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 three implications of satisfying the score FPE and a potential explanation for why the score FPE is not fulfilled 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 Introduction

Score-based generative models (SGM) [Soh+15; SE19; HJA20; Son+20a; Son+20b], also referred to as diffusion models, have led to major advances in the generation of synthetic images [DN21; Sab+22; Rom+22] and in various other downstream applications [Men+21b; Nie+22; Kaw+22; Che+22; Sai+22]. SGMs involve a forward and a backward process. In the forward process (diffusion process) increasing amounts of noise are gradually added to each data point until the original structure is lost, transforming data into pure noise. The backward process attempts to reverse the forward process, using a neural network (called a noise-conditional score model) trained to gradually remove noise, effectively transforming pure noise into clean data samples. The (noise-conditional) score models are trained with a denoising score-matching objective [HD05; Vin11; SE19; HJA20] to estimate the score (gradient of the log-likelihood function) of the data density perturbed with various amounts of noise (as in forward process).

We can interpret the training procedure of diffusion models as jointly estimating the score of the original data density and all its perturbations. Crucially, all these densities are closely related to each other as they correspond to the same data density perturbed with various amounts of noise. With sufficiently small time steps, the forward process is a diffusion [Son+20a] and the spatial-temporal
We mathematically study the implications of satisfying the score FPE. We prove that reducing the ∇

where

f

scores

we derive an associated system of PDEs characterizing the evolution of the scores (i.e., gradients) of the perturbed data densities, which we term score Fokker-Planck equation (score FPE). The ground truth scores of the perturbed data densities must in theory satisfy the score FPE.

We mathematically study the implications of satisfying the score FPE. We prove that reducing the score FPE error: (a) improves log-likelihood of the probability flow ODE diffusion mode [Son+20a], (b) improves the degree of conservativity of the models, and (c) can be achieved by enforcing higher-order score matching [Men+21a; Lu+22]. In practice, we observe that many existing pre-trained score models do not numerically satisfy the score FPE. We therefore propose a new loss function for training diffusion models combining the traditional score matching objective with a regularization term enforcing the score FPE. We show that this enables more accurate density estimation on synthetic data and it can improve the likelihood on MNIST.

2 Background

[Son+20a] unifies denoising score matching [SE19] and diffusion probabilistic models [Soh+15; HJA20] via a stochastic process x(t) of continuous time t ∈ [0, T] driven by the forward SDE

\[ dx(t) = f(x(t), t)dt + g(t)dw_t, \]

where \( f(\cdot,t) : \mathbb{R}^D \rightarrow \mathbb{R}^D \), \( g(\cdot) : \mathbb{R} \rightarrow \mathbb{R} \) and \( w_t \) is a standard Wiener process. Under some moderate conditions \( \text{[And82]} \), one can obtain a reverse time SDE from \( T \) to 0

\[ dx(t) = [f(x(t), t) - g^2(t) \nabla_x \log q_t(x(t))]dt + g(t)dw_t, \]

where \( w_t \) is a standard Wiener process in reverse time. Let \( q_t(x) \) denote the ground truth marginal density of \( x(t) \) following Eq. (1). We can train a time-conditioned neural network \( s_\theta = s_\theta(x, t) \) to approximate \( \nabla_x \log q_t(x) \) by minimizing the score matching objective \( \text{[HD05]} \)

\[ J_{\text{SM}}(\theta; \lambda(\cdot)) := \frac{1}{2} \int_0^T \lambda(t) \mathbb{E}_{x \sim q_t(x)} \left[ \| s_\theta(x, t) - \nabla_x \log q_t(x) \|^2 \right] dt. \]

Since \( q_t(x) \) is generally inaccessible, the denoising score matching (DSM) loss \( \text{[Vin11; Son+20a]} \) is exploited in practice instead:

\[ J_{\text{DSM}}(\theta; \lambda(\cdot)) := \frac{1}{2} \int_0^T \lambda(t) \mathbb{E}_{x(0) \sim \tilde{P}(x)} \mathbb{E}_{q_0(x(0)|x(0))} \left[ \| s_\theta(x(t), t) - \nabla_x \log q_0(x(t)|x(0)) \|^2 \right] dt, \]

where \( q_0(x(t)|x(0)) \) is the transition kernel from \( x(0) \) to \( x(t) \). After \( s_\theta(x, t) \approx \nabla_x \log q_t(x) \) is learned, we replace \( \nabla_x \log q_t(x) \) in Eq. (2) with \( s_\theta \) and get a parametrized reverse-time SDE for stochastic process \( x_\theta(t) \)

\[ dx_\theta(t) = [f(x_\theta(t), t) - g^2(t)s_\theta(x_\theta(t), t)]dt + g(t)dw_t, \]

(4)

Let \( P^{\text{SDE}}_{\theta, 0} \) denote the marginal distribution of \( x_\theta(t) \) with the initial distribution defined as the prior \( \pi \), where suppress the dependency on \( \pi \) for compactness. We can design \( f \) and \( g \) in Eq. (2) so that \( q_T(x) \) approximates a simple prior \( \pi \), and hence, can generate samples \( x_\theta(0) \sim P^{\text{SDE}}_{\theta, 0} \) by numerically solving Eq. (5) backward with an initial sample from the prior \( x_\theta(T) \approx \pi \). Intuitively, \( x_\theta(0) \) should be close to a sample from the data distribution.

[Son+20a] further introduces a deterministic process (with zero diffusion term) describing the evolution of samples whose trajectories share the same marginal probability densities as the forward SDE (Eq. (5)). Specifically, the process evolves through time according to the following probability flow Ordinary Differential Equation (ODE):

\[ \frac{dx}{dt} = f(x, t) - \frac{1}{2}g^2(t)\nabla_x \log q_t(x). \]

As in the SDE case, the ground truth score \( \nabla_x \log q_t(x) \) is approximated with the learned score model \( s_\theta(x, t) \approx \nabla_x \log q_t(x) \) in Eq. (6), leading to the parameterized probability flow ODE

\[ \frac{d\tilde{x}_\theta}{dt} = f(\tilde{x}_\theta, t) - \frac{1}{2}g^2(t)s_\theta(\tilde{x}_\theta, t). \]
We denote the marginal density of \( \tilde{x}_{\theta} \) as \( p_{t,\theta}^{\text{ODE}} \) with initial condition sampled from the prior \( \pi \), where the dependency on \( \pi \) is also omitted in the notation for compactness. By solving Eq. (7) backward with an initial value \( \tilde{x}_{\theta}(T) \sim \pi \) via numerical methods, we can generate a sample \( \tilde{x}_{\theta}(0) \sim p_{0,\theta}^{\text{ODE}} \) to approximate sampling from the data distribution. Indeed, due to the deterministic dynamics in Eq. (7), it is possible to compute exact likelihoods for this generative model. Let \( \tilde{x}_{\theta}(t) \in \mathbb{R}^D \) evolve reversely through time following Eq. (7) starting with \( \tilde{x}_{\theta}(T) \sim \pi \). The “Instantaneous Change of Variables” [Che+18] characterizes the temporal change of \( \log p_{t,\theta}^{\text{ODE}} \) along the trajectory \( \{ \tilde{x}_{\theta}(t) : t \in [0, T] \} \) via an ODE

\[
\frac{d}{dt} \log p_{t,\theta}^{\text{ODE}}(\tilde{x}_{\theta}(t)) = \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}(\tilde{x}_{\theta}(t), t)) - \text{div}_x(f(\tilde{x}_{\theta}(t), t)).
\]

Therefore, the log-likelihood can be exactly calculated by numerically solving the concatenated ODEs backward from \( T \) to 0, initialized with \( \tilde{x}_{\theta}(T) \sim \pi \)

\[
\frac{d}{dt} \left[ \log p_{t,\theta}^{\text{ODE}}(\tilde{x}_{\theta}(t)) \right] = \left[ \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}(\tilde{x}_{\theta}(t), t)) - \text{div}_x(f(\tilde{x}_{\theta}(t), t)) \right].
\]

### 3 The Fokker-Planck equation for score vector fields in diffusions

It is well known that the evolution of the ground truth density \( q_t(x) \) associated to Eq. (1) is governed by the Fokker-Planck equation (FPE) [Øks03] expressed in Eq. (8) (more details in Appx. E).

\[
\partial_t q_t(x) = -\sum_{j=1}^{D} \partial_{x_j} (\tilde{F}_j(x, t) q_t(x)),
\]

where \( \tilde{F}(x, t) := f(x, t) - \frac{1}{2} g^2(t) \nabla_x \log q_t(x) \). 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 \( \nabla_x \log q_t(x) \) 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 \( q_t(x) \) is sufficiently smooth for \( (x, t) \in \mathbb{R}^D \times [0, T] \). Then its score \( s(x, t) := \nabla_x \log q_t(x) \) satisfies the following system of PDEs

\[
\partial_t s - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s) + \frac{1}{2} g^2(t) \| s \|_2^2 - \langle f, s \rangle - \text{div}_x(f) \right] = 0, \quad (x, t) \in \mathbb{R}^D \times [0, T]. \tag{9}
\]

This result shows that the time-conditional scores \( s_{\theta}(x, t) \) learned by score-based models (via Eq. (4)) 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 \( f \equiv 0 \) and \( g \equiv 1 \). That is, \( x(t) \) 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 \( q_t = q_{t_0} \ast \mathcal{N}(0, t) \), and \( q_t \) can be analytically obtained from \( q_{t_0} \) [MR92] (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. We empirically support this idea in Appx. B with synthetic data whose ground truth density has a closed form expression.

Theoretically, with sufficient data and model capacity, (denoising) score matching ensures the optimal solution to Eq. (1) should satisfy Eq. (9) as it approximates the ground truth score well. However, we observe that pre-trained \( s_{\theta} \) learned via Eq. (4) do not numerically satisfy the score FPE. We hereby introduce an error term \( \epsilon_{s_{\theta}} = \epsilon_{s_{\theta}}(x, t) \) in order to quantify how \( s_{\theta} \) deviates from the score FPE

\[
\epsilon_{s_{\theta}}(x, t) := \partial_t s_{\theta} - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}) + \frac{1}{2} g^2(t) \| s_{\theta} \|_2^2 - \langle f, s_{\theta} \rangle - \text{div}_x(f) \right]. \tag{10}
\]

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

\[
\epsilon_{\text{DSM-like}}(t; \theta) := \frac{1}{D} \mathbb{E}_{x(0) \sim \pi}(\epsilon_{x(t)}(x(0)) \| x(0) \|_2)
\]

\[
\epsilon_{\text{FPE}}(t; \theta) := \frac{1}{D} \mathbb{E}_{x \sim \nu}([\epsilon_{s_{\theta}}(x(t), t)]_2), \quad \nu \sim \text{Uniform}(\{0, 1\}^D) \text{ or } \nu \sim q_t(x| x(0)).
\]

Fig. 1 plots these residuals for score models pre-trained via DSM on MNIST and CIFAR-10. Despite achieving low \( \epsilon_{\text{DSM-like}} \) score-matching loss across all \( t \) (green curve), pre-trained score models fail to satisfy the score FPE equation especially for small \( t \) (blue and orange curves).
Assumption A. We assume there are finite constants we prove in Sec. 4.3 the equivalence of conditions hold. Next, we introduce some technical conditions in Assumption A that are commonly used in theoretical likelihood of the data under $p_q$. In this section, we show that simultaneously minimizing the averaged error of the score FPE and score matching objective can reduce the KL divergence between $p_q$ and $p_{\theta,0}^{ODE}$, the data density and the one determined by parametrized probability flow ODE (Eq. (7)), respectively. In Sec. 4.2 we prove that controlling $\epsilon_{sa}$ can implicitly enforce conservativity of $s_\theta$. Moreover, if the score FPE is satisfied, we prove in Sec. 4.3 the equivalence of $s_\theta$, ground truth score $s$ and $\nabla_x \log p_{\theta,0}^{ODE}$ holds under some conditions, where $p_{\theta,0}^{ODE}$ is defined in Sec. 2 as the marginal density of parametrized diffusion process. In Sec. 4.4 we investigate the connection between higher-order score matching \cite{Men+21a, Lu+22} and the score FPE.

4 Theoretical implications and interpretations of score FPE

In this section, we first study three implications of satisfying the score FPE. More precisely, we first show in Sec. 4.1 that simultaneously minimizing the averaged error of the score FPE and score matching objective can reduce the KL divergence between $p_q$ and $p_{\theta,0}^{ODE}$, the data density and the one determined by parametrized probability flow ODE (Eq. (7)), respectively. In Sec. 4.2 we prove that controlling $\epsilon_{sa}$ can implicitly enforce conservativity of $s_\theta$. Moreover, if the score FPE is satisfied, we prove in Sec. 4.3 the equivalence of $s_\theta$, ground truth score $s$ and $\nabla_x \log p_{\theta,0}^{ODE}$ holds under some conditions, where $p_{\theta,0}^{ODE}$ is defined in Sec. 2 as the marginal density of parametrized diffusion process. In Sec. 4.4 we investigate the connection between higher-order score matching \cite{Men+21a, Lu+22} and the score FPE.

4.1 Minimizing $D_{KL}(q_0||p_{\theta,0}^{ODE})$

First of all, we consider the averaged error of the score FPE defined as follows

$$M(\theta) := \sup_{t \in [0,T]} \mathbb{E}_{x \sim q_0(x)} \left[ \int_0^T \| \epsilon_{sa}(x, \tau) \|_2^2 d\tau \right] < \infty. \quad (11)$$

In this section, we show that simultaneously minimizing $M(\theta)$ and $J_{SM}(\theta)$ enables us to decrease the KL divergence between $q_0$ and $p_{\theta,0}^{ODE}$, denoted as $D_{KL}(q_0 || p_{\theta,0}^{ODE})$ and, equivalently, improve the likelihood of the data under $p_{\theta,0}^{ODE}$.

Next, we introduce some technical conditions in Assumption A that are commonly used in theoretical studies of score-based models \cite{Son+21, Lu+22, Pid22}. For notational simplicity, let $s_\theta^{ODE}(\cdot, t)$ denote $\nabla_x \log p_{\theta,0}^{ODE}$, the time-conditional score function of the probability density $p_{\theta,0}^{ODE}$.

Assumption A. We assume there are finite constants $L > 0$ and $\delta_T > 0$ such that the following conditions hold

(a) Bounded 2nd.- non-central moments: $\mathbb{E}_{x \sim q_0(x)}(\|x\|_2^2) \leq L$;

(b) $\|s_\theta(x, t)\|_2 \leq L(1 + \|x\|_2)$, for all $x \in \mathbb{R}^D$ and $t \in [0,T]$;

(c) $\|s_\theta(x, t) - s_\theta(y, t)\|_2 \leq L \|x - y\|_2$, for all $x, y \in \mathbb{R}^D$ and $t \in [0,T]$;

(d) $\|f(x, t)\|_2 \leq L(1 + \|x\|_2)$, for all $x \in \mathbb{R}^D$ and $t \in [0,T]$;

(e) $\|f(x, t) - f(y, t)\|_2 \leq L \|x - y\|_2$, for all $x, y \in \mathbb{R}^D$ and $t \in [0,T]$;

(f) $\|s_\theta^{ODE}(x, t) - s_\theta^{ODE}(y, t)\|_2 \leq L \|x - y\|_2$, for all $x, y \in \mathbb{R}^D$ and $t \in [0,T]$;

(g) $\sup_{t \in [0,T]} \left\{ \mathbb{E}_{x \sim q_0(x)} \left[ \|s_\theta(x, t) - s_\theta^{ODE}(x, t)\|_2^2 \right] \right\} \leq \delta_T^2$, or

$\sup_{x \in \mathbb{R}^D} \|s_\theta(x, t) - s_\theta^{ODE}(x, t)\|_2^2 \leq \delta_T^2$;

(h) For all $t \in [0,T]$, there is a $k > 0$ so that $q_0(x) = \mathcal{O}(e^{-\|x\|_2^k})$ and $p_{\theta,0}^{ODE}(x) = \mathcal{O}(e^{-\|x\|_2^k})$ as $\|x\|_2 \rightarrow \infty$. 

4
Then, we review a crucial equation proposed by [Lu+22] which quantifies the exact gap between the KL divergence $D_{KL}(q_0||p_0^{ODE})$ and the score matching objective $J_{SM}(\theta)$. 

**Lemma 1 ([Lu+22]).** Set $\lambda(t) = g^2(t)$. Let $q_0$ be the data distribution, $q_t$ be the marginal density of $x(t)$ following Eq. (1) and $p_0^{ODE}$ be the marginal density determined by probability flow ODE (Eq. (1)). Assume Assumption [A(h)] is satisfied. Then

$$D_{KL}(q_0||p_0^{ODE}) = D_{KL}(q_T||p_T^{ODE}) + J_{SM}(\theta) + J_{Dff}(\theta),$$

where

$$J_{Dff}(\theta) = \frac{1}{2} \int_0^T g^2(t) E_{x \sim q_t(x)} \left[ (s_\theta(x,t) - \nabla_x \log q_t(x))^T (s_\theta^{ODE}(x,t) - s_\theta(x,t)) \right] dt.$$

We now discuss the main proposition in this section. We notice that applying the Cauchy-Schwarz inequality to $J_{Dff}(\theta)$ leads to an upper bound on $J_{Dff}(\theta)$

$$|J_{Dff}(\theta)| \leq \sqrt{J_{SM}(\theta)} \cdot \sqrt{J_{Fisher}(\theta)},$$

where $J_{Fisher}(\theta)$ is a Fisher-like divergence in terms of the two scores $s_\theta(x,t)$ and $s_\theta^{ODE}(x,t)$, and is defined as

$$J_{Fisher}(\theta) := \frac{1}{2} \int_0^T \left\| s_\theta(x,t) - s_\theta^{ODE}(x,t) \right\|^2 dt. \quad (13)$$

In Theorem [1] we will show that we can further bound $J_{Fisher}(\theta)$ above by a decreasing function in terms of $M(\theta)$:

$$J_{Fisher}(\theta) \lesssim M(\theta) + \sqrt{M(\theta)} + 1. \quad (14)$$

Therefore, Eq. (12) together with Ineq. (13) and (14) imply that we can reduce $D_{KL}(q_0||p_0^{ODE})$ once $M(\theta)$ and $J_{SM}(\theta)$ are minimized simultaneously. We now rigorously state the theorem.

**Theorem 1.** We have

$$\left( J_{Dff}(\theta) \right)^2 \leq J_{SM}(\theta) \cdot J_{Fisher}(\theta). \quad (15)$$

Moreover, if Assumption [A] is fulfilled, then there is a finite constant $C := C(L,T,g,\delta_T) > 0$, depending only on $L$, $T$, $g$, and $\delta_T$ so that we can further bound the Ineq. (15) above as

$$\left( J_{Dff}(\theta) \right)^2 \leq C^2(L,T,g,\delta_T) \cdot J_{SM}(\theta) \left( M(\theta) + \sqrt{M(\theta)} + 1 \right). \quad (16)$$

Hence,

$$D_{KL}(q_0||p_0^{ODE}) \leq D_{KL}(q_T||p_T^{ODE}) + J_{SM}(\theta) + C \cdot \sqrt{J_{SM}(\theta)} \left( M(\theta) + \sqrt{M(\theta)} + 1 \right)^{1/2}. \quad (17)$$

### 4.2 Conservativity

The ground truth score $s(x,t) = \nabla_x \log q_t(x)$ 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 [SH21]. Below we prove that we can implicitly enforce conservativity by minimizing the time-averaged residual $\epsilon_{s(x,t)}$ of the score FPE.

**Proposition 1.** If there is a $t_\theta \in [0,T]$ so that $s_\theta(x,t_\theta) = \nabla_x \log q_{t_\theta}(x)$ for all $x \in \mathbb{R}^D$, then there is a real-valued function $\Psi : \mathbb{R}^D \times [0,T] \rightarrow \mathbb{R}$ given by $\Psi_{t_\theta}(x,t) = \log q_{t_\theta}(x) + \int_{t_\theta}^t \left[ \frac{1}{2}g^2(\tau) \text{div}_x(s_\theta) + \frac{1}{2}g^2(\tau) \|s_\theta\|^2 + (f,s_\theta) - \text{div}_x(f) \right] d\tau$ so that for all $(x,t) \in \mathbb{R}^D \times [0,T]$

$$s_\theta(x,t) - \nabla_x \Psi_{t_\theta}(x,t) = \int_{t_\theta}^t \epsilon_{s_\theta}(x,\tau)d\tau. \quad (17)$$

In particular,

$$\|s_\theta(x,t) - \nabla_x \Psi_{t_\theta}(x,t)\| \leq \int_{\min\{t_\theta,t\}}^{\max\{t_\theta,t\}} \|\epsilon_{s_\theta}(x,\tau)\|_2 d\tau. \quad (18)$$
Eq. (17) indicates the error of the score FPE quantifies the degree of conservativity of \( s_\theta \). We further explain this idea via Ineq. (18). Consider a model \( s_\theta \), and assume that for a large enough timestep \( t_\theta \), it captures exactly the perturbed density \( (s_\theta(x,t_\theta) = \nabla_x \log q_{t_\theta}(x)) \) which is close to the prior (normal distribution) because \( t_\theta \approx T \). Intuitively, \( s_\theta \) 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 a log-density). Indeed, Prop 1 says that the estimated score should nearly be conservative if it approximately satisfies the score FPE, i.e., close to the gradient of a scalar function \( \Psi_\theta(x,t) \).

### 4.3 Equivalence between \( s_\theta \), \( s \), and \( \nabla_x \log p_{t_\theta}^{\text{SDE}} \)

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

**Proposition 2.** Suppose we know that in some suitable function space, \( 0 \) is the unique strong solution to the PDEs \( \partial_t v - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(v) + \frac{1}{2} g^2(t) \|v\|_2^2 + \langle f, v \rangle \right] = 0 \) with zero initial condition \( v(x,0) \equiv 0 \) and zero boundary condition. If there is some \( \theta_0 \) so that \( \epsilon_{s_{\theta_0}}(x,t) = 0 \) for all \((x,t)\), then \( s_{\theta_0} \equiv s \).

Moreover, suppose that the PDEs \( \partial_t v + \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(v) + \frac{1}{2} g^2(t) \|v\|_2^2 + \langle f, v \rangle \right] = 0 \) with zero initial and boundary condition has \( 0 \) as the unique strong solution, then \( \epsilon_{s_{\theta_0}} \equiv 0 \) implies \( s_{\theta_0} \equiv \nabla_x \log p_{t_\theta}^{\text{SDE}} \).

We hypothesize that satisfying the score FPE has a smoothing effect when \( f(x,t) \) is linear in \( x \). Suppose the assumptions of Prop. 2 hold and hence, \( s_{\theta_0} \equiv \nabla_x \log p_{t_\theta}^{\text{SDE}} \). As linearly transforming normal distributions (by \( f \)) remains normal, \([Lu+22]\) proves that \( p_{t_\theta}^{\text{SDE}} \) 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 \( \epsilon_{s_{\theta_0}} \equiv 0 \). In this case, we hypothesize that learning \( \theta \) to reduce \( \| \epsilon_{s_{\theta_0}} \| \) can reduce the gap \( \|s_{\theta} - \nabla_x \log p_{t_\theta}^{\text{SDE}}\| \), and may further modify the direction \( s_{\theta} \) toward high density region of Gaussian (smoothing effect).

### 4.4 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 \( \| \int_0^t \epsilon_{s_{\theta}}(x,\tau)d\tau \|_2 \) for all \( t \in [0,T] \). This may explain why the scores learned via Eq. (4) 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

\[
\|s - s_\theta\|_2 \leq \delta_0, \quad \sup_{R^D \times [0,T]} \|\nabla_x (s - s_\theta)\|_F \leq \delta_1, \quad \sup_{R^D \times [0,T]} \|\nabla_x \text{div}_x(s - s_\theta)\|_2 \leq \delta_2.
\]

Then for all \((x,t) \in R^D \times [0, T] \) we have

\[
\left\| \int_0^t \epsilon_{s_{\theta}}(x,\tau)d\tau \right\|_2 \leq 2\delta_0 + (\delta_2 + 2\delta_1\delta_0)C(t) + \delta_1 \int_0^t (g^2(\tau) \|s(x,\tau)\|_2 + \|f(x,\tau)\|_2) d\tau + \delta_0 \int_0^t (g^2(\tau) \|\nabla_x s(x,\tau)\|_F + \|\nabla_x f(x,\tau)\|_F) d\tau.
\]

### 5 Score FPE regularizer and techniques for efficient computation

We have demonstrated that score models learned via \( J_{\text{DSM}} \) (Eq. (4)) 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 \( J_{\text{DSM}} \) and a score FPE-regularizer

\[
R_{\text{FP}}(\theta) := \frac{1}{D} \mathbb{E}_{t \sim [0,T]} \mathbb{E}_{x \sim \nu} \|\epsilon_{s_{\theta}}(x, t)\|_2
\]
defined as
\[
\mathcal{J}_{FP}(\theta; \lambda(\cdot), \gamma) := \mathcal{J}_{DSM}(\theta; \lambda(\cdot)) + \gamma \mathcal{R}_{FP}(\theta),
\]
where \( \gamma \geq 0 \) is a hyper-parameter controlling to what extent we desire the score FPE to be satisfied.

Since \( \epsilon_{s\theta} \) in \( \mathcal{R}_{FP} \) is generally expensive to calculate for high dimensional data, we propose to exploit the finite difference method [For88] to approximate \( \partial_t s\theta \) in Sec. 5.1 and Hutchinson’s trace estimator [Hut89] for \( \text{div}_x(s\theta) \) in Sec. 5.2. In Appx. C, we discuss an additional technique based on projecting \( \epsilon_{s\theta} \) onto a random direction, thereby transforming the gradient computation in \( \epsilon_{s\theta} \) into the computation of a one-dimensional derivative.

5.1 Techniques to reduce computational cost of the \( \partial_t s\theta \) term

Typically, \( \partial_t s\theta \) is computed via automatic differentiation. However, it can be efficiently approximated by finite differences as the derivative is one-dimensional. We first review the one-dimensional finite difference method and summarize its estimation error in the following lemma.

**Lemma 2.** [For88] Let \( \alpha : [0, 1] \to \mathbb{R}^D \) 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 \( \alpha'(t) \):
\[
\frac{h_s^2 \alpha(t + h_d) + (h_d^2 - h_s^2) \alpha(t) - h_s^2 \alpha(t - h_s)}{h_s h_d (h_s + h_d)} + \mathcal{O}(\frac{h_s h_d^2 + h_d h_s^2}{h_s + h_d}).
\]
In particular, if \( h_s = h_d := h \), then the estimate becomes
\[
\frac{\alpha(t + h) - \alpha(t - h)}{2h} + \mathcal{O}(h^2).
\]
In the implementation for dataset with high dimensions, we consider \( \alpha(\cdot) := s\theta(\cdot, x) \) and hence, \( \partial_t s\theta(t, x) \) is approximated with
\[
\frac{h_s^2 s\theta(t + h_d, x) + (h_d^2 - h_s^2) s\theta(t, x) - h_s^2 s\theta(t - h_s, x)}{h_s h_d (h_s + h_d)}.
\]

5.2 Techniques to reduce the computational cost of the \( \text{div}_x(s\theta) \) term

Hutchinson’s trace estimator [Hut89] stochastically estimates the trace \( \text{tr}(A) \) of any square matrix \( A \). Its idea is choose a distribution \( p_v \) so that \( \mathbb{E}_{v \sim p_v}[v] = 0 \) and \( \mathbb{E}_{v \sim p_v}[vv^T] = I \). Hence,
\[
\text{tr}(A) = \text{tr}(A v \sim p_v[vv^T]) = \mathbb{E}_{v \sim p_v}[\text{tr}(Avv^T)] = \mathbb{E}_{v \sim p_v}[\text{tr}(vAv^T)] = \mathbb{E}_{v \sim p_v}[vAv^T].
\]
By i.i.d. sampling \( \{v_j\}_{j=1}^M \) from \( p_v \), we can use the following unbiased estimator
\[
\frac{1}{M} \sum_{j=1}^M v_j A v_j^T
\]
to estimate \( \text{tr}(A) \). We notice that \( \text{div}_x(s\theta(x, t)) = \text{tr}(\nabla_x s\theta) \). Thus, we can apply Hutchinson’s trick and replace \( \text{div}_x(s\theta) \) term with the estimation
\[
\frac{1}{M} \sum_{j=1}^M v_j \nabla_x s\theta(x, t) v_j^T.
\]
In the implementation, \( p_v \) is usually taken as a standard normal distribution or a Rademacher distribution. We follow the convention in [Son+20b] which sets \( M = 1 \) and shows its effectiveness in practice.

6 Experimental results

The effectiveness of \( \mathcal{J}_{FP} \) 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 [Son+20b] 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 $J_{FP}(\theta; \lambda(\cdot), \gamma)$ on MNIST with different $\gamma$’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 ([Son+20a] and [Son+21]). We observe a general improvement in NLL with $\gamma = 1.0$. In Fig. 3, we show examples generated with different choices of the SDE and the weight of score FPE-regularizer, $\gamma$, on MNIST.

### Table 1: NLL (in bpd) for different FP weights $\gamma$’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 + FP  | [Son+20a]        | 3.86           | 3.63           | 3.66           | 3.28           | 3.37           |
|          | [Son+21]         | 3.63           | 3.94           | 3.53           | **3.20**       | 3.23           |
| VP + FP  | [Son+20a]        | 2.95           | 3.06           | 3.09           | **2.91**       | 3.34           |
|          | [Son+21]         | 3.11           | 3.14           | **3.04**       | 3.28           | 3.28           |
| RVE + FP | [Son+20a]        | 3.45           | 3.68           | 3.77           | 3.57           | **3.13**       |
|          | [Son+21]         | 3.62           | 3.78           | 3.49           | **3.16**       | 3.36           |

![Figure 3: Illustration of generated samples.](image)

7 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

[And82] Brian DO Anderson. “Reverse-time diffusion equation models”. In: Stochastic Processes and their Applications 12.3 (1982), pp. 313–326.

[Che+18] Ricky TQ Chen et al. “Neural ordinary differential equations”. In: Advances in neural information processing systems 31 (2018).

[Che+22] Kin Wai Cheuk et al. “DiffRoll: Diffusion-based Generative Music Transcription with Unsupervised Pretraining Capability”. In: arXiv preprint arXiv:2210.05148 (2022).

[DN21] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: Advances in Neural Information Processing Systems 34 (2021), pp. 8780–8794.

[DVK21] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. “Score-based generative modeling with critically-damped langevin diffusion”. In: arXiv preprint arXiv:2112.07068 (2021).

[EG18] Lawrence C Evans and Ronald F Garzepy. Measure theory and fine properties of functions. Routledge, 2018.

[For88] Bengt Fornberg. “Generation of finite difference formulas on arbitrarily spaced grids”. In: Mathematics of computation 51.184 (1988), pp. 699–706.

[Gro19] Thomas Hakon Gronwall. “Note on the derivatives with respect to a parameter of the solutions of a system of differential equations”. In: Annals of Mathematics (1919), pp. 292–296.

[HD05] Aapo Hyvärinen and Peter Dayan. “Estimation of non-normalized statistical models by score matching.” In: Journal of Machine Learning Research 6.4 (2005).

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: Advances in Neural Information Processing Systems 33 (2020), pp. 6840–6851.

[Hut89] 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.

[Kaw+22] Bahjat Kawar et al. “Denoising diffusion restoration models”. In: arXiv preprint arXiv:2201.11793 (2022).

[Kim+22] 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.

[Kon+20] Zhifeng Kong et al. “Diffwave: A versatile diffusion model for audio synthesis”. In: arXiv preprint arXiv:2009.09761 (2020).

[Lu+22] 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.

[Men+21a] 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.

[Men+21b] Chenlin Meng et al. “Sdedit: Image synthesis and editing with stochastic differential equations”. In: arXiv preprint arXiv:2108.01073 (2021).

[MR92] Elias Masry and John A Rice. “Gaussian deconvolution via differentiation”. In: Canadian Journal of Statistics 20.1 (1992), pp. 9–21.

[Nie+22] Weili Nie et al. “Diffusion Models for Adversarial Purification”. In: arXiv preprint arXiv:2205.07460 (2022).

[Øks03] Bernt Øksendal. “Stochastic differential equations”. In: Stochastic differential equations. Springer, 2003, pp. 65–84.

[Pid22] Jakiw Pidstrigach. “Score-Based Generative Models Detect Manifolds”. In: arXiv preprint arXiv:2206.01018 (2022).

[Rom+22] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022, pp. 10684–10695.

[RPK19] 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.

[Sah+22] Chitwan Saharia et al. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: arXiv preprint arXiv:2205.11487 (2022).
[Sai+22] Koichi Saito et al. *Unsupervised vocal dereverberation with diffusion-based generative models*. 2022. DOI: [10.48550/ARXIV.2211.04124](https://arxiv.org/abs/2211.04124).

[SE19] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[SH21] Tim Salimans and Jonathan Ho. “Should EBMs model the energy or the score?” In: *Energy Based Models Workshop-ICLR 2021*. 2021.

[Soh+15] Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.

[Son+20a] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *arXiv preprint arXiv:2011.13456* (2020).

[Son+20b] 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.

[Son+21] Yang Song et al. “Maximum likelihood training of score-based diffusion models”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1415–1428.

[Vin11] 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

[Son+20a] 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 Variance Explosion (VE) SDE and Variance Preserving (VP) SDE.

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

\[
dx(t) = \sqrt{\frac{d\sigma^2(t)}{dt}} dw_t.
\]

A typical instance of VE SDE is Score Matching of Langevin dynamics (SMLD) [SET19], where \( \sigma(t) := \sigma_{\min}(\frac{\sigma_{\min}}{\sigma_{\max}})^t \) for \( t \in (0, 1) \). In our implementation, we follow the conventional setup of \( \sigma_{\min}, \sigma_{\max} \) := (0.01, 50).

In [Kim+22], they proposed a variant of VE SDE attempting to resolve the unbounded score problem [DVK21], which is called Reciprocal VE (RVE). Let \( \epsilon > 0 \) be a fixed constant. RVE SDE also has zero drift term but with a different parametrization for diffusion

\[
g(t) := \begin{cases} \sigma_{\max}(\frac{\sigma_{\max}}{\sigma_{\min}})^{t} \sqrt{2\epsilon \log(\frac{\sigma_{\min}}{\sigma_{\max}}) \frac{1}{t}}, & \text{if } t > 0 \\ \sigma_{\min}, & \text{if } t = 0 \\
\end{cases}
\]

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

\[
dx(t) = -\frac{1}{2} \beta(t) \mathbf{x}(t) dt + \sqrt{\beta(t)} dw_t.
\]

A classic example of VP SDE is Denoising Diffusion Probabilistic Modeling (DDPM) [Soh+15; HJA20], where \( \beta(t) := \beta_{\min} + t(\beta_{\max} - \beta_{\min}) \) for \( t \in [0, 1] \). We adopt the common setup of \( \beta_{\min}, \beta_{\max} = (0.1, 20) \) in our implementation.

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

|               | VE SDE       | RVE SDE          | VP SDE                  |
|---------------|--------------|------------------|-------------------------|
| \( f(\mathbf{x}, t) \) | 0            | \( -\frac{1}{2} \beta(t) \mathbf{x} \) |                        |
| \( g(t) \)   | \( \sigma_{\max}(\frac{\sigma_{\max}}{\sigma_{\min}})^{t} \sqrt{2\epsilon \log(\frac{\sigma_{\min}}{\sigma_{\max}}) \frac{1}{t}} \), \( t > 0 \) | \( \sigma_{\min} \), \( t = 0 \) | \( \sqrt{\beta(t)} \) |
| SDE           | \( dx(t) = g(t) dw_t \) | \( dx(t) = -\frac{1}{2} \beta(t) \mathbf{x}(t) dt + \sqrt{\beta(t)} dw_t \) |                        |
| score FPE     | \( \partial_t s = \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s) + \frac{1}{2} g^2(t) \|s\|^2_T \right] \) | \( \partial_t s = \frac{1}{2} \beta(t) \nabla_x \left[ \text{div}_x(s) + \|s\|^2_T + \langle x, s \rangle \right] \) |                        |

B How scores satisfy score FPE?

In this section, 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 \( \frac{1}{2} \mathcal{N}((-5, -5), I) + \frac{1}{2} \mathcal{N}((5, 5), I) \) on \( \mathbb{R}^2 \) whose samples are illustrated in Fig. 4a. The diffusion process is taken as VE SDE (Eq. (20)). The ground truth score of GMM, denoted as \( s^{\text{GMM}} \), can be expressed explicitly with a closed formula throughout the diffusion (as the diffusion process is linear in \( \mathbf{x} \)).

First of all, we examine if \( s^{\text{GMM}} \) satisfies the score FPE by computing \( \gamma_{\text{FPE}}(t; s^{\text{GMM}}) \) and plot it in Fig. 6a (blue curve). We can see that the score FPE residual of the ground truth is almost zero, which empirically supports Corollary 5.
Second, as we explain that 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 \( \tilde{s} \) to the following initial value problem of system of PDEs, we know a score at all time.

\[
\begin{align*}
\frac{\partial_t \tilde{s}}{} &= \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (\tilde{s}) + \frac{1}{2} g^2(t) \| \tilde{s} \|_2^2 \right], \\
\tilde{s}(x, 0) &= s_{\text{GMM}}^{\theta}(x, 0), \quad x \in \mathbb{R}^D.
\end{align*}
\]  

(21)

Solutions of Eq. (21) can be parametrized as neural network \( s_{\text{GMM}}^{\theta} \) \cite{RPK19}. We then can solve the PDEs by learning parameters \( \theta \) to reduce both the residuals of the initial condition \( s_{\text{GMM}}^{\theta}(x, 0) - s_{\text{GMM}}(x, 0) \) and evolution \( \epsilon_{s_{\text{GMM}}^{\theta}} := \frac{\partial_t \tilde{s} - \text{div}_x (\tilde{s}) + \frac{1}{2} g^2(t) \| \tilde{s} \|_2^2}. \) That is, \( s_{\text{GMM}}^{\theta} \) is learned with the score FPE-guided objective function:

\[
\min_{\theta} \left\{ E_{t \sim U[0, T]} E_{x \sim q_0} \left( \epsilon_{s_{\text{GMM}}^{\theta}}(x, t) \right)^2 + E_{x \sim q_0} \left( s_{\text{GMM}}^{\theta}(x, 0) - s_{\text{GMM}}(x, 0) \right)^2 \right\}.
\]

(22)

We demonstrate generated samples by the learnt \( s_{\text{GMM}}^{\theta} \) in Fig. 4 and plot its score FPE residual \( \epsilon_{s_{\text{GMM}}^{\theta}}(t; s_{\text{GMM}}^{\theta}) \) in Fig. 6 (orange curve). Interestingly, it also generates quite satisfactory samples and Fig. 5 show it estimates the ground truth score (Fig. 5) well. This supports our argument.

![Figure 4](image1.png)
(a) Ground truth data  
(b) Samples generated by \( s_{\text{GMM}}^{\theta} \)  
(c) Samples generated by \( s_{\text{GMM}}^{\theta} \)

Figure 4: Comparison of instances generated using the score functions learned by our score FPE-guided objective function (Eq. (22)) and the conventional denoising score matching (Eq. (4)), which are denoted as \( s_{\text{GMM}}^{\theta} \) and \( s_{\text{GMM}}^{\theta} \), respectively. Both scores can synthesize reasonable quality samples.

![Figure 5](image2.png)
(a) Ground truth score at \( t = 0 \)  
(b) Estimated Score \( s_{\text{GMM}}^{\theta}(\cdot, 0) \)  
(c) Estimated Score \( s_{\text{GMM}}^{\theta}(\cdot, 0) \)

Figure 5: The fluid flow graph of ground truth score and estimated scores at \( t = 0 \) by \( s_{\text{GMM}}^{\theta} \) and \( s_{\text{GMM}}^{\theta} \). The score \( s_{\text{GMM}}^{\theta} \), which is learned from the score FPE-guided objective, can also approximate the ground truth well.

In addition, we compute the residual of score FPE of a score \( s_{\text{GMM}}^{\theta} \) learned from the denoising score matching (Eq. (4)), which is plotted in Fig. 6. We observe that \( s_{\text{GMM}}^{\theta} \) also does not satisfy score FPE even though it works decently on generation (Fig. 4) and score estimation (Fig. 5).

### C A further technique to reduce computation costs for score FPE

As we explained in Sec. 5, the computation of \( \epsilon_{s_{\text{GMM}}}^{\theta}(x, t) \) in \( R_{\text{FP}}(\theta) \) is generally expensive and we propose two techniques, including finite difference trick and Hutchinson’s trace estimator, to replace expensive computations of some components in \( \epsilon_{s_{\text{GMM}}}^{\theta}(x, t) \).
For any \( v \) where \( p \) which is learned from denoising score matching, does not satisfy score FPE.

(a) FP residuals of ground truth score and the score learned from denoising score matching, does not satisfy score FPE.

(b) FP residuals of the score learned from Eq. (4)

Figure 6: Comparison of the score FPE residuals of \( s_{GMM}^{\theta}, \tilde{s}_{GMM}^{\theta}, \theta \) and \( \hat{s}_{GMM}^{\theta} \) of GMM. 6a shows that both the ground truth score (with closed form) \( s_{GMM}^{\theta} \) and the score \( \hat{s}_{GMM}^{\theta} \) obtained by solving score FPE (Eq. (21)) numerically satisfy score FPE. In contrast, 6b provides a further evidence that \( \hat{s}_{GMM}^{\theta} \), which is learned from denoising score matching, does not satisfy score FPE.

In this section, we further propose another potential trick to reduce the computation cost of differentiation. We recall that

\[
\epsilon_{s\theta}(x, t) = \partial_t s\theta - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s\theta) + \frac{1}{2} g^2(t) \|s\theta\|_2^2 - \langle f, s\theta \rangle - \text{div}_x(f) \right] \tag{23}
\]

Using automatic differentiation to compute the gradient in \( \epsilon_{s\theta}(x, t) \) (the (II) part in Eq. (23)) 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 Sec. 5.1 to reduce the computation efforts. We first recall a fundamental property before rigorously formulating the technique.

**Lemma 3.** Let \( M :\ M(x, t) : \mathbb{R}^D \times [0, T] \rightarrow \mathbb{R} \) be a continuously differentiable function of \( x \). For any \( v \in \mathbb{R}^D \),

\[
D_v M(x, t) = \langle \nabla_x M(x, t), v \rangle,
\]

where \( D_v M(x, t) \) means the directional derivative of \( M \) in \( x \) along the direction \( v \) which is defined as:

\[
D_v M(x, t) := \lim_{h \to 0} \frac{M(x + hv, t) - M(x, t)}{h} = \frac{d}{dh} M(x + hv, t) \big|_{h=0}.
\]

For simplicity, let us denote \( M(x, t) := \frac{1}{2} g^2(t) \text{div}_x(s\theta) + \frac{1}{2} g^2(t) \|s\theta\|_2^2 - \langle f, s\theta \rangle - \text{div}_x(f) \) and let \( v \in \mathbb{R}^D \) be arbitrary vector. We project \( \epsilon_{s\theta}(x, t) \) along direction \( v \) and apply the Lemma 3

\[
\langle \epsilon_{s\theta}(x, t), v \rangle = \langle \partial_t s\theta - \nabla_x M(x, t), v \rangle = \langle \partial_t s\theta, v \rangle - \langle \frac{d}{dh} M(x + hv, t) \big|_{h=0}, v \rangle.
\]

Notice that both \( \partial_t s\theta \) and \( \frac{d}{dh} M(x + hv, t) \big|_{h=0} \) are one-dimensional differentiation, which can be estimated via Lemma 2 and hence, we can avoid automatic differentiation. We hereafter propose an estimated score FP regularizer which may replace \( \mathcal{R}_{FP} \) with

\[
\hat{\mathcal{R}}_{FP}(\theta) := \frac{1}{D} E_{t \sim [0, T]} E_{x \sim p_x} E_{v \sim p_v} \|\epsilon_{s\theta}(x, t), v \|,
\]

where \( p_v \) is a distribution of random vector \( v \in \mathbb{R}^D \). We observe that the performance may degrade by using \( \hat{\mathcal{R}}_{FP} \), 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.
D 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. [2]). The network structure in Sec. [2] 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 [Son+20a] 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 [Lu+22] We found out setting the weight of score FPE to $\lambda = 0.001$ can generally work well for the toy dataset.

E Proofs and discussion

E.1 Proof of Corollary [1]

We prove the result with a more general forward SDE

$$dx = F(x, t)dt + G(x, t)d\omega_t,$$

where $F(\cdot, t): \mathbb{R}^D \rightarrow \mathbb{R}^D$ and $G(\cdot, t): \mathbb{R}^D \rightarrow \mathbb{R}^{D \times D}$.

We know that the density $q_t(x)$ satisfies the Fokker-Plank equation [Oks03]

$$\partial_t q_t(x) = -\sum_{j=1}^{D} \partial_{x_j} (\hat{F}_j(x, t)q_t(x)),$$  

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

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

$$\partial_t \log q_t(x) = \frac{1}{q_t(x)} \partial_t q_t(x)$$

$$= -\frac{1}{q_t(x)} \sum_{j=1}^{D} \partial_{x_j} (\hat{F}_j(x, t)q_t(x))$$

$$= -\frac{1}{q_t(x)} \sum_{j=1}^{D} (\partial_{x_j} \hat{F}_j(x, t)q_t(x) + \hat{F}_j(x, t)\partial_{x_j} q_t(x))$$

$$= -\sum_{j=1}^{D} (\partial_{x_j} \hat{F}_j(x, t) + \hat{F}_j(x, t)\partial_{x_j} \log q_t(x))$$

$$= - (\text{div}_x(\hat{F}) + \langle \hat{F}, s \rangle)$$

$$= - \left[ \text{div}_x(Bs) + \langle Bs, s \rangle + \langle A, s \rangle + \text{div}_x(A) \right]$$

$$= \frac{1}{2} \text{div}_x(GG^Ts) + \frac{1}{2} \|G^Ts\|^2 - \langle A, s \rangle - \text{div}_x(A).$$

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

---

[1] https://colab.research.google.com/drive/120kYB0Va1i0TD85Rj1EfFjaWdxSFU3?usp=s haring
[2] https://github.com/yang-song/score_sde_pytorch
[3] https://github.com/LuChengTHU/mle_score_ode
Remark 1. In Eq. (1) where \( G \) does not depend on \( x \), namely \( G(x, t) \equiv g(t)I \), then \( \tilde{F}(x, t) = f(x, t) - \frac{1}{2}g^2(t)\nabla_x \log q_t(x) \) and

\[
\partial_t \log q_t(x) = \frac{1}{2}g^2(t)\text{div}_x(s) + \frac{1}{2}g^2(t)\|s\|_2^2 - \langle f, s \rangle - \text{div}_x(f)
\]

\[
\partial_t s = \nabla_x \left[ \frac{1}{2}g^2(t)\text{div}_x(s) + \frac{1}{2}g^2(t)\|s\|_2^2 - \langle f, s \rangle - \text{div}_x(f) \right].
\]

E.2 Proof of Proposition 1

Integrating the following equation w.r.t. time from \( \tau = t_0 \) to \( \tau = t \) with \( t \in [0, T] \) fixed,

\[
\partial_t s_\theta = \nabla_x \left\{ \int_{t_0}^{t} \left[ \frac{1}{2}g^2(t)\text{div}_x(s_\theta) + \frac{1}{2}g^2(t)\|s_\theta\|_2^2 - \langle f, s_\theta \rangle - \text{div}_x(f) \right] \, d\tau \right\}
\]

leads to

\[
s_\theta(x, t) - s_\theta(x, t_0) = \nabla_x \left\{ \int_{t_0}^{t} \left[ \frac{1}{2}g^2(t)\text{div}_x(s_\theta) + \frac{1}{2}g^2(t)\|s_\theta\|_2^2 - \langle f, s_\theta \rangle - \text{div}_x(f) \right] \, d\tau \right\}
\]

\[
+ \int_{t_0}^{t} \epsilon_{s_\theta}(x, \tau) \, d\tau,
\]

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

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

\[
s_\theta(x, t) - \nabla_x \left\{ \log q_{t_0}(x) + \int_{t_0}^{t} \left[ \frac{1}{2}g^2(t)\text{div}_x(s_\theta) + \frac{1}{2}g^2(t)\|s_\theta\|_2^2 - \langle f, s_\theta \rangle - \text{div}_x(f) \right] \, d\tau \right\}
\]

\[
= \int_{t_0}^{t} \epsilon_{s_\theta}(x, \tau) \, d\tau.
\]

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

\[
\|s_\theta(x, t) - \nabla_x \Psi_\theta(x, t)\|_2 = \left\| \int_{t_0}^{t} \epsilon_{s_\theta}(x, \tau) \, d\tau \right\|_2.
\]

From which we obtain

\[
\|s_\theta(x, t) - \nabla_x \Psi_\theta(x, t)\|_2 = \left\| \int_{t_0}^{t} \epsilon_{s_\theta}(x, \tau) \, d\tau \right\|_2 \leq \int_{t_0}^{t} \left\| \epsilon_{s_\theta}(x, \tau) \right\|_2 \, d\tau.
\]

The upper and lower bound of integral can be respectively written as \( \max\{t, t_0\} \) and \( \min\{t, t_0\} \), and whence, the proposition is proved.

E.3 Proof of Proposition 2

Lemma 4. Let \( s_\theta \) be a score obtained from denoising score matching (Eq. (4)) and write \( s_\theta^{\text{SDE}}(\cdot, t) := \nabla_x \log p_t^{\text{SDE}} \). Then

1. [Lu+22] Eq. (5) associates with the following forward SDE whose marginal density is \( s_\theta^{\text{SDE}} \):

\[
dx_\theta(t) = \left[ f(x_\theta(t), t) + g^2(t)(s_\theta^{\text{SDE}}(x_\theta(t), t) - s_\theta(x_\theta(t), t)) \right] \, dt + g(t)w_t
\]

2. \( s_\theta^{\text{SDE}} \) satisfies the following score FPE:

\[
\partial_t s_\theta^{\text{SDE}} - \nabla_x \left[ \frac{1}{2}g^2(t)\text{div}_x(2s_\theta - s_\theta^{\text{SDE}}) + \frac{1}{2}g^2(t)(2s_\theta - s_\theta^{\text{SDE}}) - \|s_\theta^{\text{SDE}}\|_2^2 \right] - \langle f, s_\theta^{\text{SDE}} \rangle - \text{div}_x(f) = 0.
\]

Proof of Lemma 4. Consider

\[
F(x, t) := f(x, t) + g^2(t)(s_\theta^{\text{SDE}} - s_\theta) \quad \text{and} \quad G(x, t) := g(t)I
\]

in Eq. (24), and apply Corollary 1, the lemma is then established.
Proof of Proposition 2 \[ \text{We recall Eq. (10), which indicates} \]
\[
\partial_t s_\theta - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (s_\theta) + \frac{1}{2} g^2(t) \left\| s_\theta \right\|_2^2 - \left( \langle f, s_\theta \rangle - \text{div}_x (f) \right) \right] - \epsilon_{s_\theta} = 0. \tag{26} \]

First, we subtract Eq. (2) by the above equation and get
\[
\partial_t (s_\theta - s_\theta) - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (s_\theta - s_\theta) - \frac{1}{2} g^2(t) \left\| s_\theta - s_\theta \right\|_2^2 - \left( \langle f, s_\theta - s_\theta \rangle \right) \right] + \epsilon_{s_\theta} = 0. \tag{27} \]

Consider when \( \theta = \theta_0 \) and let \( u_{\theta_0} := s_\theta^{\text{SDE}} - s_\theta \). Then the PDEs become
\[
\partial_t u_{\theta_0} + \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (u_{\theta_0}) + \frac{1}{2} g^2(t) \left\| u_{\theta_0} \right\|_2^2 + \langle f, u_{\theta_0} \rangle \right] = 0.
\]
Here, \( u_{\theta_0} \) 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 \( s_\theta \) and \( s_\theta^{\text{SDE}} \) share the same initial/boundary condition. Thus, from the assumption of the uniqueness of solution, we know that \( u_{\theta_0} \equiv 0 \), and hence, \( s_\theta^{\text{SDE}} \equiv s_\theta \).

We repeat the same trick to subtract Eq. (9) by Eq. (26) from which we can obtain \( s_{\theta_0} = s \).

E.4 Proof of Proposition 3

By subtracting the following two equations
\[
\partial_t s_\theta = \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (s_\theta) + \frac{1}{2} g^2(t) \left\| s_\theta \right\|_2^2 - \langle f, s_\theta \rangle - \text{div}_x (f) \right] + \epsilon_{s_\theta},
\]
\[
\partial_t s = \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (s) + \frac{1}{2} g^2(t) \left\| s \right\|_2^2 - \langle f, s \rangle - \text{div}_x (f) \right],
\]
we obtain
\[
\partial_t (s_\theta - s) = \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x (s_\theta - s) + \frac{1}{2} g^2(t) \left( \left\| s_\theta \right\|_2^2 - \left\| s \right\|_2^2 - \langle f, s_\theta - s \rangle \right) \right] + \epsilon_{s_\theta} \tag{28}.
\]
Notice that \( \left\| s_\theta \right\|_2^2 - \left\| s \right\|_2^2 = \left\| s_\theta - s \right\|_2^2 + 2\langle s_\theta - s, s \rangle \). Integrating over time from \( \tau = 0 \) to \( \tau = t \), we obtain
\[
\int_0^t \epsilon_{s_\theta}(x, \tau) d\tau = (s_\theta(x, t) - s(x, t)) - (s_\theta(x, 0) - s(x, 0))
\]
\[
- \int_0^t \frac{1}{2} g^2(\tau) \nabla_x \text{div}_x (s_\theta - s) d\tau
\]
\[
- \int_0^t g^2(\tau) \left( \langle \nabla_x (s_\theta - s), s_\theta - s \rangle + \langle \nabla_x (s_\theta - s), s \rangle + \langle s_\theta - s, \nabla_x s \rangle \right) d\tau
\]
\[
+ \int_0^t \left( \langle \nabla_x f, s_\theta - s \rangle + \langle f, \nabla_x (s_\theta - s) \rangle \right) d\tau
\]
By applying the \( \ell_2 \)-norm and Cauchy-Schwartz inequality while noting the relation \( \| A \|_2 \leq \| A \|_F \) for a general square matrix \( A \), the statement is proved.

E.5 Proof of Theorem 1

Lemma 5 (Grönwall’s inequality [Gro19]). Assume that \( \alpha, \beta, \) and \( u \) are continuous functions on \( [0, T] \). If \( \beta \) is non-negative on \( [0, T] \) and \( u \) satisfies the integral inequality
\[
u(t) \leq \alpha(t) + \int_t^T \beta(\tau) u(\tau) d\tau, \quad \text{for all } t \in [0, T] \]
then
\[
u(t) \leq \alpha(t) + \int_t^T \alpha(\tau) \beta(\tau) \exp \left( \int_t^\tau \beta(\tau) d\tau \right) d\tau, \quad \text{for all } t \in [0, T] \]
In particular, if $\alpha$ is non-decreasing (especially, a constant independent of $t$), then
\[
u(t) \leq \alpha(t) \exp \left( \int_t^T \beta(\tau) d\tau \right), \quad \text{for all } t \in [0, T].
\]

**Proof of Grönwall’s inequality.** Consider the function
\[
v(\tau) := \exp \left( - \int_\tau^T \beta(r) dr \right) \int_\tau^T \beta(r) u(r) dr.
\]
Taking the derivative by the product rule leads to
\[
v'(\tau) = \left( - u(\tau) + \int_\tau^T \beta(r) u(r) dr \right) \beta(\tau) \exp \left( - \int_\tau^T \beta(r) dr \right)
\geq - \alpha(\tau) \beta(\tau) \exp \left( - \int_\tau^T \beta(r) dr \right).
\]
Integrating the above inequality from $\tau = t$ to $\tau = T$ proves the statement.

**Proof of Theorem 1.** We first prove the Ineq. (15). Notice that we can rearrange $\mathcal{J}_{\text{Diff}}$ as
\[
\mathcal{J}_{\text{Diff}}(\theta) = \frac{1}{2} \int_0^T g^2(t) \mathbb{E}_{x \sim q_t(x)} \left[ (s_{\theta}(x, t) - \nabla_x \log q_t(x))^T (s_{\theta}(x, t) - s_{\theta}(x, t)) \right] dt
= \int_0^T \int_{\mathbb{R}^D} \left[ g(t) \sqrt{\frac{q_t(x)}{2}} (s_{\theta}(x, t) - \nabla_x \log q_t(x)) \right]^T \left[ g(t) \sqrt{\frac{q_t(x)}{2}} (s_{\theta}(x, t) - s_{\theta}(x, t)) \right] dtdx.
\]
The claim is established by applying Cauchy-Schwartz inequality to functions $g(t) \sqrt{\frac{q_t(x)}{2}} (s_{\theta}(x, t) - \nabla_x \log q_t(x))$ and $g(t) \sqrt{\frac{q_t(x)}{2}} (s_{\theta}(x, t) - s_{\theta}(x, t))$.

We now prove the Ineq. (16), in which we just need to consider the case when $M(\theta) := \sup_{t \in [0, T]} \mathbb{E}_x \| \epsilon_{s_{\theta}}(x, t) \|_2^2 dr < \infty$. Otherwise, the result holds obviously. Recall that the probability flow ODE $\text{ODE}$ associated to Eq. (5) is defined as
\[
\frac{dx(t)}{dt} = f(x(t), t) - \frac{1}{2} g^2(t) s_{\theta}(x(t), t).
\]
By the special case of FPE (Eq. (25)) with zero drift term, we obtain the PDE characterizes the evolution of $p_{t, \theta}^{\text{ODE}}$
\[
\frac{\partial p_{t, \theta}^{\text{ODE}}}{\partial t} = \text{div}_x \left( \frac{1}{2} g^2(t) s_{\theta}(x, t) - f(x, t) p_{t, \theta}^{\text{ODE}}(x) \right)
\]
Hence,
\[
\frac{\partial \log p_{t, \theta}^{\text{ODE}}}{\partial t} = \frac{1}{p_{t, \theta}^{\text{ODE}}} \frac{\partial p_{t, \theta}^{\text{ODE}}}{\partial t}
= \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}) - \text{div}_x(f) + (s_{\theta}^{\text{ODE}} - \frac{1}{2} g^2(t) s_{\theta}) - f,
\]
where we apply the product rule of divergence in the last equality. After taking the gradient from the both sides, we obtain
\[
\frac{\partial s_{\theta}^{\text{ODE}}(x, t)}{\partial t} = \nabla_x \frac{\partial \log p_{t, \theta}^{\text{ODE}}}{\partial t}
= \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}) - \text{div}_x(f) \right] + \nabla_x \left[ (s_{\theta}^{\text{ODE}} - \frac{1}{2} g^2(t) s_{\theta}) - f \right] 
= \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}) - \text{div}_x(f) \right] + \nabla_x \left[ \frac{1}{2} g^2(t) (s_{\theta}^{\text{ODE}}, s_{\theta}) - (f, s_{\theta}^{\text{ODE}}) \right]
\]
\[\]
By rearranging Eq. (10) and combining with Eq. (29), it results in

\[
\epsilon_{s_{\theta}}(x, t) = \partial_t s_{\theta} - \nabla_x \left[ \frac{1}{2} g^2(t) \text{div}_x(s_{\theta}) - \text{div}_x(f) \right] - \nabla_x \left[ \frac{1}{2} g^2(t) \|s_{\theta}\|_2^2 - \langle f, s_{\theta}\rangle \right] \\
= \partial_t s_{\theta} - \partial_t s_{\theta}^{\text{ODE}} - \nabla_x \left[ \frac{1}{2} g^2(t) \langle s_{\theta} - s_{\theta}^{\text{ODE}}, s_{\theta}\rangle - \langle f, s_{\theta} - s_{\theta}^{\text{ODE}}\rangle\right]
\]

That is,

\[
\partial_t (s_{\theta}(x, t) - s_{\theta}^{\text{ODE}}(x, t)) = \epsilon_{s_{\theta}}(x, t) + \nabla_x \left[ \frac{1}{2} g^2(t) \langle s_{\theta} - s_{\theta}^{\text{ODE}}, s_{\theta}\rangle - \langle f, s_{\theta} - s_{\theta}^{\text{ODE}}\rangle\right]
\]

Fix a \( t \in [0, T] \), we integrate both sides of the above equation from \( \tau = T \) to \( \tau = t \)

\[
s_{\theta}(x, t) - s_{\theta}^{\text{ODE}}(x, t) = s_{\theta}(x, T) - s_{\theta}^{\text{ODE}}(x, T) + \int_T^t \epsilon_{s_{\theta}}(x, \tau) d\tau + \int_T^t \nabla_x \left[ \frac{1}{2} g^2(t) \langle s_{\theta} - s_{\theta}^{\text{ODE}}, s_{\theta}\rangle - \langle f, s_{\theta} - s_{\theta}^{\text{ODE}}\rangle\right] d\tau.
\]

Applying the \( \ell_2 \)-norm

\[
\|s_{\theta}(x, t) - s_{\theta}^{\text{ODE}}(x, t)\|_2 \leq \|s_{\theta}(x, T) - s_{\theta}^{\text{ODE}}(x, T)\|_2 \]
\[
+ \int_T^t \|\epsilon_{s_{\theta}}(x, \tau)\|_2 d\tau + \int_T^t \nabla_x \left[ \frac{1}{2} g^2(t) \langle s_{\theta} - s_{\theta}^{\text{ODE}}, s_{\theta}\rangle - \langle f, s_{\theta} - s_{\theta}^{\text{ODE}}\rangle\right]_2 d\tau. 
\]

(30)

In the last term, we may compute \( \nabla_x \left[ \frac{1}{2} g^2(t) \langle s_{\theta} - s_{\theta}^{\text{ODE}}, s_{\theta}\rangle - \langle f, s_{\theta} - s_{\theta}^{\text{ODE}}\rangle\right] \) as

\[
\frac{1}{2} g^2(\tau) \left[ \nabla_x s_{\theta} \cdot s_{\theta} - \nabla_x s_{\theta}^{\text{ODE}} \cdot s_{\theta}\right] + \frac{1}{2} g^2(\tau) \left( \nabla_x s_{\theta} \cdot (s_{\theta} - s_{\theta}^{\text{ODE}}) \right) \\
- \nabla_x f \cdot (s_{\theta} - s_{\theta}^{\text{ODE}}) - \nabla_x s_{\theta} \cdot f + \nabla_x s_{\theta}^{\text{ODE}} \cdot f
\]

(31)

Hence, we can further estimate the last term of Ineq. (30) as

\[
\int_T^t \left\| \nabla_x \left[ \frac{1}{2} g^2(t) \langle s_{\theta} - s_{\theta}^{\text{ODE}}, s_{\theta}\rangle - \langle f, s_{\theta} - s_{\theta}^{\text{ODE}}\rangle\right] \right\|_2 d\tau \\
\leq \int_T^t \frac{1}{2} g^2(\tau) \left\| \nabla_x s_{\theta} \cdot s_{\theta}\right\|_2 d\tau + \int_T^t \frac{1}{2} g^2(\tau) \left\| \nabla_x s_{\theta}^{\text{ODE}} \cdot s_{\theta}\right\|_2 d\tau \\
+ \int_T^t \frac{1}{2} g^2(\tau) \left\| \nabla_x s_{\theta} \cdot (s_{\theta} - s_{\theta}^{\text{ODE}})\right\|_2 d\tau + \int_T^t \left\| \nabla_x f \cdot (s_{\theta} - s_{\theta}^{\text{ODE}})\right\|_2 d\tau \\
+ \int_T^t \left\| \nabla_x s_{\theta} \cdot f\right\|_2 d\tau + \int_T^t \left\| \nabla_x s_{\theta}^{\text{ODE}} \cdot f\right\|_2 d\tau \\
\leq \int_T^t \frac{1}{2} g^2(\tau) \left\| \nabla_x s_{\theta}\right\|_{\text{op}} \|s_{\theta}\|_2 d\tau + \int_T^t \frac{1}{2} g^2(\tau) \left\| \nabla_x s_{\theta}^{\text{ODE}}\right\|_{\text{op}} \|s_{\theta}\|_2 d\tau \\
+ \int_T^t \frac{1}{2} g^2(\tau) \left\| \nabla_x s_{\theta}\right\|_{\text{op}} \|s_{\theta} - s_{\theta}^{\text{ODE}}\|_2 d\tau + \int_T^t \left\| \nabla_x f\right\|_{\text{op}} \|s_{\theta} - s_{\theta}^{\text{ODE}}\|_2 d\tau \\
+ \int_T^t \left\| \nabla_x s_{\theta}\right\|_{\text{op}} \|f\|_2 d\tau + \int_T^t \left\| \nabla_x s_{\theta}^{\text{ODE}}\right\|_{\text{op}} \|f\|_2 d\tau \\
\leq \left[L^2 \int_0^T g^2(\tau) d\tau (1 + \|x\|_2) \right] + \left[ \int_T^T \left( \frac{L}{2} g^2(\tau) + L \right) \|s_{\theta} - s_{\theta}^{\text{ODE}}\|_2 d\tau \right] + \left[ 2L^2 T (1 + \|x\|_2) \right] \\
\leq C(L, T, g)(1 + \|x\|_2) + \int_T^T \left( \frac{L}{2} g^2(\tau) + L \right) \|s_{\theta} - s_{\theta}^{\text{ODE}}\|_2 d\tau
\]
where \( \|A\|_{op} := \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \) denotes the operator norm of the matrix \( A \). In the second-to-last inequality, we apply Assumption A(a) together with the Rademacher’s theorem \([EG18]\) which bounds the total differentiations of \( s_\theta, s_\theta^{ODE} \), and \( f \) by their Lipschitz constants. Moreover, we summarize constant terms into \( C := C(L,T,g) \), which depends on \( L, T, \) and the function \( g \).

Combining this estimation with Ineq. (30), we have

\[
\|s_\theta(x, t) - s_\theta^{ODE}(x, t)\|_2 \leq \|s_\theta(x, T) - s_\theta^{ODE}(x, T)\|_2 + \int_0^T \|\epsilon_{s_\theta}(x, \tau)\|_2 \, d\tau + C(L, T, g)(1 + \|x\|_2) \\
+ \int_0^T \left( \frac{L}{2} g^2(\tau) + L \right) \|s_\theta(x, \tau) - s_\theta^{ODE}(x, \tau)\|_2 \, d\tau.
\]

Consider the following functions in Lemma 5

\[
u(t) := \|s_\theta(x, t) - s_\theta^{ODE}(x, t)\|_2 \\
\alpha(t) := \|s_\theta(x, T) - s_\theta^{ODE}(x, T)\|_2 + \int_0^T \|\epsilon_{s_\theta}(x, \tau)\|_2 \, d\tau + C(L, T, g)(1 + \|x\|_2) \\
\beta(t) := \frac{L}{2} g^2(t) + L.
\]

We remark that \( \alpha \equiv \alpha(t) \) is actually independent of \( t \). Then the lemma implies

\[
u(t) \leq \alpha \exp \left( \int_t^T \beta(\tau) \, d\tau \right) \\
\leq C(L, T, g) \left[ \|s_\theta(x, T) - s_\theta^{ODE}(x, T)\|_2 + \int_0^T \|\epsilon_{s_\theta}(x, \tau)\|_2 \, d\tau + (1 + \|x\|_2) \right],
\]

where we bound \( \exp \left( \int_0^T \beta(\tau) \, d\tau \right) \) by \( \exp \left( \int_0^T \beta(t) \, d\tau \right) \) which is a constant, and we absorb all constant terms into \( C(L, T, g) \).

We are going to square both sides of the above estimation and take the expectation over \( q_t(x) \). For the sake of simplicity, we denote \( e_\theta(x) := \int_0^T \|\epsilon_{s_\theta}(x, \tau)\|_2 \, d\tau \) and \( \delta_\theta(x) := \|s_\theta(x, T) - s_\theta^{ODE}(x, T)\|_2 \), and hence, we obtain

\[
\mathbb{E}_{q_t(x)} \left[ \nu^2(t) \right] \\
\leq C(L, T, g) \cdot \mathbb{E}_{q_t(x)} \left[ \delta_\theta(x) + e_\theta(x) + (1 + \|x\|_2) \right]^2 \\
\leq C(L, T, g) \cdot \left\{ \mathbb{E}_{q_t(x)} \left[ \delta^2_\theta(x) \right] + \mathbb{E}_{q_t(x)} \left[ e^2_\theta(x) \right] + \mathbb{E}_{q_t(x)} \left[ (1 + \|x\|_2)^2 \right] \right\} \\
+ \mathbb{E}_{q_t(x)} \left[ \delta_\theta(x) e_\theta(x) \right] + \mathbb{E}_{q_t(x)} \left[ \delta_\theta(x) (1 + \|x\|_2) \right] + \mathbb{E}_{q_t(x)} \left[ e_\theta(x) (1 + \|x\|_2) \right] \right\} \\
\tag{32}
\]

The last three terms of the above inequality can be further bounded via Cauchy–Schwarz inequality

\[
\mathbb{E}_{q_t(x)} \left[ \delta_\theta(x) e_\theta(x) \right] \leq \sqrt{\mathbb{E}_{q_t(x)} \left[ \delta^2_\theta(x) \right]} \sqrt{\mathbb{E}_{q_t(x)} \left[ e^2_\theta(x) \right]} \\
\mathbb{E}_{q_t(x)} \left[ \delta_\theta(x) (1 + \|x\|_2) \right] \leq \sqrt{\mathbb{E}_{q_t(x)} \left[ \delta^2_\theta(x) \right]} \sqrt{\mathbb{E}_{q_t(x)} \left[ (1 + \|x\|_2)^2 \right]} \\
\mathbb{E}_{q_t(x)} \left[ e_\theta(x) (1 + \|x\|_2) \right] \leq \sqrt{\mathbb{E}_{q_t(x)} \left[ e^2_\theta(x) \right]} \sqrt{\mathbb{E}_{q_t(x)} \left[ (1 + \|x\|_2)^2 \right]}.
\]

It is noticed that Assumption A(a) indeed implies the following estimation which bounds 1st- and 2nd- central moments for all \( t \in [0, T] \)

\[
\sup_{t \in [0,T]} \left\{ \mathbb{E}_{x \sim q_t(x)} \left[ \|x\|_2 \right] \right\}, \sup_{t \in [0,T]} \left\{ \mathbb{E}_{x \sim q_t(x)} \left[ \|x\|_2^2 \right] \right\} \leq L \tag{33}
\]
as by Cauchy Schwartz inequality that $E_{x \sim q_0(x)}[\|x\|_2] \leq E_{x \sim q_0(x)}[\|x\|_2^2] \leq L$ and the transition density $q_0(x(t)|x(0))$ has bounded covariance matrices as a function in $t \in [0,T]$. With Ineq. (33) and Assumption A.(g), Ineq. (32) becomes

$$E_{q_t(x)} \left[ \|s_\theta(x, t) - s_\theta^{ODE}(x, t)\|^2 \right] \leq C(L, T, g) \cdot \left\{ \delta_T^2 + E_{q_t(x)}[\sigma_\theta^2(x)] + (1 + 3L) \right\}$$

$$+ \delta_T \sqrt{E_{q_t(x)}[\sigma_\theta^2(x)] + \delta_T \sqrt{1 + 3L + \sqrt{1 + 3L} \sqrt{E_{q_t(x)}[\sigma_\theta^2(x)]}}}$$

$$\leq C(L, T, g, \delta_T) \cdot \left( E_{q_t(x)}[\sigma_\theta^2(x)] + \sqrt{E_{q_t(x)}[\sigma_\theta^2(x)]} + 1 \right)$$

$$\leq C(L, T, g, \delta_T) \cdot \left( M(\theta) + \sqrt{M(\theta)} + 1 \right).$$

Again, we abuse of the notation and summarize constants into $C = C(L, T, g, \delta_T)$. Therefore, after combining the Ineq. (15) and the estimation above, we obtain (with a fusion of constant term)

$$J_{Diff}(\theta) \leq J_{SM}(\theta) \cdot J_{Fisher}(\theta)$$

$$\leq C(L, T, g, \delta_T) \cdot J_{SM}(\theta) \left( M(\theta) + \sqrt{M(\theta)} + 1 \right)$$

\[\blacksquare\]

**Remark 2.** We remark that one can easily extend the proposition and obtain a sharper bound. We provide an approach as an instance. Let us assume there is a constant $\delta_{ODE} > 0$ to control the distance between $\nabla_x s_\theta - s_\theta^{ODE}$ and $\nabla_x s_\theta$ instead (in this case, we do not require Assumption A.(f)). That is,

$$\sup_{R^D \times [0,T]} \|\nabla_x (s_\theta - s_\theta^{ODE})\|_2 \leq \delta_{ODE}. \quad (34)$$

Notice that Eq. (31) can be rewritten as

$$\frac{1}{2} g^2(\tau) \left( \nabla_x (s_\theta - s_\theta^{ODE}) \cdot s_\theta \right) + \frac{1}{2} g^2(\tau) \left( \nabla_x s_\theta \cdot (s_\theta - s_\theta^{ODE}) \right)$$

$$= \nabla_x f \cdot (s_\theta - s_\theta^{ODE}) - \nabla_x (s_\theta - s_\theta^{ODE}) \cdot f.$$

Following the same argument as the proof of Theorem 1 together with the help of Ineq. (34), $J_{Fisher}(\theta)$ can be upper bounded by a constant which depends monotonically increasingly on $\delta_{ODE}$. Therefore, we can get a sharper estimation if $\delta_{ODE}$ is smaller.