$$

The upper and lower bound of integral can be respectively written as $\max\{t, t_{\boldsymbol{\theta}}\}$ and $\min\{t, t_{\boldsymbol{\theta}}\}$, and whence, the proposition is proved. ■

E.3 Proof of Proposition 2

Lemma 3. *Let $\boldsymbol{s}_{\boldsymbol{\theta}}$ be a score obtained from denoising score matching (Eq. 3) and write $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}} := \nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{SDE}$. Then*

1. *[7] Eq. 4 associates with the following forward SDE whose marginal density is $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}$:*

$$d\boldsymbol{x}_{\boldsymbol{\theta}}(t) = \Big[\boldsymbol{f}(\boldsymbol{x}_{\boldsymbol{\theta}}(t), t) + g^2(t)\big(\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x}_{\boldsymbol{\theta}}(t), t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_{\boldsymbol{\theta}}(t), t)\big)\Big] dt + g(t)\boldsymbol{w}_t \tag{26}$$

2. *$\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}$ satisfies the following score FPE:*

$$\partial_t \hat{\boldsymbol{s}}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}\big(2\boldsymbol{s}_{\boldsymbol{\theta}} - \hat{\boldsymbol{s}}_{\boldsymbol{\theta}}\big) + \frac{1}{2} g^2(t) \big(2\langle \boldsymbol{s}_{\boldsymbol{\theta}}, \hat{\boldsymbol{s}}_{\boldsymbol{\theta}} \rangle - \left\|\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}\right\|_2^2\big) - \langle \boldsymbol{f}, \hat{\boldsymbol{s}}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] = 0. \tag{27}$$

Proof of Lemma 3. Consider

$$\boldsymbol{F}(\boldsymbol{x}, t) := \boldsymbol{f}(\boldsymbol{x}, t) + g^2(t)(\hat{\boldsymbol{s}}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}) \quad \text{and} \quad \boldsymbol{G}(\boldsymbol{x}, t) := g(t)\boldsymbol{I}$$

in Eq. 15, and apply Corollary 1, the lemma is then established. ■

Proof of Proposition 2. We recall Eq. 6, which indicates

$$\partial_t \boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(t) \left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] - \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}} = 0. \tag{28}$$

First, we subtract Eq. 27 by the above equation and get

$$\partial_t (\hat{\boldsymbol{s}}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \hat{\boldsymbol{s}}_{\boldsymbol{\theta}}) - \frac{1}{2} g^2(t) \left\|\boldsymbol{s}_{\boldsymbol{\theta}} - \hat{\boldsymbol{s}}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \hat{\boldsymbol{s}}_{\boldsymbol{\theta}} \rangle\Big] + \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}} = 0. \tag{29}$$

Consider when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and let $\boldsymbol{u}_{\boldsymbol{\theta}_0} := \hat{\boldsymbol{s}}_{\boldsymbol{\theta}_0} - \boldsymbol{s}_{\boldsymbol{\theta}_0}$. Then the PDEs become

$$\partial_t \boldsymbol{u}_{\boldsymbol{\theta}_0} + \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{u}_{\boldsymbol{\theta}_0}) + \frac{1}{2} g^2(t) \left\|\boldsymbol{u}_{\boldsymbol{\theta}_0}\right\|_2^2 + \langle \boldsymbol{f}, \boldsymbol{u}_{\boldsymbol{\theta}_0} \rangle\Big] = 0.$$

Here, $\boldsymbol{u}_{\boldsymbol{\theta}_0}$ is a solution to the PDEs. It is noticed that this system of PDEs has a zero initial condition and zero boundary condition as both $\boldsymbol{s}_{\boldsymbol{\theta}_0}$ and $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}_0}$ share the same initial/boundary condition. Thus, from the assumption of the uniqueness of solution, we know that $\boldsymbol{u}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{0}$, and hence, $\hat{\boldsymbol{s}}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{s}_{\boldsymbol{\theta}_0}$.

We repeat the same trick to subtract Eq. 5 by Eq. 28 from which we can obtain $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{s}$. ■

E.4 Proof of Proposition 3

By subtracting the following two equations

$$\begin{aligned}
\partial_t \boldsymbol{s}_{\boldsymbol{\theta}} &= \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(t) \left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] + \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}} && (30) \\
\partial_t \boldsymbol{s} &= \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \frac{1}{2} g^2(t) \left\|\boldsymbol{s}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big], && (31)
\end{aligned}$$

we obtain

$$\partial_t (\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}) = \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}) + \frac{1}{2} g^2(t) \big(\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \left\|\boldsymbol{s}\right\|_2^2 \big) - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s} \rangle\Big] + \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}} \qquad (32)$$

Notice that $\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \left\|\boldsymbol{s}\right\|_2^2 = \left\|\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}\right\|_2^2 + 2\langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}, \boldsymbol{s} \rangle$. Integrating over time from $\tau = 0$ to $\tau = t$, we obtain

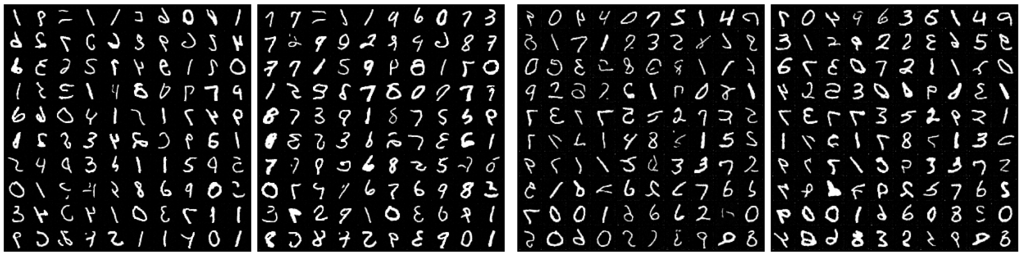
$$\begin{aligned}
\int_0^t \boldsymbol{\epsilon}_{\boldsymbol{s}_{\boldsymbol{\theta}}}(\boldsymbol{x}, \tau) d\tau = \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}(\boldsymbol{x}, t)\big) - \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, 0) - \boldsymbol{s}(\boldsymbol{x}, 0)\big) && (33) \\
- \int_0^t \frac{1}{2} g^2(\tau) \nabla_{\boldsymbol{x}} \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}) d\tau && (34) \\
- \int_0^t g^2(\tau) \Big[\langle \nabla_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}), \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s} \rangle + \langle \nabla_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}), \boldsymbol{s} \rangle + \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}, \nabla_{\boldsymbol{x}} \boldsymbol{s} \rangle\Big] d\tau && (35) \\
+ \int_0^t \Big[\langle \nabla_{\boldsymbol{x}} \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s} \rangle + \langle \boldsymbol{f}, \nabla_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}) \rangle\Big] d\tau && (36)
\end{aligned}$$

By applying the ℓ_2-norm and Cauchy-Schwartz inequality while noting the relation $\|A\|_2 \leq \|A\|_F$ for a general square matrix A, the statement is proved.

■

F Demonstration of generated MNIST examples

In Fig. 6, we show examples generated with different choices of the SDE and the weight of score FPE-regularizer, γ, on MNIST.



(a) VE ($\gamma = 0.0$) (b) VE ($\gamma = 1.0$) (c) VP ($\gamma = 0.0$) (d) VP ($\gamma = 1.0$)

Figure 6: Illustration of generated samples.