FP-Diffusion: Improving Score-based Diffusion Models by Enforcing the Underlying Score Fokker-Planck Equation

Chieh-Hsin Lai [1] **Yuhta Takida** [1] **Naoki Murata** [1] **Toshimitsu Uesaka** [1] **Yuki Mitsufuji** [1 2] **Stefano Ermon** [3]

Abstract

Score-based generative models (SGMs) 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 linked together by the *Fokker-Planck equation* (FPE), a partial differential equation (PDE) governing the spatial-temporal evolution of a density undergoing a diffusion process. In this work, we derive a corresponding equation called the *score FPE* that characterizes the noise-conditional scores of the perturbed data densities (i.e., their gradients). Surprisingly, despite the impressive empirical performance, we observe that scores learned through denoising score matching (DSM) fail to fulfill the underlying score FPE, which is an inherent self-consistency property of the ground truth score. We prove that satisfying the score FPE is desirable as it improves the likelihood and the degree of conservativity. Hence, we propose to regularize the DSM objective to enforce satisfaction of the score FPE, and we show the effectiveness of this approach across various datasets.

1. Introduction

Score-based generative models (SGMs), also referred to as diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2020b;a), have led to major advances in the generation of synthetic images (Dhariwal & Nichol, 2021; Saharia et al., 2022; Rombach et al., 2022; Kim et al., 2022) and audio (Kong et al., 2020). In addition, SGMs have been applied to various downstream tasks such as media content editing (Meng et al., 2021b; Cheuk et al., 2022), or restoration (Kawar et al., 2022; Saito et al., 2022; Murata et al., 2023). An SGM involves a stochastic forward and backward process. In the forward process, also known as the diffusion process, noise with gradually increasing variances is added to each data point until the original structure is lost, transforming data into pure noise. The backward process attempts to reverse the diffusion process by using a neural network (called a noise-conditional score model) that is trained to gradually denoise the data, effectively transforming pure noise into clean data samples. The neural network is trained with a denoising score matching objective (Hyvärinen & Dayan, 2005; Vincent, 2011) to estimate the score (i.e., the gradient of the log-likelihood function) of the data density perturbed with various amounts of noise (as in forward process).

The training can be interpreted as a joint estimation of the scores 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 (Song et al., 2020b) and the spatial-temporal evolution of the data density is thus governed by the classic Fokker-Planck partial differential equation (PDE) (Øksendal, 2003). In principle, this implies that with knowledge of the density for a *single* noise level, we could recover all the densities by solving the Fokker-Planck equation (FPE) without any additional learning.

Our contributions Building on the above notions, we derive an associated system of PDEs that characterizes the evolution of the *scores* (i.e., gradients) of the perturbed data densities; we term it as *score Fokker-Planck equation* (*score FPE*). In theory, the ground truth scores of the perturbed data densities must satisfy the score FPE (*self-consistency property*). Hence, we mathematically study the implications of satisfying the score FPE. We prove the following effects of reducing the score FPE error: (a) improvement in the log-likelihood of the probability flow ordinary differential equation (ODE) diffusion mode (Song et al., 2020b), (Theorems 4.2 and 4.3); and (b) improvement in the degree of conservativity of the models (Proposition 4.4). In addition, we prove that (c) score FPE error reduction can be achieved by enforcing higher-order score matching (Meng

[1]Sony AI, Tokyo, Japan [2]Sony Group Corporation, Tokyo, Japan [3]Department of Computer Science, Stanford University, Stanford, CA, USA. Correspondence to: Chieh-Hsin Lai <Chieh-hsin.lai@sony.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

et al., 2021a; Lu et al., 2022) (Proposition 4.6). In practice, we observe that many existing, pre-trained score models do not numerically satisfy the score FPE. Therefore, we propose a new loss function for training diffusion models by combining the traditional score matching objective with a regularization term derived from the underlying score FPE to enforce the consistency of models. Our proposed new method is called *FP-Diffusion*. We show that FP-Diffusion enables more accurate density estimation on synthetic data and improves the likelihood on the MNIST, Fashion MNIST, CIFAR-10 and ImageNet32 (ImageNet downsampled to 32×32) (Chrabaszcz et al., 2017) datasets.

2. Background

Song et al. (2020b) unified denoising score matching (Song & Ermon, 2019) and diffusion probabilistic models (Sohl-Dickstein et al., 2015; Ho et al., 2020) via a stochastic process $\boldsymbol{x}(t)$ with continuous time $t \in [0, T]$. The process is driven by the following forward SDE

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

where $\boldsymbol{f}(\cdot, t)\colon \mathbb{R}^D \to \mathbb{R}^D$, $g(\cdot)\colon \mathbb{R} \to \mathbb{R}$ are pre-assigned[1] and $\boldsymbol{w}_t$ is a standard Wiener process. Under moderate conditions (Anderson, 1982), a reverse time SDE from T to 0 can be obtained as

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

where $\bar{\boldsymbol{w}}_t$ is a standard Wiener process in reverse time, and $q_t(\boldsymbol{x})$ denotes the ground truth marginal density of $\boldsymbol{x}(t)$ following Eq. (1). We can train a time-conditional 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 a score matching objective (Hyvärinen & Dayan, 2005) $\mathcal{J}_{\mathrm{SM}}(\boldsymbol{\theta}; \lambda(\cdot)) :=$

$$\frac{1}{2}\int_0^T \lambda(t)\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}\Big[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})\|_2^2\Big]dt.$$

As $q_t(\boldsymbol{x})$ is generally inaccessible, the denoising score matching (DSM) loss (Vincent, 2011; Song et al., 2020b) $\mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}; \lambda(\cdot))$ is exploited in practice instead

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

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

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

Let $p^{\mathrm{SDE}}_{t,\boldsymbol{\theta}}$ denote the marginal distribution of $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(t)$ with an initial distribution defined as the prior π, where we suppress the dependence on π for compactness. We can design $\boldsymbol{f}$ and g in Eq. (2), such that $q_T(\boldsymbol{x})$ approximates a simple prior π; samples $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(0) \sim p^{\mathrm{SDE}}_{0,\boldsymbol{\theta}}$ can be generated 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.

Song et al. (2020b) also introduced a deterministic process (with a zero diffusion term) that describes the evolution of samples whose trajectories share the same marginal probability densities as the forward SDE (Eq. (4)). Specifically, the process evolves through time according to the following probability flow ODE

$$\frac{d\boldsymbol{x}}{dt}(t) = \boldsymbol{f}(\boldsymbol{x}(t), t) - \frac{1}{2}g^2(t)\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}(t)). \tag{5}$$

As in the SDE case, the ground truth score in Eq. (5) is approximated with the learned score model $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \approx \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$. This yields to the following parameterized probability flow ODE

$$\frac{d\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}}{dt}(t) = \boldsymbol{f}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t) - \frac{1}{2}g^2(t)\boldsymbol{s}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t) \tag{6}$$

We denote the marginal density of $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}$ as $p^{\mathrm{ODE}}_{t,\boldsymbol{\theta}}$ with an initial condition sampled from the prior π, For compactness, we omit the dependence on π in the notation. By solving Eq. (6) numerically using an initial value $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(T) \sim \pi$, we can generate a sample $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(0) \sim p^{\mathrm{ODE}}_{0,\boldsymbol{\theta}}$ to approximate sampling from the data distribution. Indeed, the deterministic dynamics in Eq. (6) make it possible to compute exact likelihoods for this generative model. Let $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t) \in \mathbb{R}^D$ evolve in reverse time via Eq. (6), starting with $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(T) \sim \pi$. The "instantaneous change of variables" (Chen et al., 2018) characterizes the temporal changes in $\log p^{\mathrm{ODE}}_{t,\boldsymbol{\theta}}$ along the trajectory $\{\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t) : t \in [0, T]\}$ via the following ODE:

$$\begin{aligned}&\frac{d \log p^{\mathrm{ODE}}_{t,\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t))}{dt} \\ =&\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t)\big) - \mathrm{div}_{\boldsymbol{x}}\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t)\big).\end{aligned}$$

Hence, the log-likelihood can be exactly calculated by numerically solving the concatenated ODEs backward from T to 0, after initialization with $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(0) \sim q_0(\boldsymbol{x})$

$$\begin{aligned}&\frac{d}{dt}\begin{bmatrix}\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t) \\ \log p^{\mathrm{ODE}}_{t,\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t))\end{bmatrix} \\ =&\begin{bmatrix}\boldsymbol{f}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t) - \frac{1}{2}g^2(t)\boldsymbol{s}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t) \\ \frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t)\big) - \mathrm{div}_{\boldsymbol{x}}\big(\boldsymbol{f}(\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}(t), t)\big)\end{bmatrix}.\end{aligned}$$

[1] With specific choices of $\boldsymbol{f}$ and g, there are two common instantiations of the stochastic differential equation (SDE): VE and VP. See Appendix A for details.

3. Score Fokker-Planck equation for diffusion

It is well known that the evolution of the ground truth density $q_t(\boldsymbol{x})$ associated with Eq. (1) is governed by the Fokker-Planck equation (FPE) (Øksendal, 2003)

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

where $\tilde{\boldsymbol{F}}(\boldsymbol{x},t) := \boldsymbol{f}(\boldsymbol{x},t) - \frac{1}{2}g^2(t)\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$. As there is a one-to-one mapping (up to a constant) between densities and their scores, we derive (in Appendix G) an equivalent system of PDEs for the ground truth *scores* $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$. We designate it as the *score Fokker-Planck equation* or simply the *score FPE*.

Proposition 3.1 (Score FPE)**.** *Assume the* ground truth *density $q_t(\boldsymbol{x})$ is sufficiently smooth on $\mathbb{R}^D \times [0,T]$ with its score denoted as $\boldsymbol{s}(\boldsymbol{x},t) := \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$. Then for all $(\boldsymbol{x},t) \in \mathbb{R}^D \times [0,T]$, its log-density satisfies the PDE*

$$\begin{aligned}\partial_t \log q_t(\boldsymbol{x}) =& \frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}(\boldsymbol{x},t)) + \frac{1}{2}g^2(t)\,\|\boldsymbol{s}(\boldsymbol{x},t)\|_2^2 \\ &- \langle \boldsymbol{f}(\boldsymbol{x},t), \boldsymbol{s}(\boldsymbol{x},t)\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f}(\boldsymbol{x},t))\end{aligned} \tag{7}$$

and its score $\boldsymbol{s}$ satisfies the following system of PDEs

$$\begin{aligned}\partial_t \boldsymbol{s}(\boldsymbol{x},t) = \nabla_{\boldsymbol{x}}\Big[&\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}(\boldsymbol{x},t)) + \frac{1}{2}g^2(t)\,\|\boldsymbol{s}(\boldsymbol{x},t)\|_2^2 \\ &- \langle \boldsymbol{f}(\boldsymbol{x},t), \boldsymbol{s}(\boldsymbol{x},t)\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f}(\boldsymbol{x},t))\Big].\end{aligned} \tag{8}$$

For notational simplicity, let $\mathcal{L}[\cdot] := \frac{1}{2}g^2\mathrm{div}_{\boldsymbol{x}}(\cdot) + \frac{1}{2}g^2\,\|\cdot\|_2^2 - \langle \boldsymbol{f}, \cdot\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})$ be the operator mapping vector fields to real-valued functions. Thus, Eq. (7) and Eq. (8) can be expressed as $\partial_t \log q_t(\boldsymbol{x}) = \mathcal{L}[\boldsymbol{s}](\boldsymbol{x},t)$ and $\partial_t \boldsymbol{s}(\boldsymbol{x},t) = \nabla_{\boldsymbol{x}}\mathcal{L}[\boldsymbol{s}](\boldsymbol{x},t)$, respectively.

Proposition 3.1 shows that the time-conditional scores $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$ 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 it intuitively by considering the special case when $\boldsymbol{f} \equiv \boldsymbol{0}$ and $g \equiv 1$, i.e., when, $\boldsymbol{x}(t)$ 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 that q_t can be analytically obtained from q_{t_0} (Masry & Rice, 1992) (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 provide empirical evidence to substantiate Proposition 3.1 from two distinct perspectives, as presented in Section 6.1 and Appendix B.1, respectively.

3.1. Pre-trained scores fail to satisfy score FPEs

Theoretically, with sufficient data and model capacity, score matching ensures that the optimal solution to Eq. (3) should satisfy Eq. (8) as it should approximate the ground truth score well. However, in our experiments, we observe that pre-trained scores $\boldsymbol{s}_{\boldsymbol{\theta}}$ learned via Eq. (3) do not fulfill the score FPE. Therefore, we introduce an error term $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}] := \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t)$ to quantify how much $\boldsymbol{s}_{\boldsymbol{\theta}}$ deviates from the score FPE

$$\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t) := \partial_t \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \nabla_{\boldsymbol{x}}\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t). \tag{9}$$

Set $T = 1$, we define the average residual of the score FPE, computed over $\boldsymbol{x}$, as a function of $t \in [0,1]$

$$r_{\text{FP, trans.}}[\boldsymbol{s}_{\boldsymbol{\theta}}](t) := \frac{1}{D}\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{\boldsymbol{x}(t)|\boldsymbol{x}(0)}\Big[\,\|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t)\|_2\Big].$$

We further consider the following averaged residual for DSM

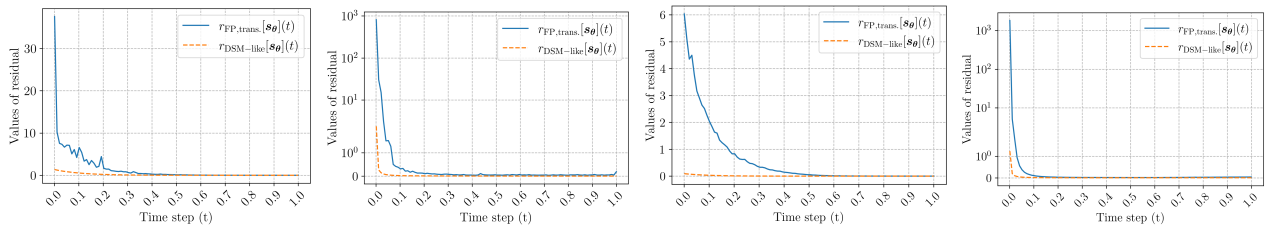
$$\begin{aligned}r_{\text{DSM-like}}[\boldsymbol{s}_{\boldsymbol{\theta}}](t) := \frac{1}{D}\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{\boldsymbol{x}(t)|\boldsymbol{x}(0)}\Big[&\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}(t),t) \\ &- \nabla_{\boldsymbol{x}(t)}\log q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))\|_2\Big].\end{aligned}$$

Compared to the integrand in the standard DSM loss in Eq. (3), $r_{\text{DSM-like}}[\boldsymbol{s}_{\boldsymbol{\theta}}]$ uses the ℓ_2-norm (instead of the MSE) and drops the time-weighting function $\lambda(t)$ to be consistent with the averaged residuals of the score FPE.

Figure 1 plots these residuals for score models that were pre-trained via DSM on the MNIST and CIFAR-10 datasets. Despite achieving a low $r_{\text{DSM-like}}$ across all t (orange curve), the pre-trained score models fail to satisfy the score FPE equation, especially for small t (blue curve). This implies that models learned by DSM do not satisfy the score FPE.

4. Theoretical implications of score FPE

In this section, we first study three implications of satisfying the score FPE. Specifically, we show in Section 4.1 that simultaneous minimization of quantities related to the score FPE and the conventional score matching objective can reduce the KL divergence between the data density q_0 and the density $p_{0,\boldsymbol{\theta}}^{\text{ODE}}$, determined by the parametrized probability flow ODE (Eq. (6)). In Section 4.2 we prove that controlling of $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}]$ implicitly enforces the *conservativity* of $\boldsymbol{s}_{\boldsymbol{\theta}}$. Moreover, in Section 4.3 we prove that if the score FPE is satisfied, then under certain conditions, $\boldsymbol{s}_{\boldsymbol{\theta}}$, ground truth score $\boldsymbol{s}$, $\nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{\text{SDE}}$, and $\nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{\text{ODE}}$ must match. Here $p_{t,\boldsymbol{\theta}}^{\text{SDE}}$ and $p_{t,\boldsymbol{\theta}}^{\text{ODE}}$ were defined in Section 2 as the marginal density of the parametrized diffusion process and the probability flow ODE, respectively. Finally, in Section 4.4, we investigate the connection between higher-order score matching (Meng et al., 2021a; Lu et al., 2022) and the score FPE. We provide the proofs of all theorems in Appendix G.



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

Figure 1. Comparison of the numerical scales of $r_{\text{DSM-like}}[\boldsymbol{s}_{\boldsymbol{\theta}}](t)$ and $r_{\text{FP, trans.}}[\boldsymbol{s}_{\boldsymbol{\theta}}](t)$ for pre-trained scores $\boldsymbol{s}_{\boldsymbol{\theta}}$ on MNIST and CIFAR-10. We treat these errors as functions of time. The pre-trained models do not numerically satisfy the score FPE, in contrast to their DSM-like errors. We attempt to explain this phenomenon in Sections. 4.2 and 4.4.

4.1. Minimization $\mathcal{D}_{\text{KL}}\big(q_0\big\|p^{\text{ODE}}_{0,\boldsymbol{\theta}}\big)$

In this section, we show that under certain regularity conditions (see Assumptions F.1 and F.2), simultaneous minimization of $\mathcal{J}_{\text{SM}}(\boldsymbol{\theta})$ and certain score FPE related quantities (see Eqs. (11) and (12)) can decrease the KL divergence between q_0 and $p^{\text{ODE}}_{0,\boldsymbol{\theta}}$, denoted as $\mathcal{D}_{\text{KL}}\big(q_0\big\|p^{\text{ODE}}_{0,\boldsymbol{\theta}}\big)$. This is equivalent to improving the likelihood of data under $p^{\text{ODE}}_{0,\boldsymbol{\theta}}$.

First, we review an equation proposed by Lu et al. (2022) that quantifies the exact gap between $\mathcal{D}_{\text{KL}}\big(q_0\big\|p^{\text{ODE}}_{0,\boldsymbol{\theta}}\big)$ and the score matching loss $\mathcal{J}_{\text{SM}}(\boldsymbol{\theta})$. For compactness, we denote $\boldsymbol{s}^{\text{ODE}}_{\boldsymbol{\theta}}(\boldsymbol{x},t) := \nabla_{\boldsymbol{x}} \log p^{\text{ODE}}_{t,\boldsymbol{\theta}}(\boldsymbol{x})$.

Lemma 4.1 (Lu et al. (2022))**.** *Set $\lambda(t) = g^2(t)$. Let q_0 be the data distribution, and q_t be the marginal density of $\boldsymbol{x}(t)$ following Eq. (1). Assume that Assumption F.1 is satisfied. Then,*

$$\mathcal{D}_{\text{KL}}\big(q_0\big\|p^{\text{ODE}}_{0,\boldsymbol{\theta}}\big) = \mathcal{D}_{\text{KL}}\big(q_T\big\|p^{\text{ODE}}_{T,\boldsymbol{\theta}}\big) + \mathcal{J}_{SM}(\boldsymbol{\theta}) + \mathcal{J}_{Diff}(\boldsymbol{\theta}),$$

where

$$\mathcal{J}_{Diff}(\boldsymbol{\theta}) = \frac{1}{2}\int_0^T g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\Big[\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}(\boldsymbol{x},t)\big)^\top \big(\boldsymbol{s}^{\text{ODE}}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)\big)\Big]dt.$$

We now introduce the main theoretical results in this section. First, we note that application of the Cauchy-Schwartz inequality to $\mathcal{J}_{\text{Diff}}(\boldsymbol{\theta})$ gives

$$|\mathcal{J}_{\text{Diff}}(\boldsymbol{\theta})| \le \sqrt{\mathcal{J}_{\text{SM}}(\boldsymbol{\theta})} \cdot \sqrt{\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta})}.$$

Here, $\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta})$ is a Fisher-like divergence in terms of the two scores $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$ and $\boldsymbol{s}^{\text{ODE}}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$, defined as $\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) :=$

$$\frac{1}{2}\int_0^T g^2(t)\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}^{\text{ODE}}_{\boldsymbol{\theta}}(\boldsymbol{x},t)\right\|_2^2 dt.$$

Next, in Theorem. 4.2, we show that under Assumption F.1, $\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta})$ can be bounded from above by the averaged residual of the score FPE $M(\boldsymbol{\theta})$:

$$\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) \lesssim M(\boldsymbol{\theta}) + \sqrt{M(\boldsymbol{\theta})} + C_1, \tag{10}$$

where $C_1 > 0$ is a constant, $\lesssim$ denotes multiplicative constants independent of $\boldsymbol{\theta}$ are concealed, and $M(\boldsymbol{\theta}) :=$

$$\sup_{t\in[0,T]} \mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}\left[\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2^2\, d\tau\right]. \tag{11}$$

Furthermore, we can compute

$$M(\boldsymbol{\theta}) \le \sup_{\boldsymbol{x}}\left[\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2^2\, d\tau\right],$$

meaning that this upper bound measures the worst time-averaged score FPE error. In Appendix G.3, we consider more interpretable quantities than $M(\boldsymbol{\theta})$ by introducing the density weighting $p_\tau(\boldsymbol{x})$ in τ-integrand and derive similar estimations as in Ineq. (10).

Moreover, we prove in Theorem. 4.3 that with a different regularity condition (Assumption F.2), $\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta})$ is upper bounded by $M(\boldsymbol{\theta})$ and a "time-derivative taming" term that can be derived from Eq. (7) which is defined as

$$m(\boldsymbol{\theta}) := \sup_{\boldsymbol{x}} \int_0^T |\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)|\, d\tau. \tag{12}$$

More specifically,

$$\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) \lesssim M(\boldsymbol{\theta}) + m(\boldsymbol{\theta}) + C_2, \tag{13}$$

where C_2 is another constant, distinct from C_1.

Hence, Lemma. 4.1 together with Ineq. (10) or (13) implies that $\mathcal{D}_{\text{KL}}\big(q_0\big\|p^{\text{ODE}}_{0,\boldsymbol{\theta}}\big)$ decreases when "$M(\boldsymbol{\theta})$ and $\mathcal{J}_{\text{SM}}(\boldsymbol{\theta})$" or "$M(\boldsymbol{\theta})$, $m(\boldsymbol{\theta})$, and $\mathcal{J}_{\text{SM}}(\boldsymbol{\theta})$" are reduced simultaneously. We now rigorously state these theorems.

Theorem 4.2. *We have*

$$\Big(\mathcal{J}_{Diff}(\boldsymbol{\theta})\Big)^2 \le \mathcal{J}_{SM}(\boldsymbol{\theta}) \cdot \mathcal{J}_{Fisher}(\boldsymbol{\theta}). \tag{14}$$

Moreover, if Assumption. F.1 is fulfilled, then there is another finite constant $C_1 > 0$ independent of $\boldsymbol{\theta}$ such that we can

further bound Ineq. (14) above by

$$\left(\mathcal{J}_{Diff}(\boldsymbol{\theta})\right)^2 \lesssim \mathcal{J}_{SM}(\boldsymbol{\theta}) \cdot \left(M(\boldsymbol{\theta}) + \sqrt{M(\boldsymbol{\theta})} + C_1\right). \quad (15)$$

Thus, $\mathcal{D}_{\mathrm{KL}}\left(q_0 \| p^{\mathrm{ODE}}_{0,\boldsymbol{\theta}}\right) \lesssim \mathcal{D}_{\mathrm{KL}}\left(q_T \| p^{\mathrm{ODE}}_{T,\boldsymbol{\theta}}\right)$

$$+\mathcal{J}_{SM}(\boldsymbol{\theta}) + \mathcal{J}^{1/2}_{SM}(\boldsymbol{\theta})\left(M(\boldsymbol{\theta}) + \sqrt{M(\boldsymbol{\theta})} + C_1\right)^{1/2}.$$

Theorem 4.3. *If Assumption. F.2 is satisfied, then there is another finite constant $C_2 > 0$ independent of $\boldsymbol{\theta}$ such that*

$$\left(\mathcal{J}_{Diff}(\boldsymbol{\theta})\right)^2 \lesssim \mathcal{J}_{SM}(\boldsymbol{\theta}) \cdot \left(M(\boldsymbol{\theta}) + m(\boldsymbol{\theta}) + C_2\right). \quad (16)$$

It is noticed that constants C_1 and C_2 involve regularity bounds of the ground truth density and Lipschitz constants of networks. Hence, the upper bounds in Ineq. (15) and (16) are difficult to compare.

As the ground truth score should follow the score FPE, it is intuitive that reduction of the score FPE residual encourages the network-parametrized score to approach the ground truth score (a special case is proved in Proposition 4.5). Theorems 4.2 and 4.3 support that reduction of these quantities related to the score FPE may also reduce the gap (in the KL divergence) of their corresponding densities. In Section 7, we empirically support these claims.

4.2. 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 (Salimans & Ho, 2021). Below, we prove that we can implicitly enforce conservativity by minimizing the time-averaged error $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)$ of the score FPE.

Proposition 4.4. *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 exists a real-valued function $\Psi_{\boldsymbol{\theta}} \colon \mathbb{R}^D \times [0, T] \to \mathbb{R}$ (with an explicit expression) that satisfies*

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = \int_{t_{\boldsymbol{\theta}}}^{t} \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) d\tau, \quad (17)$$

for all $(\boldsymbol{x}, t) \in \mathbb{R}^D \times [0, T]$. In particular,

$$\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\|_2 \leq \left| \int_{t}^{t_{\boldsymbol{\theta}}} \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\|_2 \, d\tau \right|. \quad (18)$$

Eq. (17) indicates that the error of the score FPE quantifies the degree of conservativity of $\boldsymbol{s}_{\boldsymbol{\theta}}$. We further explain this idea via Ineq. (18), from which we easily obtain $\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\|_2 \leq \left| \int_t^{t_{\boldsymbol{\theta}}} \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\|_2 \, d\tau \right| \leq \int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\|_2 \, d\tau$, for any $\boldsymbol{x}$ and t. Thus, if the $\boldsymbol{\theta}$-parametrized score approximately satisfies the score FPE, giving a small score FPE error $\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\|_2 \, d\tau$, then the estimated score should nearly be conservative, i.e., close to the gradient of a scalar function $\Psi_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$. We empirically support this fact in Section 6.2.

Proposition 4.4 necessitates a precise alignment of scores at a given timestep. However, we propose a modification that allows for a small discrepancy by incorporating an error term into the score matching process. As a result, we present an expanded proposition, namely Proposition G.4, which is detailed in Appendix G.5.

4.3. Equivalence of scores

We now investigate another implication of satisfying the score FPE which connects the score $\boldsymbol{s}_{\boldsymbol{\theta}}$ with the ground truth $\boldsymbol{s}$, $\boldsymbol{s}^{\mathrm{SDE}}_{\boldsymbol{\theta}}$, and $\boldsymbol{s}^{\mathrm{ODE}}_{\boldsymbol{\theta}}$. The following proposition provides conditions under which all of these scores are identical if we train to reach a zero residual for the score FPE for all $(\boldsymbol{x}, t)$.

Proposition 4.5. *(1) Suppose in some suitable function space, $\mathbf{0}$ is the unique strong solution to the PDEs $\partial_t \boldsymbol{v} - \nabla_{\boldsymbol{x}}\left[\frac{1}{2}g^2 \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{v}) + \frac{1}{2}g^2\left(\|\boldsymbol{v}\|_2^2 + 2\langle \boldsymbol{v}, \boldsymbol{s}\rangle\right) - \langle \boldsymbol{f}, \boldsymbol{v}\rangle\right] = 0$ with a zero initial condition $\boldsymbol{v}(\boldsymbol{x}, 0) \equiv 0$ and a zero boundary condition. If there is some $\boldsymbol{\theta}_0$ so that for all $(\boldsymbol{x}, t)$ $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}_0}](\boldsymbol{x}, t) = 0$ and that $\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{x}, 0) = \boldsymbol{s}(\boldsymbol{x}, 0)$, then $\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{x}, t) = \boldsymbol{s}(\boldsymbol{x}, t)$, for all $(\boldsymbol{x}, t)$.*

(2) Moreover, suppose the PDEs $\partial_t \boldsymbol{v} + \nabla_{\boldsymbol{x}}\left[\frac{1}{2}g^2 \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{v}) + \frac{1}{2}g^2\|\boldsymbol{v}\|_2^2 + \langle \boldsymbol{f}, \boldsymbol{v}\rangle\right] = 0$ with zero initial and boundary condition have $\mathbf{0}$ as the unique strong solution. Then $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}_0}] \equiv 0$ and $\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{x}, 0) \equiv \boldsymbol{s}^{\mathrm{SDE}}_{\boldsymbol{\theta}_0}(\boldsymbol{x}, 0)$ implies $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{s}^{\mathrm{SDE}}_{\boldsymbol{\theta}_0}$.

(3) Lastly, if there is some $\boldsymbol{\theta}_0$ such that $\partial_t \boldsymbol{v} - \nabla_{\boldsymbol{x}}\left[\langle \frac{1}{2}g^2 \boldsymbol{s}_{\boldsymbol{\theta}_0} - \boldsymbol{f}, \boldsymbol{v}\rangle\right] = 0$ with zero initial and boundary conditions admit $\mathbf{0}$ as the unique strong solution, then $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}_0}] \equiv 0$ and $\boldsymbol{s}_{\boldsymbol{\theta}_0}(\boldsymbol{x}, 0) \equiv \boldsymbol{s}^{\mathrm{ODE}}_{\boldsymbol{\theta}_0}(\boldsymbol{x}, 0)$ implies $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{s}^{\mathrm{ODE}}_{\boldsymbol{\theta}_0}$.

Proposition 4.5 implies that if the parametric scores match with the ground truth score at the initial time, the only global minimum is the ground truth score. Essentially, the scores at any given time can be obtained solely by achieving a flawless alignment of scores at a single timestep through the dynamics of PDE. This indicates the score FPE residual is a proper quantity to measure the gaps between the ground truth and parametric scores. Indeed, this proposition is an extreme case of "the continuous dependence of PDE solutions on parameters $\boldsymbol{\theta}$" (Artstein, 1975). A more sophisticated analysis (Lunardi, 2012; Papageorgiou, 1994) can be applied to prove for instance, that as $\|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}]\| \to 0$, $\left\|\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}^{\mathrm{SDE}}_{\boldsymbol{\theta}}\right\| \to 0$ if $\boldsymbol{f} \equiv 0$ (with a careful choice of norms). However, such technical generalization is outside this work's scope.

4.4. Higher-order score matching

Higher-order derivatives of the score can yield additional information about the data distribution (Meng et al., 2021a; Lu et al., 2022). We prove that bounding of the higher-order score matching loss can further control the FPE residual $\left\|\int_0^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)d\tau\right\|_2$ for all $t \in [0,T]$. This partially explains why scores learned via $\mathcal{J}_{\text{DSM}}$ do not satisfy the score FPE, as DSM only matches gradients, while higher-order derivatives may still deviate from the ground truth.

Proposition 4.6. *Assume that on $\mathbb{R}^D \times [0,T]$, higher-order score matchings admit the following error bounds: $\|\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\theta}}\|_2 \leq \delta_0$, $\|\nabla_{\boldsymbol{x}}(\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\theta}})\|_F \leq \delta_1$, $\|\nabla_{\boldsymbol{x}}\text{div}_{\boldsymbol{x}}(\boldsymbol{s}-\boldsymbol{s}_{\boldsymbol{\theta}})\|_2 \leq \delta_2$.*

Then for all $(\boldsymbol{x},t) \in \mathbb{R}^D \times [0,T]$, $\left\|\int_0^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)d\tau\right\|_2$

$$\leq 2\delta_0 + \frac{1}{2}(\delta_2 + 2\delta_1\delta_0)\int_0^t g^2(\tau)d\tau$$
$$+ \delta_1 \int_0^t \left(g^2(\tau)\|\boldsymbol{s}(\boldsymbol{x},\tau)\|_2 + \|\boldsymbol{f}(\boldsymbol{x},\tau)\|_2\right)d\tau$$
$$+ \delta_0 \int_0^t \left(g^2(\tau)\|\nabla_{\boldsymbol{x}}\boldsymbol{s}(\boldsymbol{x},\tau)\|_F + \|\nabla_{\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x},\tau)\|_F\right)d\tau.$$

5. Training with score FPE-regularizer

We showed in Section 3.1 that score models learned via $\mathcal{J}_{\text{DSM}}$ (Eq. (3)) do not satisfy the score FPE, a property that ground truth scores should satisfy *a priori*. Motivated by this fact and Theorem 4.3, we hence devise a novel regularization term which is called *score FPE-regularizer* and defined as $\mathcal{R}_{\text{FP}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{FP}}(\boldsymbol{\theta};\alpha,\beta,\lambda_{\text{FP}}(\cdot),m) :=$

$$\mathbb{E}_{t\sim\mathcal{U}[0,T]}\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{\boldsymbol{x}(t)|\boldsymbol{x}(0)}\Big[\alpha \cdot \frac{1}{D^m}\|\lambda_{\text{FP}}(t)\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t)\|_2^m + \beta \cdot |\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t)|\Big].$$

Here, $\alpha,\beta \geq 0$ are parameters controlling the regularization strength, $\lambda_{\text{FP}}(\cdot)$ is the time-weighting function for the score FPE residual, and m is an integer. $\mathcal{R}_{\text{FP}}$ consists of the score FPE residual and time-derivative taming term, which respectively imitate Eq. (11) and Eq. (12). With the score FPE-regularizer, we propose a new loss $\mathcal{J}_{\text{FP}}$ which comprises $\mathcal{J}_{\text{DSM}}$ and $\mathcal{R}_{\text{FP}}$ with $\mathcal{J}_{\text{FP}}(\boldsymbol{\theta}) = \mathcal{J}_{\text{FP}}(\boldsymbol{\theta};\lambda(\cdot),\alpha,\beta,\lambda_{\text{FP}}(\cdot),m) :=$

$$\mathcal{J}_{\text{DSM}}(\boldsymbol{\theta};\lambda(\cdot)) + \mathcal{R}_{\text{FP}}(\boldsymbol{\theta};\alpha,\beta,\lambda_{\text{FP}}(\cdot),m), \tag{19}$$

We refer to a model trained with our proposed $\mathcal{J}_{\text{FP}}$ as *FP-Diffusion*. We remark that Eq. (19) returns the vanilla DSM loss (Eq. (3)) with $\alpha = \beta = 0.0$. Hereafter, we take $\lambda(\cdot) = g^2(\cdot)$ in $\mathcal{J}_{\text{DSM}}$.

Because $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}]$ in $\mathcal{R}_{\text{FP}}$ is generally expensive to calculate for high dimensional data, we propose efficient approximations for $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ and $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$.

Finite difference (Fornberg, 1988) for $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ can be efficiently approximated by finite difference method as the derivative is one-dimensional. For high dimensional datasets, we set $(h_s, h_d) = (0.001, 0.0005)$ and approximate $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$ by

$$\frac{h_s^2 \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t+h_d) + (h_d^2 - h_s^2)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - h_d^2 \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t-h_s)}{h_s h_d (h_s + h_d)}.$$

Hutchinson's estimator (Hutchinson, 1989) for $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$ Hutchinson's trace estimator stochastically estimates the trace of any square matrix. As $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) = \text{tr}\big(\nabla_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{\theta}}\big)$, we can apply Hutchinson's trick and replace the $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$ term with an estimation

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

where $\boldsymbol{v}_j \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. We set $M = 1$, following Song et al. (2020a) which works well in practice.

In Appendix C, we present supplementary information encompassing a theoretical analysis of the error involved in estimating the score FPE, as well as a potential technique that can enhance computational efficiency in this regard. Moreover, we supplement with runtime comparison in Appendix E.2.

6. Empirical implications of score FPE

In this section, we investigated two implications of the score FPE. First, we examined the solvability of scores through a Cauchy problem associated with the score FPE (as stated in Proposition 3.1). Second, we investigated how reducing the score FPE residual enhances the conservativity of a model (as described in Proposition 4.4).

6.1. Scores learning by solving Cauchy problems

Here, we consider the data distribution as a 2D GMM $\frac{1}{5}\mathcal{N}\big((-5,-5),\boldsymbol{I}\big) + \frac{4}{5}\mathcal{N}\big((5,5),\boldsymbol{I}\big)$. The diffusion process is taken as VE SDE (Eq. (22)). The ground truth score of a 2D GMM, denoted as $\boldsymbol{s}^{\text{GMM}}$, can be expressed explicitly in a closed form throughout the diffusion (as the diffusion process is linear in $\boldsymbol{x}$). In Section 3, we explained that the score at all times can theoretically be solved, given the score at a single time step. That is, the score is a solution $\tilde{\boldsymbol{s}}$ to the following Cauchy problem on the system of PDEs:

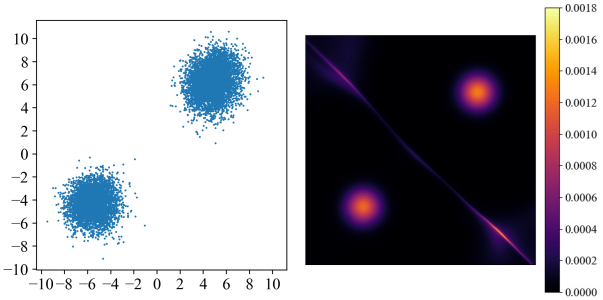
$$\begin{cases} \partial_t \tilde{\boldsymbol{s}}(\boldsymbol{x},t) = \nabla_{\boldsymbol{x}}\mathcal{L}[\tilde{\boldsymbol{s}}](\boldsymbol{x},t), & (\boldsymbol{x},t) \in \mathbb{R}^D \times (0,T] \\ \tilde{\boldsymbol{s}}(\boldsymbol{x},0) = \boldsymbol{s}^{\text{GMM}}(\boldsymbol{x},0), & \boldsymbol{x} \in \mathbb{R}^D, \end{cases} \tag{20}$$

where we recall $\mathcal{L}[\tilde{\boldsymbol{s}}] = \frac{1}{2}g^2\text{div}_{\boldsymbol{x}}(\tilde{\boldsymbol{s}}) + \frac{1}{2}g^2\|\tilde{\boldsymbol{s}}\|_2^2$. We fulfill this idea by parametrizing solutions of Eq. (20) via neural

networks $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$ (Raissi et al., 2019; Blechschmidt & Ernst, 2021) and learning an optimal $\boldsymbol{\theta}$ to minimize:

$$\mathbb{E}_{t\sim\mathcal{U}[0,T]}\mathbb{E}_{\boldsymbol{x}(0)}\mathbb{E}_{q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))}\left\|\boldsymbol{\epsilon}[\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}](\boldsymbol{x},t)\right\|_2 + \mathbb{E}_{\boldsymbol{x}(0)}\left\|\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}(\boldsymbol{x},0)-\boldsymbol{s}^{\text{GMM}}(\boldsymbol{x},0)\right\|_2. \tag{21}$$

Interestingly, as shown in Figures 2(a) and (b), respectively, $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$ generates satisfactory samples and enables good density estimation. This supports our argument that all temporal score information can be obtained by solving the score FPE. Generally, an initial condition to match the ground truth score is impractical. Nevertheless, this opens up the possibility of learning diffusion models from noisy data by substituting the exact score matching (with the ground truth) at the initial time with a noise-contaminated score matching (i.e., denoising score matching trick).



(a) Generated samples by $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$ (b) Estimated density by $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$

Figure 2. (a) visualizes instances generated by $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$. (b) shows the estimated density via the probability flow ODE of $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$. Scores at all times can be obtained by solving a Cauchy problem of the score FPE.

6.2. Reduction of score FPE residual implies conservativity

We take the 2D GMM described in Section 6.1 as the data distribution. It is known that a vector field $\boldsymbol{F} := (F_1, F_2, F_3)\colon \mathbb{R}^3 \to \mathbb{R}^3$ is conservative if and only if its "curl", $\left(\partial_{\boldsymbol{x}_2}F_3 - \partial_{\boldsymbol{x}_3}F_2, \partial_{\boldsymbol{x}_3}F_1 - \partial_{\boldsymbol{x}_1}F_3, \partial_{\boldsymbol{x}_1}F_2 - \partial_{\boldsymbol{x}_2}F_1\right)$, is zero. Thus, considering the mean squared error (MSE) of their curls quantifies the degree of conservativity of the score. We compared these values for the following four cases: scores trained (a) from Eq. (3), (b) from Eq. (19) with $(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m) = (0.001, 0.0, 1.0, 1)$, (c) and from Eq. (19) with $(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m) = (0.01, 0.0, 1.0, 1)$, along with (d) the ground truth score. Figure 3 plots the MSEs of the curls of the trained and ground truth scores for each timestep. The time-averaged MSEs of curls of the four scores are 2.22, 1.89, 0.60, and $3.73e-13$, respectively. We observed that the ground truth score is numerically conservative by its nature, and that scores trained with the score FPE-regularizer tend to be conservative, which empirically supports Proposition 4.4.

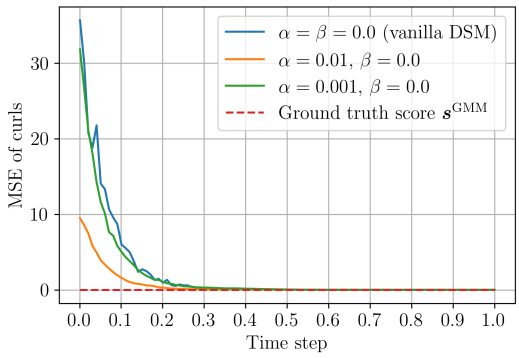


Figure 3. Comparison of the MSEs of curls. Scores trained with the score FPE-regularizer tend to be conservative.

7. Density Estimation Experiments

We examined the effectiveness of $\mathcal{J}_{\text{FP}}$ on three synthetic datasets, MNIST, Fashion MNIST, CIFAR-10, and ImageNet32. Appendix D gives the implementation details and Appendix E.3 visualizes randomly generated examples. We released our code at `https://github.com/sony/FP-Diffusion`.

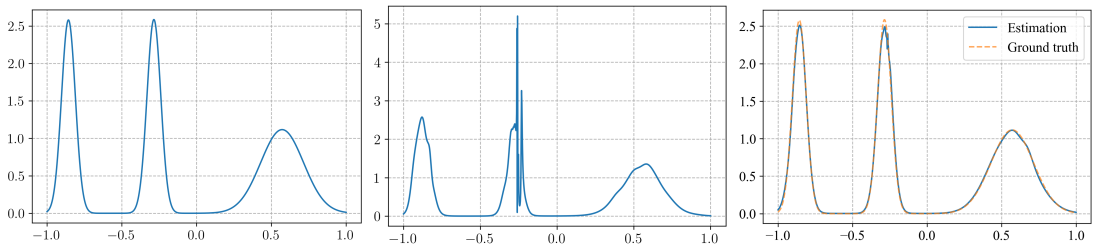
7.1. Synthetic datasets

We compared and visualized density estimation via models trained with vanilla $\mathcal{J}_{\text{DSM}}$ (Eq. (3)) and the proposed $\mathcal{J}_{\text{FP}}$ (Eq. (19)) with fixed $(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m) = (0.0015, 0.0, 1.0, 1)$. Here, the forward SDE is taken as a VE type. We examined the models' performance across three synthetic datasets: a 1D GMM with three modes $\frac{3}{10}\mathcal{N}\left(-\frac{6}{7}, (\frac{1}{70})^2\right) + \frac{3}{10}\mathcal{N}\left(-\frac{2}{7}, (\frac{1}{70})^2\right) + \frac{4}{10}\mathcal{N}\left(\frac{4}{7}, (\frac{1}{7})^2\right)$, a 2D checkerboard, Swiss rolls, and a 2D Gaussian mixture models (GMM) with eight modes whose means are located equidistant on the unit circle and with a standard deviation 1. We refer to Appendix D.1 for more details.

For all datasets, scores trained with the score FPE-regularizer, as shown in Figure 4(b) and Figures 5(b), (d), and (f), can approximate the data density well, with improvement over vanilla score matching, as shown in Figure 4(c) and Figures 5(a), (c), and (e). This reinforces the implication of Theorem 4.2 that the score FPE-regularizer may improve density estimation of the probability flow ODE, as it enforces a known self-consistency property of the ground truth score.

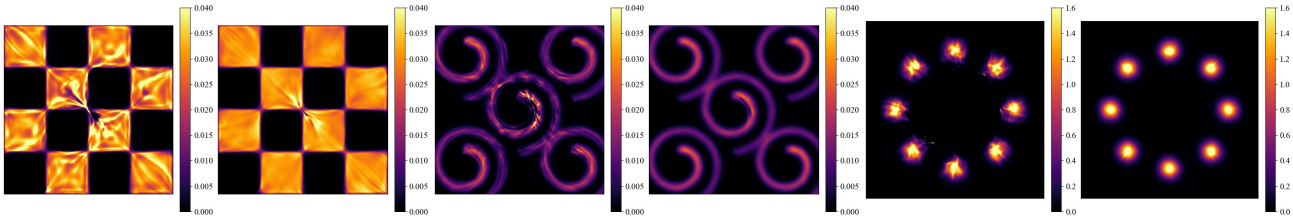
7.2. MNIST and Fashion MNIST

We trained models with the proposed $\mathcal{J}_{\text{FP}}$ on MNIST and Fashion MNIST with different α's values from scratch, and we evaluated the test set negative log-likelihood (NLL) in terms of bits/dim (bpd). In FP-Diffusion, the rest of parameters were fixed as $(\beta, \lambda_{\text{FP}}(\cdot), m) = (0.0, 1.0, 1)$. Table 1



(a) Data density (b) Vanilla DSM ($\alpha = \beta = 0.0$) (c) FP-Diffusion (with $\alpha = 0.0015$)

Figure 4. (a) demonstrates the ground truth data density. We compare (b) estimated density by probability flow ODE with $\boldsymbol{s_\theta}$ trained with $\alpha = \beta = 0.0$, and (c) with $\alpha = 0.0015$. Score FPE-regularizer improves density estimation.



(a) Vanilla DSM ($\alpha = \beta = 0.0$) (b) FP-Diffusion (with $\alpha = 0.0015$) (c) Vanilla DSM ($\alpha = \beta = 0.0$) (d) FP-Diffusion (with $\alpha = 0.0015$) (e) Vanilla DSM ($\alpha = \beta = 0.0$) (f) FP-Diffusion (with $\alpha = 0.0015$)

Figure 5. Estimated densities on a 2D checkerboard, multiple Swiss rolls, and eight GMMs, respectively. (a), (c), and (e) show estimated densities via the probability flow ODE of $\boldsymbol{s_\theta}$ trained with $\alpha = \beta = 0.0$ (vanilla DSM). In contrast, (b), (d), and (e) show the densities via training with $\alpha = 0.0015$. The score FPE-regularizer estimated the data densities well.

reports the averaged NLLs over five repeated runs of likelihood computations and three different initializations for training across two instantiations of the forward SDE (including VE, and VP) of various choices of α. A lower NLL indicates a better performance. We observed a general improvement in the NLL with $\alpha = 0.1$. To better understand the choice of the hyper-parameter α, ignoring dirt effects that result from different training initializations, we compare the NLLs via fine-tuning from pre-trained models in Appendix E.1.

Table 1. NLL comparisons on MNIST and Fashion MNIST

	MNIST		**Fashion MNIST**	
Method	VE	VP	VE	VP
Vanilla (Song et al., 2020b)	3.73	3.24	4.76	4.46
FP-Diffusion ($\alpha = 0.001$)	3.64	3.17	4.67	4.50
FP-Diffusion ($\alpha = 0.01$)	3.58	3.12	4.61	4.36
FP-Diffusion ($\alpha = 0.1$)	3.42	**2.98**	**4.40**	**4.21**
FP-Diffusion ($\alpha = 1.0$)	**3.30**	3.11	4.44	4.36
FP-Diffusion ($\alpha = 10.0$)	3.31	3.21	4.46	4.67

7.3. CIFAR-10 and ImageNet32

We fine tuned the pre-trained VE models from the checkpoints of Song et al. (2020b); Lu et al. (2022) by training them for 0.1M additional iterations on CIFAR-10 and ImageNet32, respectively. Here, we set the hyper-parameters of FP-Diffusion as $(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m) = (0.15, 0.01, g^2(\cdot), 2)$, but we also explored different choices as described in Appendix E.1. Table 2 reports the averaged NLLs of probability flow ODE on the test dataset over five repeated runs. Compared with vanilla DSM (Song et al., 2020b), FP-Diffusion significantly improved the NLL. Moreover, FP-Diffusion was competitive with higher-order DSM (Lu et al., 2022), where we re-computed the NLL based on their checkpoints but also indicated their reported results in the parentheses. On CIFAR-10, we noticed that FP-Diffusion trained with VE and VE-deep architectures may obtain inferior FID scores: 10.83 and 4.51, respectively; compared with FID scores of vanilla models: 3.33 and 2.44. However, the difference is generally imperceptible (see Appendix E.3 for illustration and quantitative measurements). Moreover, Appendix B.2 provides empirical evidence demonstrating that FP-Diffusion exhibits superior adherence to the score FPE compared to the vanilla model.

Table 2. NLL comparisons on CIFAR-10 and ImageNet32

	CIFAR-10		**ImageNet32**
Method	VE	VE-deep	VE
FP-Diffusion	**3.36**	3.32	**3.77**
Vanilla (Song et al., 2020b)	3.61 (3.66)	3.42	4.01 (4.21)
2nd DSM (Lu et al., 2022)	3.44	3.35	3.82 (4.06)
3rd DSM (Lu et al., 2022)	3.38	**3.31** (**3.27**)	3.80 (4.02)

8. Related work and discussions

Conservativity is a key property for understanding the consistency between the learned density and the ground truth

density, as the latter is inherently conservative. Despite the dominance of diffusion models over energy-based methods (which are inherently conservative), exploring conservativity can offer valuable insights into their underlying mechanisms and potentially enhance both types of models.

Salimans & Ho (2021) adopted a special parameterization to ensure conservativity and compared with energy-based models. In contrast, Chao et al. (2022) imposed a penalty to reach zero curl (i.e., conservativity), independently of the model architecture. In addition, they empirically showed that non-conservative scores may incur rotational vector fields tangent to the true score function, leading to inefficient updates during the sampling processes. However, how the non-conservativity affects the sampling theoretically and empirically is not well studied in the literature of diffusion models. Although enforcement of conservativity of diffusion models is not the main purpose of this work, conservativity is one of the outcomes by reducing the score FPE residual. Nevertheless, score FPE provides a different framework from the PDE perspective, which is theoretically solid and may stimulate further study.

On the other hand, researchers have also attempted to theoretically explain the success of diffusion models by studying the gap between the data and learned densities. De Bortoli et al. (2021) proved error bounds for these densities in terms of the total variation. Song et al. (2021) showed the likelihood of the diffusion model can be bounded by the score matching objective with a specific choice of temporal weighting. Chen et al. (2022) provided a convergence analysis for any data distribution with second-order moment in KL divergence. Kwon et al. (2022) found that minimization of the score matching loss may implicitly reduce the Wasserstein-2 distance between the data and learned density. Meng et al. (2021a) introduced the concept of estimating higher-order gradients of a data distribution. Later, Lu et al. (2022) extended the idea and showed that the likelihood from the deterministic trajectory of a diffusion model may be improved by matching higher-order scores.

Shen et al. (2022) showed that the asymptotic fixed point of the velocity field associated with the classic FPE (Fokker, 1914; Planck, 1917) (governing the density evolution) can recover the solution of FPE in the Wasserstein-2 sense. However, its study was neither adapted to diffusion models nor generative models.

9. Conclusion

We introduce the score FPE and theoretically study its relationship with likelihood improvements, conservativity, higher-order score matching, and scores induced by a parametric reverse diffusion. Moreover, we propose to regularize models by enforcing consistency properties of the ground truth score through the score FPE, and show this achieves better density estimation and likelihoods on various datasets. We empirically support our theory by finding that reduction of the score FPE residual improves the conservativity of a model. The Cauchy problem defined with the score FPE can be used to obtain time-conditioned scores directly by PDEs solving. Incorporating more advanced numerical methods for solving PDEs is an interesting avenue for future research.

Acknowledgements

We would like to thank Lucas Mauch for their variable comments during the preparation of this manuscript. Additionally, we sincerely appreciate anonymous reviewers for their insightful feedback and suggestions.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Artstein, Z. Continuous dependence on parameters: On the best possible results. *Journal of Differential Equations*, 19(2):214–225, 1975.

Blechschmidt, J. and Ernst, O. G. Three ways to solve partial differential equations with neural networks—a review. *GAMM-Mitteilungen*, 44(2):e202100006, 2021.

Chao, C.-H., Sun, W.-F., Cheng, B.-W., and Lee, C.-Y. Quasi-conservative score-based generative models. *arXiv preprint arXiv:2209.12753*, 2022.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Cheuk, K. W., Sawata, R., Uesaka, T., Murata, N., Takahashi, N., Takahashi, S., Herremans, D., and Mitsufuji, Y. Diffroll: Diffusion-based generative music transcription with unsupervised pretraining capability. *arXiv preprint arXiv:2210.05148*, 2022.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Evans, L. C. and Garzepy, R. F. *Measure theory and fine properties of functions*. Routledge, 2018.

Fokker, A. D. Die mittlere energie rotierender elektrischer dipole im strahlungsfeld. *Annalen der Physik*, 348(5): 810–820, 1914.

Fornberg, B. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of computation*, 51 (184):699–706, 1988.

Gronwall, T. H. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, pp. 292–296, 1919.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pp. 2722–2730. PMLR, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.

Kim, D., Kim, Y., Kang, W., and Moon, I.-C. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Kwon, D., Fan, Y., and Lee, K. Score-based generative modeling secretly minimizes the wasserstein distance. *arXiv preprint arXiv:2212.06359*, 2022.

Lu, C., Zheng, K., Bao, F., Chen, J., Li, C., and Zhu, J. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International Conference on Machine Learning*, pp. 14429–14460. PMLR, 2022.

Lunardi, A. *Analytic semigroups and optimal regularity in parabolic problems*. Springer Science & Business Media, 2012.

Masry, E. and Rice, J. A. Gaussian deconvolution via differentiation. *Canadian Journal of Statistics*, 20(1):9–21, 1992.

Meng, C., Song, Y., Li, W., and Ermon, S. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34: 25359–25369, 2021a.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021b.

Murata, N., Saito, K., Lai, C.-H., Takida, Y., Uesaka, T., Mitsufuji, Y., and Ermon, S. Gibbsddrm: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration, 2023.

Øksendal, B. Stochastic differential equations. In *Stochastic differential equations*, pp. 65–84. Springer, 2003.

Papageorgiou, N. On the solution set of nonlinear evolution inclusions depending on a parameter. 1994.

Pidstrigach, J. Score-based generative models detect manifolds. *arXiv preprint arXiv:2206.01018*, 2022.

Planck, V. Über einen satz der statistischen dynamik und seine erweiterung in der quantentheorie. *Sitzungberichte der*, 1917.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Roosta-Khorasani, F. and Ascher, U. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Saito, K., Murata, N., Uesaka, T., Lai, C.-H., Takida, Y., Fukui, T., and Mitsufuji, Y. Unsupervised vocal dereverberation with diffusion-based generative models. *arXiv preprint arXiv:2211.04124*, 2022.

Salimans, T. and Ho, J. Should ebms model the energy or the score? In *Energy Based Models Workshop-ICLR 2021*, 2021.

Shen, Z., Wang, Z., Kale, S., Ribeiro, A., Karbasi, A., and Hassani, H. Self-consistency of the fokker planck equation. In *Conference on Learning Theory*, pp. 817–841. PMLR, 2022.

Skorski, M. Modern analysis of hutchinson's trace estimator. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–5. IEEE, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Contents

A. Instantiations of forward SDEs and corresponding score FPEs

B. How scores satisfy score FPEs

B.1. Supportive experiments for Section 3
B.2. FP-Diffusion enforces satisfaction of score FPE

C. More details on techniques for efficient score FPE computation

C.1. Trick to reduce computation cost of $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$
C.2. Trick to reduce computation cost of $\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$
C.3. Error analysis of score FPE estimation
C.4. Potential technique to compute $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}]$ more efficiently

D. Implementation details

D.1. Synthetic dataset
D.2. MNIST and Fashion MNIST
D.3. CIFAR-10 and ImageNet32

E. Supplemental results

E.1. Sensitivity to hyper-parameters
E.2. Runtime discussion
E.3. Illustration and quality of generated samples

F. Theoretical assumptions

G. Proofs and discussions

G.1. Proof of Proposition 3.1
G.2. Proof of Theorem 4.2
G.3. Discussions on Theorem 4.2
G.4. Proof of Theorem 4.3
G.5. Proof and discussion of Proposition 4.4
G.6. Proof of Proposition 4.5
G.7. Proof of Proposition 4.6
G.8. Proof of Proposition C.2

A. Instantiations of forward SDEs and corresponding score FPEs

Song et al. (2020b) categorizes the forward SDE into three types based on the behavior of the variance during evolution. Here, we focus on two types: the Variance Explosion (VE) SDE and Variance Preserving (VP) SDE.

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

$$d\boldsymbol{x}(t) = \sqrt{\frac{d\sigma^2(t)}{dt}} d\boldsymbol{w}_t. \tag{22}$$

A typical instance of a VE SDE is Score Matching of Langevin Dynamics (SMLD) (Song & Ermon, 2019), 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 $(\sigma_{\text{min}}, \sigma_{\text{max}}) := (0.01, 50)$.

VP SDE Let β be a non-negative function of t. A VP SDE has a linear drift term $\boldsymbol{f}(\boldsymbol{x}, t) = -\frac{1}{2}\beta(t)\boldsymbol{x}$ and a 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 a VP SDE is Denoising Diffusion Probabilistic Modeling (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020), 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.

Table 3 summarizes the aforementioned SDE instantiations and their associated score FPEs.

Table 3. Summary of forward SDEs and their score FPEs

	VE SDE	**VP SDE**
$\boldsymbol{f}(\boldsymbol{x}, t)$	$\boldsymbol{0}$	$-\frac{1}{2}\beta(t)\boldsymbol{x}$
$g(t)$	$\sigma_{\text{min}}\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)^t\sqrt{2\log\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right)}$	$\sqrt{\beta(t)}$
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	$\partial_t \boldsymbol{s} = \nabla_{\boldsymbol{x}}\left[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \frac{1}{2}g^2(t)\,\lVert\boldsymbol{s}\rVert_2^2\right]$	$\partial_t \boldsymbol{s} = \frac{1}{2}\beta(t)\nabla_{\boldsymbol{x}}\left[\text{div}_{\boldsymbol{x}}(\boldsymbol{s}) + \lVert\boldsymbol{s}\rVert_2^2 + \langle\boldsymbol{x}, \boldsymbol{s}\rangle\right]$

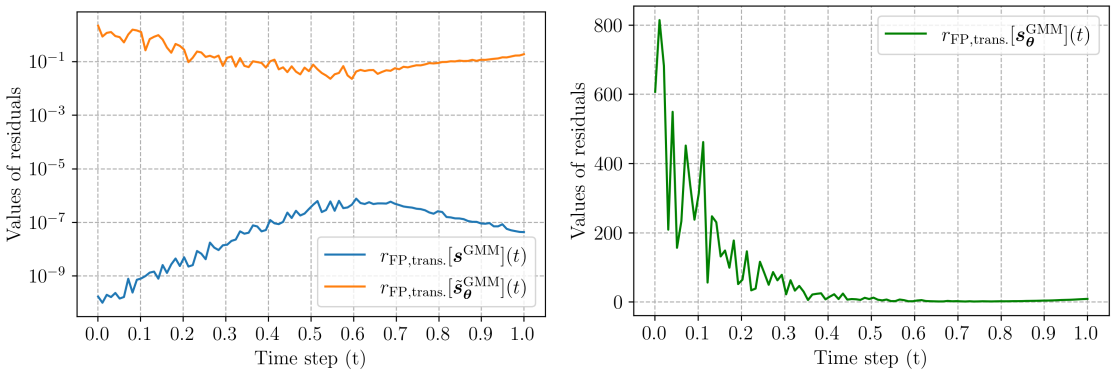
B. How scores satisfy score FPEs

B.1. Supportive experiments for Section 3

In this section, we further demonstrate how score functions should satisfy the score FPE empirically. We treat the data distribution as a 2D GMM $\frac{1}{5}\mathcal{N}\big((-5, -5), \boldsymbol{I}\big) + \frac{4}{5}\mathcal{N}\big((5, 5), \boldsymbol{I}\big)$ as in Section 6.1 and use the same notations. The diffusion process is taken as a VE SDE (Eq. (22)).

We examine whether $\boldsymbol{s}^{\text{GMM}}$ satisfies the score FPE by computing $r_{\text{FP}}[\boldsymbol{s}^{\text{GMM}}](t)$. Figure 6(a) shows its residual as a function of time (blue curve) and supplements with the time residual of $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$, obtained by solving Eq. (21). The score FPE residual of the ground truth is almost zero, which empirically supports Proposition 3.1.

In addition, Figure 6(b) shows the computed residual of the score FPE as a function of time for a score $\boldsymbol{s}_{\theta}^{\text{GMM}}$ learned by DSM (Eq. (3)). We observed that $\boldsymbol{s}_{\theta}^{\text{GMM}}$ also does not satisfy the score FPE. This phenomenon matches with the results shown in Figure 1 for realistic datsets.



(a) FP residuals of the ground truth score and the score learned from Eq. (21)

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

Figure 6. Comparison of the score FPE residuals of $\boldsymbol{s}^{\text{GMM}}$, $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$ and $\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{GMM}}$ for a 2D GMM. (a) shows that both the (closed-form) ground truth score $\boldsymbol{s}^{\text{GMM}}$ and the score $\tilde{\boldsymbol{s}}_{\boldsymbol{\theta}}^{\text{GMM}}$ obtained by solving the score FPE (Eq. (20)) numerically satisfy the score FPE. On the other hand, (b) provides further evidence that $\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{GMM}}$, which is learned from DSM, does not satisfy the score FPE.

B.2. FP-Diffusion enforces satisfaction of score FPE

In this section, we compared the score FPE residuals as a function of time (i.e., $r_{\text{FP, trans.}}[\boldsymbol{s}_{\boldsymbol{\theta}}](t)$) between a vanilla DSM model (trained for 0.1M additional iterations) and FP-Diffusion respectively trained on CIFAR-10 with various $\alpha \in \{1.0, 0.5, 0.15\}$ and a fixed $(\beta, \lambda_{\text{FP}}(\cdot), m) = (0.01, g^2(\cdot), 2)$. The forward SDE was taken as the VE type. For demonstration purpose, we only plotted the residuals at representative timesteps $\{10^{-5}, 0.1, 0.2, \cdots, 1.0\}$ in Figure 7 and recorded their corresponding numerical values in Table 4. We remark that $\alpha = 0.15$ achieves the best NLL reported in Table 2. Thanks to the score FPE-regularizer, FP-Diffusion generally obtains smaller residuals compared with the vanilla model (i.e., scores satisfy the score FPE better).

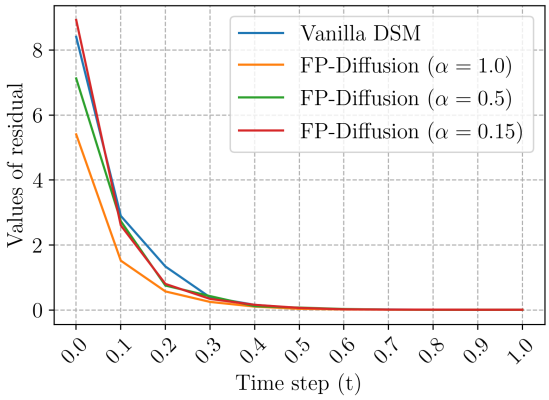


Figure 7. Comparison of the score FPE residuals as a function of time of vanilla DSM model and FP-Diffusion trained with $\alpha = 1.0$, 0.5, and 0.15, respectively.

Table 4. Values of score FPE residuals as a function of time of vanilla DSM model and FP-Diffusion trained with $\alpha = 1.0$, 0.5, and 0.15, respectively. For each timestep, we mark the largest value in bold.

Model/Time	10^{-5}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Vanilla (Song et al., 2020b)	8.401	**2.889**	**1.331**	0.383	**0.152**	**0.067**	**0.023**	**0.008**	**0.005**	**0.002**	**0.003**
FP-Diffusion ($\alpha = 1.0$)	5.399	1.512	0.565	0.242	0.097	0.030	0.009	0.003	0.001	0.000	0.001
FP-Diffusion ($\alpha = 0.5$)	7.121	2.728	0.745	**0.421**	0.110	0.046	0.015	0.005	0.002	0.001	0.000
FP-Diffusion ($\alpha = 0.15$)	**8.922**	2.597	0.796	0.335	0.132	0.058	0.014	0.006	0.004	0.001	0.001

C. More details on techniques for efficient score FPE computation

As explained in Section 5, the computation of $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t)$ in $\mathcal{R}_{\text{FP}}(\boldsymbol{\theta})$ is generally expensive; hence, we applied two techniques, the finite difference trick and Hutchinson's trace estimator, to replace the expensive computations of certain components in $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t)$.

C.1. Trick to reduce computation cost of $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$

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

Lemma C.1. *(Fornberg, 1988) Let $\alpha\colon [0,1] \to \mathbb{R}^D$ be a vector-valued function that is continuously differentiable up to third order derivatives. Let h_s and h_d be step-size hyper-parameters. 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

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

In implementation for a high-dimensional dataset, we consider $\alpha(\cdot) := \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \cdot)$; hence, $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ is approximated as

$$\frac{h_s^2 \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t+h_d) + (h_d^2 - h_s^2)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - h_d^2 \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t-h_s)}{h_s h_d (h_s + h_d)},$$

where we set $(h_s, h_d) = (0.001, 0.0005)$.

C.2. Trick to reduce computation cost of $\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$

Hutchinson's trace estimator (Hutchinson, 1989) stochastically estimates the trace $\mathrm{tr}(\boldsymbol{A})$ of any square matrix $\boldsymbol{A}$. The idea is to 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, $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}\mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\boldsymbol{v}\boldsymbol{v}^T]) = \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\mathrm{tr}(\boldsymbol{A}\boldsymbol{v}\boldsymbol{v}^T)] = \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\mathrm{tr}(\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}^T)] = \mathbb{E}_{\boldsymbol{v}\sim p_{\boldsymbol{v}}}[\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}^T]$. By i.i.d. sampling $\{\boldsymbol{v}_j\}_{j=1}^M$ from $p_{\boldsymbol{v}}$, we can use an unbiased estimator

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

to estimate $\mathrm{tr}(\boldsymbol{A})$. Note that $\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)) = \mathrm{tr}\big(\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}\big)$. Thus, we can apply Hutchinson's trick and replace the $\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}})$ term with the following 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 implementation, $p_{\boldsymbol{v}}$ is usually taken as a standard normal distribution or a Rademacher distribution.

We set $M = 1$ in our implementation.

C.3. Error analysis of score FPE estimation

In this section, we provide a theoretical error analysis of score FPE by using finite difference approximation and Hutchinson's estimation.

Here we consider the case of taking the Rademacher distribution for Hutchinson's estimation. The proof is differed to Appendix G.8. As for the case of normal distribution, we may utilize a statistical bound in (Roosta-Khorasani & Ascher, 2015) and obtain a similar statistical estimate by following the identical argument.

Proposition C.2. *Let $\boldsymbol{s}(\boldsymbol{x}, t)$ be a vector field on $\mathbb{R}^D \times [t_0, T]$, where $t_0 > 0$ and g be a continuous function defined on $[t_0, T]$ which achieves minimum at $t^* \in [t_0, T]$ and that $g^* := g^2(t^*) > 0$. Denote*

$$\mathscr{V}[\boldsymbol{s}] := \partial_t \boldsymbol{s} - \frac{1}{2} g^2 \mathrm{tr}(\nabla \boldsymbol{s})$$

and

$$\hat{\mathscr{V}}[\boldsymbol{s}] := \mathrm{FD}(\boldsymbol{s}) - \frac{1}{2} g^2 \mathrm{tr}_{H^{(M)}}(\nabla \boldsymbol{s}).$$

Here we denote a finite difference approximation in t with parameters $h_s, h_d > 0$ as $\mathrm{FD}(\boldsymbol{s})$, and a Hutchinson's estimator with M samples from Rademacher distribution as $\mathrm{tr}_{H^{(M)}}(\nabla \boldsymbol{s})$ (see Section 5 for their explicit definitions). Let $\mathscr{T}[\boldsymbol{s}] := \partial_t \boldsymbol{s} - \mathrm{FD}(\boldsymbol{s})$, $\mathscr{H}[\boldsymbol{s}] := \frac{1}{2} g^2 \big(\mathrm{tr}(\nabla \boldsymbol{s}) - \mathrm{tr}_{H^{(M)}}(\nabla \boldsymbol{s})\big)$ and $\mathscr{E}[\boldsymbol{s}] := \mathscr{V}[\boldsymbol{s}] - \hat{\mathscr{V}}[\boldsymbol{s}]$. Notice from Lemma C.1 that there is a constant $C > 0$ so that $\|\mathscr{F}[\boldsymbol{s}]\|_D < CDh$, where $h = \frac{h_s^2 h_d + h_d h_s^2}{h_s + h_d}$ and $\|\cdot\|_D$ indicates the ℓ_D-norm. For any $\epsilon \in (0, \frac{3}{8} g^*)$, if $h \in (0, \frac{\epsilon}{2CD})$, then

$$\mathbb{P}\big(\|\mathscr{E}[\boldsymbol{s}]\|_D < \epsilon\big) \geq 1 - \exp^{-\left(\frac{M\epsilon^2}{2(g^* - \frac{8}{3}\epsilon)}\right)}.$$

C.4. Potential technique to compute $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}]$ more efficiently

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

$$\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t) = \underbrace{\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}}_{\text{(I)}} - \underbrace{\nabla_{\boldsymbol{x}}\Big[\frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(t) \|\boldsymbol{s}_{\boldsymbol{\theta}}\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big]}_{\text{(II)}} \tag{23}$$

The use of automatic differentiation to compute the gradient in $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t)$ (part (II) in Eq. (23)) is generally cumbersome for high dimensional data. We thus propose to use random projection to replace the gradient computation (multi-dimensional) with a directional derivative (one-dimensional). Then, we can apply the finite difference trick introduced in above to further reduce the computation effort. We first recall a fundamental property before rigorously formulating the technique.

Lemma C.3. *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 $D_{\boldsymbol{v}} M(\boldsymbol{x}, t)$ denotes the directional derivative of M in $\boldsymbol{x}$ along the direction $\boldsymbol{v}$ and is defined as follows:

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

For simplicity, we let $M(\boldsymbol{x}, t) := \frac{1}{2} g^2(t) \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2} g^2(t) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})$ and let $\boldsymbol{v} \in \mathbb{R}^D$ be an arbitrary vector. We project $\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t)$ along direction $\boldsymbol{v}$ and apply Lemma C.3:

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

Note that both $\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}$ and $\frac{d}{dh} M(\boldsymbol{x} + h\boldsymbol{v}, t)\big|_{h=0}$ entail one-dimensional differentiation and can be estimated via Lemma C.1, thus avoiding automatic differentiation. That is, we may have the estimation

$$\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t) \approx \mathbb{E}_{\boldsymbol{v} \sim p_{\boldsymbol{v}}} \langle \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t), \boldsymbol{v} \rangle, \tag{24}$$

where $p_{\boldsymbol{v}}$ is the distribution of a random vector $\boldsymbol{v} \in \mathbb{R}^D$. However, the performance may be degraded by using the estimate in Eq. (24) possibly because of the inaccurate approximation of the exact score FPE. Hence, we will need further study on lowering the computation costs while preventing performance degradation.

D. Implementation details

In this section, we describe the details of our implementation on synthetic dataset, MNIST/Fashion MNIST, and CIFAR-10/ImageNet32.

D.1. Synthetic dataset

We conducted our experiments on 4 NVIDIA GeForce RTX 3090 GPUs.

2D GMM. For all experiments of the 2D GMM $\frac{1}{5}\mathcal{N}\big((-5, -5), \boldsymbol{I}\big) + \frac{4}{5}\mathcal{N}\big((5, 5), \boldsymbol{I}\big)$, we exploited a network structure similar to the one in a particular repository[2] for all . In that repository, we modified the forward SDE modified to be a VE SDE or VP SDE (see Appendix A), but we simply replaced all convolutional layers with fully connected layers. We trained for $2,000$ epochs with a learning rate of 10^{-3} and a batch size of 500.

Checkerboard, Swiss rolls, eight-mode 2D GMM, and 1D GMM. The neural network setups for the results shown in Figures 5 and 4 were the same as the toy model structures provided in the repository of Lu et al. (2022) [3]. We show our detailed data preparation below, as modified from the same repository. We trained the models for 0.1M iterations with a learning rate of 10^{-3} and a batch size of 500. For both training and inference, the start time was 10^{-3}.

Listing 1. Checkerboard dataset

```
import numpy
import torch
x1 = np.random.rand(batch_size) * 4 - 2
x2_ = np.random.rand(batch_size) - np.random.randint(0, 2, batch_size) * 2
x2 = x2_ + (np.floor(x1) % 2)
checkerboard = torch.from_numpy(np.concatenate([x1[:, None], x2[:, None]], 1).float() * 2
```

[2] https://colab.research.google.com/drive/120kYYBOVa1i0TD85RjlEkFjaWDxSFUx3?usp=sharing
[3] https://github.com/LuChengTHU/mle_score_ode

Listing 2. Swiss rolls dataset

```
import numpy
import torch
import sklearn
data = sklearn.datasets.make_swiss_roll(n_samples=batch_size, noise=1.0)[0]
data = data.astype("float32")[:, [0, 2]]
data /= 4.
data = torch.from_numpy(data).float()
r = 4.5
data1 = data.clone() + torch.tensor([-r, -r])
data2 = data.clone() + torch.tensor([-r, r])
data3 = data.clone() + torch.tensor([r, -r])
data4 = data.clone() + torch.tensor([r, r])
swiss_roll = torch.cat([data, data1, data2, data3, data4], axis=0)
```

Listing 3. 8 modes 2D GMM dataset

```
import numpy
import torch
num_mixture = 8
radius = 1.0
sigma = 0.1
mix_probs = [1/num_mixture] * num_mixture
std = torch.stack([torch.ones(dim) * sigma for i in range(len(mix_probs))], dim=0)
mix_probs = torch.tensor(mix_probs)
mix_idx = torch.multinomial(mix_probs, n, replacement=True)
thetas = np.linspace(0, 2 * np.pi, num_mixture, endpoint=False)
xs = radius * np.sin(thetas, dtype=np.float32)
ys = radius * np.cos(thetas, dtype=np.float32)
center = np.vstack([xs, ys]).T
center = torch.tensor(centers)
centers = centers[mix_idx]
stds = std[mix_idx]
eight_GMM = torch.randn_like(centers) * stds + centers
```

D.2. MNIST and Fashion MNIST

We conducted our experiments on 4 NVIDIA GeForce RTX 3090 GPUs.

We trained score networks on MNIST and Fashion MNIST from scratch for 200 epochs with a learning rate of 10^{-3} and batch size of 32 by using the setup as in the repository[4], with the forward SDE modified to be a VE SDE or VP SDE. For both training and inference, the start time was 10^{-3}.

D.3. CIFAR-10 and ImageNet32

For CIFAR-10 and ImageNet32, we followed the same model architectures and experimental setups as in Song et al. (2020b)[5] and Lu et al. (2022)[6], respectively. More precisely, we used NCSN++ cont. for the VE and NCSN++ cont. deep for the VE-deep. We conducted our experiments on 4 NVIDIA A100 GPUs (40 GiB). The batch size was fixed as 48. Instead of training from scratch, we ued the pre-trained VE models provided by the two repositories and fine tuned them by training for 0.1M additional iterations. As we have found that a smaller batch size may decrease the NLL of the probability flow ODE, for a fair comparison, we also trained the vanilla DSM models for 0.1M additional iterations. Table 2 reports the results.

We used uniform dequantization (Ho et al., 2019) for likelihood evaluation. To reduce the variance, we computed the NLL (in bpd) over five repeated runs and took their average. For both training and inference, we chose a start time of 10^{-5}.

[4] https://colab.research.google.com/drive/120kYYBOVa1i0TD85RjlEkFjaWDxSFUx3?usp=sharing
[5] https://github.com/yang-song/score_sde_pytorch
[6] https://github.com/LuChengTHU/mle_score_ode/

E. Supplemental results

E.1. Sensitivity to hyper-parameters

Fine-tuning pre-trained models on MNIST and Fashion MNIST. In Section 7.2, we trained models from scratch on MNIST and Fashion MNIST, and compared vanilla DSM models with FP-Diffusion with various α's. However, different initializations and optimization dynamics may lead to variances of NLLs. To avoid this issue, we followed our training strategy for CIFAR-10 and ImageNet32 by fine tuning pre-trained models instead. More precisely, we first trained vanilla DSM models for 100 epochs and fine tuned with different α's for additional 100 epochs. Table 5 shows the NLL comparisons, where we can observe general improvements with FP-Diffusion.

Table 5. NLL comparisons on MNIST and Fashion MNIST (trained by fine-tuning)

	MNIST		**Fashion MNIST**	
Method	VE	VP	VE	VP
Vanilla (Song et al., 2020b)	3.67	3.27	4.78	4.49
FP-Diffusion ($\alpha = 0.001$)	3.63	3.15	4.61	4.50
FP-Diffusion ($\alpha = 0.01$)	3.60	3.10	4.58	4.46
FP-Diffusion ($\alpha = 0.1$)	3.47	**3.01**	4.42	**4.23**
FP-Diffusion ($\alpha = 1.0$)	**3.23**	3.09	**4.39**	4.27
FP-Diffusion ($\alpha = 10.0$)	3.30	3.17	4.41	4.43

Fine-tuning pre-trained models on CIFAR-10. We compared the proposed FP-Diffusion with the VE SDE trained on CIFAR-10 with different hyper-parameter choices. Table 6 reports the test set NLL results, which were computed by averaging over five repeated runs to reduce variances. We found that $(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m) = (0.15, 0.01, g^2(\cdot), 2)$ generally works well on more complicated datasets such as CIFAR-10 and ImageNet32. We remark that the choice of β makes the scale of $|\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t)|$ in Eq. (19) be comparable with $\mathcal{J}_{\text{DSM}}$, where $\beta \approx 10^{-2}$ is generally a reasonable value for both CIFAR-10 and ImageNet32. Additionally, we motivate the choice of $\lambda_{\text{FP}}(\cdot)$ as $g^2(\cdot)$ in Appendix G.3.

General recipe for searching hyper-parameters. A general recipe for the selection of hyper-parameters is (1) to set α be at the scale of 10^{-1}; (2) may need β (roughly at the scale of 10^{-2}) for more complicated datasets. However. hyper-parameters to obtain the best results may depend on the optimization and the structure of datasets as the landscapes of loss functions are totally different. We generally observed that a larger α (roughly at the scale of 10^0) is preferable for relatively sparse datasets such as MNIST and Fashion MNIST. In contrast, a smaller α (roughly at the scale of 10^{-1}) and β (roughly at the scale of 10^{-2}) is preferable for more complicated datasets such as CIFAR-10 and ImageNet32. As for synthetic datasets, $(\alpha, \beta) \approx (10^{-2}, 0.0)$ works well. Nevertheless, FP-Diffusion overall improves NLLs against the vanilla training.

Table 6. FP-Diffusion with different of hyper-parameter choices for the VE SDE trained on CIFAR-10.

$(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m)$	**NLL (bpd) on CIFAR-10**
Vanilla DSM $\alpha = \beta = 0.0$	3.61
$(0.15, 0.01, g^2(\cdot), 2)$	**3.36**
$(1.0, 0.01, g^2(\cdot), 2)$	3.40
$(0.5, 0.01, g^2(\cdot), 2)$	3.38
$(0.2, 0.01, g^2(\cdot), 2)$	3.37
$(0.1, 0.01, g^2(\cdot), 2)$	3.37
$(0.0, 0.01, *, *)$	3.37
$(1.0, 0.0, g^2(\cdot), 2)$	3.57

E.2. Runtime discussion

In this section, we compared the runtime of vanilla diffusion model (trained with $\mathcal{J}_{\text{DSM}}$) and the proposed FP-Diffusion on CIFAR-10. We fixed the training batch size as 48 and examined their runtime on VE type model NCSN++ cont. with PyTorch. The hardware was 4 NVIDIA A100 GPUs (40 GiB). We believe the computation time of FP-Diffusion can be improved with a more optimized code and setup of the environment.

Table 7. Runtime comparison of vanilla diffusion and FP-Diffusion trained on CIFAR-10. The forward SDE was taken as the VE type.

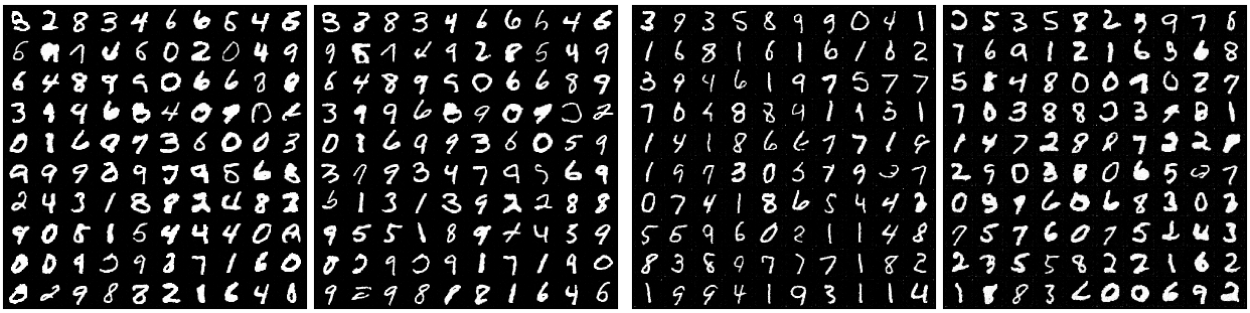
Method	**Time per iteration (sec)**	**Memory (GiB)**
Vanilla (Song et al., 2020b)	0.17	23.48
FP-Diffusion $(\alpha, \beta, \lambda_{\text{FP}}(\cdot), m) = (0.15, 0.01, g^2(\cdot), 2)$	2.08	49.01

E.3. Illustration and quality of generated samples

We visualized randomly generated samples with models trained on MNIST, Fashion MNIST, and CIFAR-10 in Figures 8, 9, and 10, respectively. In Table 8, we reported numerical results of sample quality with models trained on CIFAR-10. Even though FP-Diffusion has inferior numerical measurements compared with vanilla models, the differences are imperceptible by comparing their generated samples.

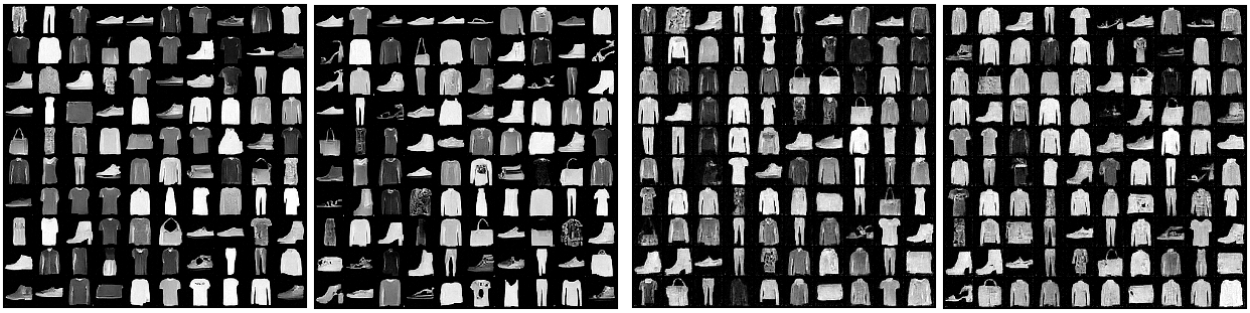
Table 8. Sample quality on CIFAR-10.

	FID ↓		**IS ↑**	
Method	VE	VE-deep	VE	VE-deep
Vanilla (Song et al., 2020b)	**3.33**	**2.44**	**9.19**	**9.80**
FP-Diffusion	10.83	3.33	8.88	9.14



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

Figure 8. Samples generated with models trained on MNIST by using (a, c) vanilla DSM and (b, d) FP-Diffusion with the setup described in Section 7.2.



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

Figure 9. Samples generated with models trained on Fashion MNIST by using (a, c) vanilla DSM and (b, d) FP-Diffusion with the setup described in Section 7.2.

F. Theoretical assumptions

Here, we introduce some regularity conditions to establish Theorems .4.2 and 4.3 which are commonly used in theoretical studies of score-based models (Song et al., 2021; Lu et al., 2022; Pidstrigach, 2022; Kwon et al., 2022).

(a) VE (vanilla $\alpha = \beta = 0.0$) on CIFAR-10 (b) VE (FP-Diffusion) on CIFAR-10 (c) VE-deep (vanilla $\alpha = \beta = 0.0$) on CIFAR-10 (d) VE-deep (FP-Diffusion) on CIFAR-10

Figure 10. Illustration of generated samples with VE and VE-deep models trained on CIFAR-10. (a) and (c) show samples generated by vanilla DSM. (b) and (d) show samples generated by FP-Diffusion with the setup described in Section 7.3.

Assumption F.1. We assume there are finite constants $L > 0$, which is sufficiently large (may assume $L \geq 1$), and $\delta_T > 0$ such that the following conditions hold for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$ and $t \in [0, T]$

(a) Bounded 2nd non-central moment: $\mathbb{E}_{q_0(\boldsymbol{x})}[\|\boldsymbol{x}\|_2^2] \leq L$, or $\sup_{t\in[0,T]} \left\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\|\boldsymbol{x}\|_2^2]\right\} \leq L$ to streamline the proof,

(b) $\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\|_2 \leq L(1 + \|\boldsymbol{x}\|_2)$,

(c) $\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{y}, t)\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$,

(d) $\|\boldsymbol{f}(\boldsymbol{x}, t)\|_2 \leq L(1 + \|\boldsymbol{x}\|_2)$,

(e) $\|\boldsymbol{f}(\boldsymbol{x}, t) - \boldsymbol{f}(\boldsymbol{y}, t)\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$,

(f) $\left\|\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t)\right\|_2 \leq L(1 + \|\boldsymbol{x}\|_2)$,

(g) $\left\|\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{y}, t)\right\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$,

(h) $\|\boldsymbol{s}(\boldsymbol{x}, t)\|_2 \leq L(1 + \|\boldsymbol{x}\|_2)$,

(i) $\|\boldsymbol{s}(\boldsymbol{x}, t) - \boldsymbol{s}(\boldsymbol{y}, t)\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$,

and that

(j) $\sup_{t\in[0,T]} \left\{\mathbb{E}_{q_t(\boldsymbol{x})}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, T)\right\|_2^2\right]\right\} \leq \delta_T^2$, or $\sup_{\boldsymbol{x}\in\mathbb{R}^D} \left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, T)\right\|_2^2 \leq \delta_T^2$,

(k) For any $t \in [0, T]$, there is a $k > 0$ so that as $\|x\|_2 \to \infty$, $q_t(\boldsymbol{x}) = \mathcal{O}\big(e^{-\|x\|_2^k}\big)$ and $p_t^{\text{ODE}}(\boldsymbol{x}) = \mathcal{O}\big(e^{-\|x\|_2^k}\big)$.

Assumption F.2. We assume there is a finite constant $L > 0$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$ and $t \in [0, T]$ the following conditions hold

(c') $\|\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{y}, t)\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$,

(e') $\|\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{y}, t)\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$,

(g') $\left\|\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) - \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{y}, t)\right\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2$.

G. Proofs and discussions

G.1. Proof of Proposition 3.1

Proof. We prove the result with a more general forward SDE

$$d\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}, t)dt + \boldsymbol{G}(\boldsymbol{x}, t)d\boldsymbol{w}_t, \tag{25}$$

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 (Øksendal, 2003)

$$\partial_t q_t(\boldsymbol{x}) = -\sum_{j=1}^{D} \partial_{x_j}\big(\tilde{\boldsymbol{F}}_j(\boldsymbol{x}, t)q_t(\boldsymbol{x})\big), \tag{26}$$

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

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

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 G.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

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

G.2. Proof of Theorem 4.2

Lemma G.2 (Grönwall's inequality (Gronwall, 1919))**.** *Assume that α, β, and u are continuous functions on $[0, T]$. If β is non-negative on $[0, T]$ and if u satisfies the integral inequality*

$$u(t) \le \alpha(t) + \int_t^T \beta(\tau)u(\tau)d\tau, \qquad \textit{for all } t \in [0, T]$$

then

$$u(t) \le \alpha(t) + \int_t^T \alpha(\tau)\beta(\tau)\exp\big(\int_t^\tau \beta(r)dr\big)d\tau, \qquad \textit{for all } t \in [0,T]$$

In particularly, if α is non-decreasing (especially, a constant independent of t), then

$$u(t) \le \alpha(t)\exp\big(\int_t^T \beta(\tau)d\tau\big), \qquad \textit{for all } t \in [0,T].$$

Proof. Grönwall's inequality

Consider the function

$$v(\tau) := \exp\Big(-\int_\tau^T \beta(r)dr\Big)\int_\tau^T \beta(r)u(r)dr.$$

Taking the derivative by the product rule leads to

$$\begin{aligned} v'(\tau) &= \Big(-u(\tau) + \int_\tau^T \beta(r)u(r)dr\Big)\beta(\tau)\exp\Big(-\int_\tau^T \beta(r)dr\Big) \\ &\ge -\alpha(\tau)\beta(\tau)\exp\Big(-\int_\tau^T \beta(r)dr\Big). \end{aligned}$$

Integrating the above inequality from $\tau = t$ to $\tau = T$ proves the statement. □

Proof. Theorem 4.2

We first prove the Ineq. (14). Notice that we can rearrange $\mathcal{J}_{\text{Diff}}$ as

$$\begin{aligned} \mathcal{J}_{\text{Diff}}(\boldsymbol{\theta}) &= \frac{1}{2}\int_0^T g^2(t)\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}\Big[\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x})\big)^\top\big(\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)\big)\Big]dt \\ &= \int_0^T\int_{\mathbb{R}^D}\Big[g(t)\sqrt{\frac{q_t(\boldsymbol{x})}{2}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x})\big)\Big]^\top\Big[g(t)\sqrt{\frac{q_t(\boldsymbol{x})}{2}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)\big)\Big]dtd\boldsymbol{x}. \end{aligned}$$

The claim is established by applying Cauchy-Schwartz inequality to functions $g(t)\sqrt{\frac{q_t(\boldsymbol{x})}{2}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x})\big)$ and $g(t)\sqrt{\frac{q_t(\boldsymbol{x})}{2}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)\big)$.

Now, we prove the Ineq. (15), in which we just need to consider the case when $M(\boldsymbol{\theta}) := \sup_{t\in[0,T]}\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}\big[\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2\, d\tau\big] < \infty$; otherwise, the result holds obviously. Throughout the proof, we simply use the notation $\lesssim$ to express $\lesssim_{T,\delta_T,g,L}$, which indicates the estimation depends only on T, δ_T, g, L.

Recall that the probability flow ODE (Song et al., 2020b) associated to Eq. (4) is defined as

$$\frac{d\boldsymbol{x}}{dt}(t) = \boldsymbol{f}(\boldsymbol{x}(t),t) - \frac{1}{2}g^2(t)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}(t),t).$$

By the special case of FPE (Eq. (26)) with zero diffusion term, we obtain the PDE characterizes the evolution of $p_{t,\boldsymbol{\theta}}^{\text{ODE}}$

$$\frac{\partial p_{t,\boldsymbol{\theta}}^{\text{ODE}}}{\partial t}(\boldsymbol{x},t) = \text{div}_{\boldsymbol{x}}\Big(\big(\frac{1}{2}g^2(t)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{f}(\boldsymbol{x},t)\big)p_{t,\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x})\Big)$$

Hence,

$$\begin{aligned} \frac{\partial \log p_{t,\boldsymbol{\theta}}^{\text{ODE}}}{\partial t} &= \frac{1}{p_{t,\boldsymbol{\theta}}^{\text{ODE}}}\frac{\partial p_{t,\boldsymbol{\theta}}^{\text{ODE}}}{\partial t} \\ &= \frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) - \text{div}_{\boldsymbol{x}}(\boldsymbol{f}) + \langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \frac{1}{2}g^2(t)\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{f}\rangle, \end{aligned}$$

where we apply the product rule of divergence in the last equality. After taking the gradient from the both sides, we obtain[7]

$$\begin{aligned}\frac{\partial \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}}{\partial t}(\boldsymbol{x},t) &= \nabla_{\boldsymbol{x}} \frac{\partial \log p_{t,\boldsymbol{\theta}}^{\text{ODE}}}{\partial t} \\ &= \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) - \text{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] + \nabla_{\boldsymbol{x}}\Big[\langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \frac{1}{2}g^2(t)\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{f}\rangle\Big] \\ &= \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) - \text{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] + \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\rangle\Big]\end{aligned} \tag{27}$$

By rearranging Eq. (9) and combining with Eq. (27), it results in

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

That is,

$$\partial_t\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\big) = \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t) + \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\rangle\Big] \tag{28}$$

Fix a $t \in [0,T]$. We integrate both sides of the above equation from $\tau = T$ to $\tau = t$

$$\begin{aligned}\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t) &= \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},T) \\ &\quad + \int_T^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)d\tau + \int_T^t \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\rangle\Big]d\tau.\end{aligned}$$

Applying the ℓ_2-norm

$$\begin{aligned}\big\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\big\|_2 &\le \big\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},T)\big\|_2 \\ &\quad + \int_t^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2\, d\tau \\ &\quad + \int_t^T \Big\|\nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\rangle\Big]\Big\|_2 d\tau.\end{aligned} \tag{29}$$

In the last term, we may compute $\nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\rangle\Big]$ as

$$\begin{aligned}&\frac{1}{2}g^2(\tau)\Big(\nabla_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{\theta}}\cdot\boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\cdot\boldsymbol{s}_{\boldsymbol{\theta}}\Big) + \frac{1}{2}g^2(\tau)\Big(\nabla_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{\theta}}\cdot\big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\big)\Big) \\ &\qquad - \nabla_{\boldsymbol{x}}\boldsymbol{f}\cdot\big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\big) - \nabla_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{\theta}}\cdot\boldsymbol{f} + \nabla_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\cdot\boldsymbol{f}\end{aligned} \tag{30}$$

[7]Indeed, Eq. (27) can also be derived from Proposition 3.1.

Hence, we can further estimate the last term of Ineq. (29) as

$$
\begin{aligned}
&\int_t^T \left\| \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \rangle \Big] \right\|_2 d\tau \\
\leq & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau + \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \cdot \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau \\
+ & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \big) \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{f} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \big) \right\|_2 d\tau \\
+ & \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \boldsymbol{f} \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \cdot \boldsymbol{f} \right\|_2 d\tau \\
\leq & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau + \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau \\
+ & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{f} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau \\
+ & \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_{\mathrm{op}} \left\| \boldsymbol{f} \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_{\mathrm{op}} \left\| \boldsymbol{f} \right\|_2 d\tau \\
\leq & \left[L^2 \big(\int_0^T g^2(\tau) d\tau \big) (1 + \|\boldsymbol{x}\|_2) \right] + \left[\int_t^T \big(\frac{L}{2} g^2(\tau) + L \big) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau \right] + \left[2L^2 T (1 + \|\boldsymbol{x}\|_2) \right] \\
\leq & \, C_1(L, T, g)(1 + \|\boldsymbol{x}\|_2) + \int_t^T \big(\frac{L}{2} g^2(\tau) + L \big) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau
\end{aligned}
$$

where $\|\boldsymbol{A}\|_{\mathrm{op}} := \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$ denotes the operator norm of the matrix $\boldsymbol{A}$. In the second-to-last inequality, we apply Assumption F.1 together with the Rademacher's theorem (Evans & Garzepy, 2018) which bounds the total differentiations of $\boldsymbol{s}_{\boldsymbol{\theta}}$, $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}$, and $\boldsymbol{f}$ by their Lipschitz constants. Moreover, we summarize constant terms into $C_1 := C_1(L, T, g)$, which depends on L, T, and the function g.

Combining this estimation with Ineq. (29), we have

$$
\begin{aligned}
\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, t) \right\|_2 \leq & \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, T) \right\|_2 + \int_0^T \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2 d\tau + C_1(L, T, g)(1 + \|\boldsymbol{x}\|_2) \\
& + \int_t^T \big(\frac{L}{2} g^2(\tau) + L \big) \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, \tau) \right\|_2 d\tau.
\end{aligned}
$$

Consider the following functions in Lemma G.2

$$
\begin{aligned}
u(t) &:= \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, t) \right\|_2 \\
\alpha(t) &:= \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, T) \right\|_2 + \int_0^T \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2 d\tau + C_1(L, T, g)(1 + \|\boldsymbol{x}\|_2) \\
\beta(t) &:= \frac{L}{2} g^2(t) + L.
\end{aligned}
$$

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

$$
\begin{aligned}
u(t) &\leq \alpha \exp\big(\int_t^T \beta(\tau) d\tau \big) \\
&\lesssim \Big[\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, T) \right\|_2 + \int_0^T \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2 d\tau + (1 + \|\boldsymbol{x}\|_2) \Big],
\end{aligned}
$$

where we bound $\exp\big(\int_t^T \beta(\tau) d\tau \big)$ by $\exp\big(\int_0^T \beta(\tau) d\tau \big)$ which is a constant, and we absorb all constant terms.

We are going to square both sides of the above estimation and take the expectation over $q_t(\boldsymbol{x})$. For the sake of simplicity, we denote $e_{\boldsymbol{\theta}}(\boldsymbol{x}) := \int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2 \, d\tau$ and $\delta_{\boldsymbol{\theta}}(\boldsymbol{x}) := \left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},T)\right\|_2$, and hence, we obtain

$$\begin{aligned}
\mathbb{E}_{q_t(\boldsymbol{x})}\left[u^2(t)\right] &\lesssim \mathbb{E}_{q_t(\boldsymbol{x})}\left(\delta_{\boldsymbol{\theta}}(\boldsymbol{x}) + e_{\boldsymbol{\theta}}(\boldsymbol{x}) + (1+\|\boldsymbol{x}\|_2)\right)^2 \\
&\lesssim \Bigg\{\mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right] + \mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right] + \mathbb{E}_{q_t(\boldsymbol{x})}\left[(1+\|\boldsymbol{x}\|_2)^2\right] \\
&\quad + \mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})e_{\boldsymbol{\theta}}(\boldsymbol{x})\right] + \mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})(1+\|\boldsymbol{x}\|_2)\right] + \mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}(\boldsymbol{x})(1+\|\boldsymbol{x}\|_2)\right]\Bigg\}.
\end{aligned} \tag{31}$$

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

$$\begin{aligned}
\mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})e_{\boldsymbol{\theta}}(\boldsymbol{x})\right] &\le \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]} \\
\mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})(1+\|\boldsymbol{x}\|_2)\right] &\le \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[\delta_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[(1+\|\boldsymbol{x}\|_2)^2\right]} \\
\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}(\boldsymbol{x})(1+\|\boldsymbol{x}\|_2)\right] &\le \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[(1+\|\boldsymbol{x}\|_2)^2\right]}.
\end{aligned}$$

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

$$\sup_{t\in[0,T]}\left\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\|\boldsymbol{x}\|_2]\right\} \le L \quad \text{and} \quad \sup_{t\in[0,T]}\left\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\|\boldsymbol{x}\|_2^2]\right\} \le L \tag{32}$$

as by Cauchy Schwartz inequality that $\mathbb{E}_{\boldsymbol{x}\sim q_0(\boldsymbol{x})}[\|\boldsymbol{x}\|_2] \le \left(\mathbb{E}_{\boldsymbol{x}\sim q_0(\boldsymbol{x})}[\|\boldsymbol{x}\|_2^2]\right)^{1/2} \le L^{1/2}$ and the transition density $q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$ has bounded covariance matrices as a function in $t \in [0,T]$. With Ineq. (32) and Assumption F.1.(j), Ineq. (31) becomes

$$\begin{aligned}
&\mathbb{E}_{q_t(\boldsymbol{x})}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\right\|_2^2\right] \\
&\lesssim \Bigg\{\delta_T^2 + \mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right] + (1+3L) \\
&\quad + \delta_T\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]} + \delta_T\sqrt{1+3L} + \sqrt{1+3L}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]}\Bigg\} \\
&\lesssim \left(\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right] + \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\left[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\right]} + C_1(L,T,g,\delta_T)\right) \\
&\lesssim \left(M(\boldsymbol{\theta}) + \sqrt{M(\boldsymbol{\theta})} + C_1(L,T,g,\delta_T)\right).
\end{aligned}$$

Here we use that $e_{\boldsymbol{\theta}}^2(\boldsymbol{x}) = \left(\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2 \, d\tau\right)^2 \le T\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\|_2^2 \, d\tau$. Again, we we abuse of the notation and summarize constants into $C_1 = C_1(L,T,g,\delta_T)$. Therefore, after combining the Ineq. (14) and the estimation above, we obtain (with a fusion of constant term)

$$\begin{aligned}
\left(\mathcal{J}_{\text{Diff}}(\boldsymbol{\theta})\right)^2 &\lesssim \mathcal{J}_{\text{SM}}(\boldsymbol{\theta}) \cdot \mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) \\
&\lesssim \mathcal{J}_{\text{SM}}(\boldsymbol{\theta}) \cdot \left(M(\boldsymbol{\theta}) + \sqrt{M(\boldsymbol{\theta})} + C_1(L,T,g,\delta_T)\right)
\end{aligned}$$

□

G.3. Discussions on Theorem 4.2

Tighter bounds of Theorem 4.2. We remark that one can easily extend Theorem 4.2 and obtain a sharper bound by checking carefully the tightness of each estimation. We provide an approach as an instance. Let us assume there is a constant

$\delta_{\text{ODE}} > 0$ to control the distance between $\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}$ and $\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}$ instead (in this case, we do not require Assumption F.1.(g)). That is,

$$\sup_{\mathbb{R}^D \times [0,T]} \left\| \nabla_{\boldsymbol{x}} \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) \right\|_2 \leq \delta_{\text{ODE}}. \tag{33}$$

Notice that Eq. (30) can be rewritten as

$$\begin{aligned} \frac{1}{2} g^2(\tau) \Big(\nabla_{\boldsymbol{x}} \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) \cdot \boldsymbol{s}_{\boldsymbol{\theta}} \Big) + \frac{1}{2} g^2(\tau) \Big(\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) \Big) \\ - \nabla_{\boldsymbol{x}} \boldsymbol{f} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) - \nabla_{\boldsymbol{x}} \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) \cdot \boldsymbol{f}. \end{aligned}$$

Following the same argument as the proof of Theorem. 4.2 together with the help of Ineq. (33), $\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta})$ can be upper bounded by a constant which depends monotonically increasingly on δ_{ODE}. Therefore, we can get a sharper estimation if δ_{ODE} is smaller.

Variants of Theorem 4.2. $M(\boldsymbol{\theta})$ describes one of the worst case scenarios by searching the largest time-averaged residuals among all time slices. Theoretically it is a legit quantity. While we can obtain similar results as Theorems 4.2 and 4.3 (however, need additional assumptions) with more interpretable bounds. Here we just focus on the discussion on Theorem 4.2, and a similar argument can be adopted to Theorem 4.3.

First, we introduce a variant of $M(\boldsymbol{\theta})$ which is defined as

$$\widetilde{M}(\boldsymbol{\theta}) := \sup_{t \in [t_0, T]} \mathbb{E}_{\boldsymbol{x} \sim q_t(\boldsymbol{x})} \left[\int_{t_0}^{T} q_\tau(\boldsymbol{x}) \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2^2 d\tau \right]. \tag{34}$$

$\widetilde{M}(\boldsymbol{\theta})$ is more interpretable in the sense that for any time slice t and sample $\boldsymbol{x}_t \sim q_t(\boldsymbol{x})$, if τ is deviating from t when taking the time average, the marginal density $q_\tau(\boldsymbol{x}_t)$ can reduce the effect of the score FPE residual at that by providing a lighter weight.

Notice that if we assume that there is positive constants Q^* and q^* so that $q^* \leq q_t(\boldsymbol{x}) \leq Q^*$ for all $\boldsymbol{x} \in \text{supp}(q_t)$ and $t \in [t_0, T]$, then $q^* M(\boldsymbol{\theta}) \leq \widetilde{M}(\boldsymbol{\theta}) \leq Q^* M(\boldsymbol{\theta})$.

Theorem G.3 (A variant of Theorem 4.2)**.** *Let $t_0 > 0$ be a constant indicating the terminal time of the backward diffusion process (or equivalently the initial time of the forward process). Suppose in addition to Assumption F.1 that we have bounded 4th non-central moment: $\mathbb{E}_{q_0(\boldsymbol{x})}[\|\boldsymbol{x}\|_2^4] \leq L$. If there is positive constants Q^* and q^* so that $Q^* \geq q_t(\boldsymbol{x}) \geq q^*$ for all $\boldsymbol{x} \in \text{supp}(q_t)$ and $t \in [t_0, T]$. Then we can obtain an upper bound for $\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta})$ in terms of $\widetilde{M}(\boldsymbol{\theta})$ as*

$$\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) \lesssim \widetilde{M}(\boldsymbol{\theta}) + \sqrt{\widetilde{M}(\boldsymbol{\theta})} + C_3,$$

where $C_3 > 0$ is a constant independent of $\boldsymbol{\theta}$ and different from C_1 and C_2.

Proof. Theorem G.3

We start from Eq. (28) in the proof of Theorem 4.2.

$$\partial_t \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) \big) = \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t) + \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big].$$

Multiplying $q_t(\boldsymbol{x})$ from the both sides and integrating from $\tau = T$ to $\tau = t$, we get the following equation with integration by part

$$\begin{aligned} \text{LHS} &:= \int_T^t \partial_\tau \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, \tau) \big) q_\tau(\boldsymbol{x}) \, d\tau \\ &= \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, \tau) \big) q_\tau(\boldsymbol{x}) \Big|_{\tau=T}^{\tau=t} - \int_T^t \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, \tau) \big) \partial_\tau q_\tau(\boldsymbol{x}) \, d\tau \\ &= \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) \big) q_t(\boldsymbol{x}) - \Big[\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, T) \big) q_T(\boldsymbol{x}) + \int_T^t \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, \tau) \big) \partial_\tau q_\tau(\boldsymbol{x}) \, d\tau \Big] \end{aligned}$$

and

$$\text{RHS} := \int_T^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau) q_\tau(\boldsymbol{x})\, d\tau + \int_T^t \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big] q_\tau(\boldsymbol{x})\, d\tau.$$

Hence, we have

$$\begin{aligned}\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\big) q_t(\boldsymbol{x}) = &\Big[\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},T)\big) q_T(\boldsymbol{x}) + \int_T^t \big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},\tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},\tau)\big) \partial_\tau q_\tau(\boldsymbol{x}) d\tau\Big] \\ &+ \int_T^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau) q_\tau(\boldsymbol{x}) d\tau + \int_T^t \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big] q_\tau(\boldsymbol{x}) d\tau.\end{aligned}$$

Applying the ℓ_2-norm and using the fact that $Q^* \geq q_t(\boldsymbol{x}) \geq q^*$ for all $\boldsymbol{x} \in \text{supp}(q_t)$ and $t \in [t_0, T]$ (or alternatively, we may choose sufficiently large $L > 0$ so that $L \geq q_t(\boldsymbol{x}) \geq 1/L$),

$$\begin{aligned}q_t(\boldsymbol{x}) \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t) \right\|_2 \lesssim &\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},T) \right\|_2 + \int_T^t \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},\tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},\tau) \right\| \left| \partial_\tau q_\tau(\boldsymbol{x}) \right| d\tau \\ &+ \int_t^T q_\tau(\boldsymbol{x}) \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau) \right\|_2 d\tau \\ &+ \int_t^T \left\| \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big] \right\|_2 d\tau.\end{aligned} \tag{35}$$

With the classic Fokker-Planck equation and $\boldsymbol{s}(\boldsymbol{x},\tau) = \nabla_{\boldsymbol{x}} q_\tau(\boldsymbol{x}) / q_\tau(\boldsymbol{x})$

$$\partial_\tau q_\tau(\boldsymbol{x}) = -\big[(\text{div}_{\boldsymbol{x}}(\boldsymbol{f}) - \frac{1}{2} g^2(\tau) \text{div}_{\boldsymbol{x}}(\boldsymbol{s})) q_\tau + \langle \boldsymbol{f} - \frac{1}{2} g^2(\tau) \boldsymbol{s}, \boldsymbol{s} \rangle q_\tau\big],$$

we can bound the term $\int_T^t \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},\tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},\tau) \right\| \left| \partial_\tau q_\tau(\boldsymbol{x}) \right| d\tau \leq C_3 (1 + \|\boldsymbol{x}\|_2)^2$, where $C_3 := C_3(L, T, g, Q^*, q^*) > 0$ is a constant depending on L, T, Q^*, q^*, and the function g.

In the last term of Ineq. (29), we may compute $\nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big]$ as

$$\begin{aligned}\frac{1}{2} g^2(\tau) \Big(\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \cdot \boldsymbol{s}_{\boldsymbol{\theta}} \Big) + \frac{1}{2} g^2(\tau) \Big(\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) \Big) \\ - \nabla_{\boldsymbol{x}} \boldsymbol{f} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \big) - \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \boldsymbol{f} + \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \cdot \boldsymbol{f}\end{aligned}$$

Hence, we can further estimate the last term of Ineq. (35) as

$$
\begin{aligned}
&\int_t^T \left\| \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \rangle \Big] \right\|_2 d\tau \\
\leq & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau + \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \cdot \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau \\
+ & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \big) \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{f} \cdot \big(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \big) \right\|_2 d\tau \\
+ & \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \cdot \boldsymbol{f} \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \cdot \boldsymbol{f} \right\|_2 d\tau \\
\leq & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau + \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 d\tau \\
+ & \int_t^T \frac{1}{2} g^2(\tau) \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{f} \right\|_{\mathrm{op}} \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau \\
+ & \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_{\mathrm{op}} \left\| \boldsymbol{f} \right\|_2 d\tau + \int_t^T \left\| \nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_{\mathrm{op}} \left\| \boldsymbol{f} \right\|_2 d\tau \\
\leq & \left[L^2 \big(\int_0^T g^2(\tau) d\tau \big) (1 + \|\boldsymbol{x}\|_2) \right] + \left[\int_t^T \big(\frac{L}{2} g^2(\tau) + L \big) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau \right] + \left[2L^2 T (1 + \|\boldsymbol{x}\|_2) \right] \\
\leq & \, C_3(L, T, g)(1 + \|\boldsymbol{x}\|_2) + \int_t^T \big(\frac{L}{2} g^2(\tau) + L \big) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}} \right\|_2 d\tau
\end{aligned}
$$

where $\|\boldsymbol{A}\|_{\mathrm{op}} := \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$ denotes the operator norm of the matrix $\boldsymbol{A}$. In the second-to-last inequality, we apply Assumption F.1 together with the Rademacher's theorem which bounds the total differentiations of $\boldsymbol{s}_{\boldsymbol{\theta}}$, $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}$, and $\boldsymbol{f}$ by their Lipschitz constants. Moreover, we summarize constant terms into $C_3 := C_3(L, T, g, Q^*, q^*)$.

By dividing both side of Ineq. (35) with $q_t(\boldsymbol{x})$ and using that $\frac{1}{Q^*} \leq \frac{1}{q_t(\boldsymbol{x})} \leq \frac{1}{q^*}$ for all $t \in [t_0, T]$ and $\boldsymbol{x} \in \mathrm{supp}(q_t)$, we combine above estimations which leads to

$$
\begin{aligned}
\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, t) \right\|_2 \leq & \frac{1}{q^*} \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, T) \right\|_2 + \frac{1}{q^*} \int_{t_0}^T \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2 d\tau \\
& + C_3(L, T, g, q^*) [(1 + \|\boldsymbol{x}\|_2) + (1 + \|\boldsymbol{x}\|_2)^2] \\
& + \int_t^T \frac{1}{q^*} \big(\frac{L}{2} g^2(\tau) + L \big) \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, \tau) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, \tau) \right\|_2 d\tau.
\end{aligned}
$$

Consider the following functions in the Lemma G.2 (Grönwall's inequality)

$$
\begin{aligned}
u(t) &:= \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, t) \right\|_2 \\
\alpha(t) &:= \frac{1}{q^*} \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, T) \right\|_2 + \frac{1}{q^*} \int_{t_0}^T q_\tau(\boldsymbol{x}) \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2 d\tau \\
&\quad + C_3(L, T, g, q^*) [(1 + \|\boldsymbol{x}\|_2) + (1 + \|\boldsymbol{x}\|_2)^2] \\
\beta(t) &:= \frac{1}{q^*} \big(\frac{L}{2} g^2(t) + L \big).
\end{aligned}
$$

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

$$
\begin{aligned}
u(t) &\leq \alpha \exp\big(\int_t^T \beta(\tau) d\tau \big) \\
&\lesssim \Big[\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{ODE}}(\boldsymbol{x}, T) \right\|_2 + \int_{t_0}^T q_\tau(\boldsymbol{x}) \left\| \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right\|_2 d\tau + (1 + \|\boldsymbol{x}\|_2) + (1 + \|\boldsymbol{x}\|_2)^2 \Big],
\end{aligned}
$$

where we bound $\exp\big(\int_t^T \beta(\tau)d\tau\big)$ by $\exp\big(\int_{t_0}^T \beta(\tau)d\tau\big)$ which is a constant, and we absorb all constant terms.

We are going to square both sides of the above estimation and take the expectation over $q_t(\boldsymbol{x})$. For the sake of simplicity, we denote $e_{\boldsymbol{\theta}}(\boldsymbol{x}) := \int_{t_0}^T q_\tau(\boldsymbol{x}) \left\|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\right\|_2 d\tau$ and $\delta_{\boldsymbol{\theta}}(\boldsymbol{x}) := \left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},T)\right\|_2$, and hence, we obtain

$$
\begin{aligned}
\mathbb{E}_{q_t(\boldsymbol{x})}\big[u^2(t)\big] &\lesssim \mathbb{E}_{q_t(\boldsymbol{x})}\left(\delta_{\boldsymbol{\theta}}(\boldsymbol{x}) + e_{\boldsymbol{\theta}}(\boldsymbol{x}) + \big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)\right)^2 \\
&\lesssim \Bigg\{\mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big] + \mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big] + \mathbb{E}_{q_t(\boldsymbol{x})}\Big[\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)^2\Big] \\
&\quad + \mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})e_{\boldsymbol{\theta}}(\boldsymbol{x})\big] + \mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)\big] + \mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}(\boldsymbol{x})\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)\big]\Bigg\}.
\end{aligned} \tag{36}
$$

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

$$
\begin{aligned}
\mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})e_{\boldsymbol{\theta}}(\boldsymbol{x})\big] &\le \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]} \\
\mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}(\boldsymbol{x})\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)\big] &\le \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[\delta_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)^2\big]} \\
\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}(\boldsymbol{x})\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)\big] &\le \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[\big(2 + 3\left\|\boldsymbol{x}\right\|_2 + \left\|\boldsymbol{x}\right\|_2^2\big)^2\big]}.
\end{aligned}
$$

It is noticed that bounded 4th non-central moment $\mathbb{E}_{q_0(\boldsymbol{x})}[\left\|\boldsymbol{x}\right\|_2^4] \le L$ indeed implies the following estimation which bounds all lower non-central moments for all $t \in [t_0, T]$

$$
\sup_{t\in[t_0,T]}\Big\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\left\|\boldsymbol{x}\right\|_2]\Big\}, \quad \sup_{t\in[t_0,T]}\Big\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\left\|\boldsymbol{x}\right\|_2^2]\Big\}, \quad \sup_{t\in[t_0,T]}\Big\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\left\|\boldsymbol{x}\right\|_2^3]\Big\}, \quad \sup_{t\in[t_0,T]}\Big\{\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}[\left\|\boldsymbol{x}\right\|_2^4]\Big\} \le L \tag{37}
$$

as by Cauchy Schwartz inequality and the transition density $q_{0t}(\boldsymbol{x}(t)|\boldsymbol{x}(0))$ has bounded covariance matrices as a function in $t \in [t_0, T]$. With Ineq. (37) and Assumption F.1.(j), Ineq. (36) becomes

$$
\begin{aligned}
&\mathbb{E}_{q_t(\boldsymbol{x})}\Big[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\right\|_2^2\Big] \\
&\lesssim \Bigg\{\delta_T^2 + \mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big] + (4 + 32L) \\
&\quad + \delta_T\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]} + \delta_T\sqrt{4+32L} + \sqrt{4+32L}\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}\Bigg\} \\
&\lesssim \Big(\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big] + \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]} + C_3(L,T,g,\delta_T)\Big) \\
&\lesssim \Big(\widetilde{M}(\boldsymbol{\theta})) + \sqrt{\widetilde{M}(\boldsymbol{\theta})} + C_3(L,T,g,q^*,\delta_T)\Big),
\end{aligned}
$$

Again, we we abuse of the notation and summarize constants into $C_3 = C_3(L,T,g,Q^*,q^*,\delta_T)$. Therefore, after combining the Ineq. (14) and the estimation above, we obtain (with a fusion of constant term)

$$
\begin{aligned}
\Big(\mathcal{J}_{\text{Diff}}(\boldsymbol{\theta})\Big)^2 &\lesssim \mathcal{J}_{\text{SM}}(\boldsymbol{\theta}) \cdot \mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) \\
&\lesssim \mathcal{J}_{\text{SM}}(\boldsymbol{\theta}) \cdot \Big(\widetilde{M}(\boldsymbol{\theta})) + \sqrt{\widetilde{M}(\boldsymbol{\theta})} + C_3(L,T,g,Q^*,q^*,\delta_T)\Big)
\end{aligned}
$$

□

Tighter bounds of Theorem G.3. Indeed, we can obtain tighter upper bound from Ineq. (38). Let us revisit the argument in Ineq. (38):

$$
\mathbb{E}_{q_t(\boldsymbol{x})}\Big[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\right\|_2^2\Big] \lesssim \Big(\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big] + \sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]} + C_3(L,T,g,Q^*,q^*,\delta_T)\Big).
$$

This implies

$$\begin{aligned}\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) &= \frac{1}{2}\int_{t_0}^{T} g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\Big[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)-\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\right\|_2^2\Big]dt \\ &\lesssim \int_{t_0}^{T} g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]dt + \int_{t_0}^{T} g^2(t)\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}dt + C_3(L,T,g,Q^*,q^*,\delta_T).\end{aligned}$$

By applying Cauchy-Schwartz inequality,

$$\begin{aligned}\int_{t_0}^{T} g^2(t)\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}dt &= \int_{t_0}^{T}\Big(g(t)\sqrt{\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]}\Big)\cdot g(t)dt \\ &\le \Big(\int_{t_0}^{T} g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]dt\Big)^{1/2}\Big(\int_{t_0}^{T} g^2(t)dt\Big)^{1/2}.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) &:= \frac{1}{2}\int_{t_0}^{T} g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\Big[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t)-\boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\right\|_2^2\Big]dt \\ &\lesssim \int_{t_0}^{T} g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]dt + \|g\|_{L^2([0,T])}\sqrt{\int_{t_0}^{T} g^2(t)\mathbb{E}_{q_t(\boldsymbol{x})}\big[e_{\boldsymbol{\theta}}^2(\boldsymbol{x})\big]dt} + C_3(L,T,g,Q^*,q^*,\delta_T). \qquad (38)\end{aligned}$$

We further define

$$\widetilde{\widetilde{M}}(\boldsymbol{\theta}) := \int_{t_0}^{T} g^2(t)\mathbb{E}_{\boldsymbol{x}\sim q_t(\boldsymbol{x})}\left[\int_{t_0}^{T} q_\tau(\boldsymbol{x})\left\|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},\tau)\right\|_2^2 d\tau\right]dt.$$

Then Ineq. (38) can be rewritten as

$$\mathcal{J}_{\text{Fisher}}(\boldsymbol{\theta}) \lesssim \widetilde{\widetilde{M}}(\boldsymbol{\theta}) + \|g\|_{L^2([0,T])}\sqrt{\widetilde{\widetilde{M}}(\boldsymbol{\theta})} + C_3(L,T,g,Q^*,q^*,\delta_T).$$

Indeed, $g^2(\cdot)$ in $\widetilde{\widetilde{M}}$ motivates the choice of time-weighting function $\lambda_{\text{FP}}(\cdot)$ as $g^2(\cdot)$ in the score FPE-regularizer for the training of more complicated datasets such as CIFAR-10 and ImageNet32.

Let $g_{\max} := \max_{t\in[t_0,T]} g^2(t)$. At last, we summarize a relation between all bounds $\widetilde{\widetilde{M}}(\boldsymbol{\theta})$, $\widetilde{M}(\boldsymbol{\theta})$, and $M(\boldsymbol{\theta})$:

$$\widetilde{\widetilde{M}}(\boldsymbol{\theta}) \le T g_{\max}\cdot\widetilde{M}(\boldsymbol{\theta}) \quad \text{and} \quad q^* M(\boldsymbol{\theta}) \le \widetilde{M}(\boldsymbol{\theta}) \le Q^* M(\boldsymbol{\theta}).$$

G.4. Proof of Theorem 4.3

Proof. Now we prove Ineq. (16). As the argument of Theorem 4.2, we also start with Ineq. (14) and attempt to seek for its upper bound.

By rearranging Eq. (9) and combining with Eq. (27), it results in

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

That is,

$$\partial_t\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x},t)\big) = \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x},t) + \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}},\boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \langle \boldsymbol{f},\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}\rangle\Big]$$

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

$$\begin{aligned} \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) &= \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, T) \\ &+ \int_T^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) d\tau + \int_T^t \nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big] d\tau. \end{aligned} \tag{39}$$

In the last term, we may compute $\nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big]$ as

$$\begin{aligned} &\nabla_{\boldsymbol{x}} \Big[\frac{1}{2} g^2(t) \left\| \boldsymbol{s}_{\boldsymbol{\theta}} \right\|_2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle - \frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle + \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big] \\ =&\nabla_{\boldsymbol{x}} \Big[\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}] + \text{div}_{\boldsymbol{x}}(\boldsymbol{f}) - \frac{1}{2} g^2(t) \text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) - \frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle + \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big] \\ =&\nabla_{\boldsymbol{x}} \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}] + \nabla_{\boldsymbol{x}} \Big[\text{div}_{\boldsymbol{x}}(\boldsymbol{f}) - \frac{1}{2} g^2(t) \text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) - \frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle + \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big], \end{aligned} \tag{40}$$

where $\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t) := \frac{1}{2} g^2(t) \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \right\|_2 - \langle \boldsymbol{f}(\boldsymbol{x}, t), \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \rangle + \frac{1}{2} g^2(t) \text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)) - \text{div}_{\boldsymbol{x}}(\boldsymbol{f}(\boldsymbol{x}, t))$. We apply the Taylor expansion at any fixed point $\boldsymbol{x}_0$ to $\nabla_{\boldsymbol{x}} \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}]$ and get

$$\mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}_0, t) - \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t) = \nabla_{\boldsymbol{x}} \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, t) \cdot (\boldsymbol{x}_0 - \boldsymbol{x}) + \mathcal{O}(\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2). \tag{41}$$

Now set $\boldsymbol{x}_0 := \boldsymbol{x} + \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t)$ and re-denote it as $\boldsymbol{x}_{\boldsymbol{\theta}}$. Combining Eq. (39), Eq. (40), and Eq. (41), and taking the dot product with $\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}$ from the both side of Eq. (39), we obtain

$$\begin{aligned} \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) \right\|_2^2 \leq& \left| \langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, T), \boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x} \rangle \right| + \left| \langle \int_T^t \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) d\tau, \boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x} \rangle \right| \\ &+ \int_T^t \left| \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}_0, \tau) - \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}_{\boldsymbol{\theta}}, \tau) \right| d\tau + \mathcal{O}(\|\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}\|_2^2) \\ &+ \left| \langle \int_T^t \nabla_{\boldsymbol{x}} \Big[\text{div}_{\boldsymbol{x}}(\boldsymbol{f}) - \frac{1}{2} g^2(t) \text{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) - \frac{1}{2} g^2(t) \langle \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}, \boldsymbol{s}_{\boldsymbol{\theta}} \rangle + \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}} \rangle \Big], \boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x} \rangle \right| \end{aligned} \tag{42}$$

With Assumption F.1 (b)-(f),

$$\begin{aligned} \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) \right\|_2^2 \leq& \left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, T) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, T) \right\| \|\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}\| + \int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\| \, d\tau \, \|\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}\| \\ &+ 2 \sup_{\boldsymbol{x}} \int_0^T \left| \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right| d\tau + \mathcal{O}(\|\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}\|_2^2) + \big(1 + \|\boldsymbol{x}\|\big) \|\boldsymbol{x}_{\boldsymbol{\theta}} - \boldsymbol{x}\| \\ \lesssim& \int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\| \, d\tau \cdot \big(1 + \|\boldsymbol{x}\|\big) + \sup_{\boldsymbol{x}} \int_0^T \left| \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right| d\tau \\ &+ \big(1 + \|\boldsymbol{x}\|\big) + \big(1 + \|\boldsymbol{x}\|\big)^2 \end{aligned} \tag{43}$$

Taking the expectation over $q_t(\boldsymbol{x})$ and applying Cauchy-Schwartz inequality, we obtain

$$\begin{aligned} \mathbb{E}_{q_t(\boldsymbol{x})} \Big[\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{s}_{\boldsymbol{\theta}}^{\text{ODE}}(\boldsymbol{x}, t) \right\|_2^2 \Big] \lesssim& \delta_T \mathbb{E}_{q_t(\boldsymbol{x})} \big[\big(1 + \|\boldsymbol{x}\|\big) \big] + \mathbb{E}_{q_t(\boldsymbol{x})} \Big[\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\| \, d\tau \Big] \cdot \mathbb{E}_{q_t(\boldsymbol{x})} \Big[\big(1 + \|\boldsymbol{x}\|\big) \Big] \\ &+ 2 \sup_{\boldsymbol{x}} \int_0^T \left| \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right| d\tau + \mathbb{E}_{q_t(\boldsymbol{x})} \big[\big(1 + \|\boldsymbol{x}\|_2\big) \big] + \mathbb{E}_{q_t(\boldsymbol{x})} \big[\big(1 + \|\boldsymbol{x}\|_2\big)^2 \big] \\ \lesssim& \mathbb{E}_{q_t(\boldsymbol{x})} \Big[\int_0^T \|\boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau)\| \, d\tau \Big] + \sup_{\boldsymbol{x}} \int_0^T \left| \mathcal{L}[\boldsymbol{s}_{\boldsymbol{\theta}}](\boldsymbol{x}, \tau) \right| d\tau + C_2(L, T, \delta_T, g). \end{aligned} \tag{44}$$

□

G.5. Proof and discussion of Proposition 4.4

Proof. 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

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

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

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

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

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

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

Hence, the proposition is proved. □

Proposition 4.4 requests a perfect match of scores at some single timestep $t_{\boldsymbol{\theta}} \in [0, T]$: $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t_{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{x}} \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x})$ for all $\boldsymbol{x}$. However, we can involve an error term when the scores are not matched exactly and formulate an extended version of Proposition 4.4 as the following.

Proposition G.4. *Suppose that there is a constant $\delta > 0$ so that for any $\boldsymbol{\theta}$, there is a single timestep $t_{\boldsymbol{\theta}} \in [0, T]$ such that $\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{x}} \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x}) \right\|_2 \leq \delta$, then there exists a real-valued function $\Psi_{\boldsymbol{\theta}} \colon \mathbb{R}^D \times [0, T] \to \mathbb{R}$ (with an explicit expression) that satisfies*

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

Proof. The proof is almost identical to the original one. We start with

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

By inserting the term $\nabla_{\boldsymbol{x}} \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x})$, we have

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

We now let $\Psi_{\boldsymbol{\theta}}(\boldsymbol{x},t) = \log q_{t_{\boldsymbol{\theta}}}(\boldsymbol{x}) + \int_{t_{\boldsymbol{\theta}}}^{t} \left[\frac{1}{2}g^2(\tau)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(\tau)\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\right] d\tau$. By taking the norm of the above equation, we establish the claim.

□

G.6. Proof of Proposition 4.5

Lemma G.5. *Let $\boldsymbol{s}_{\boldsymbol{\theta}}$ be a score obtained from denoising score matching (Eq. (3)) and write $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}(\cdot,t) := \nabla_{\boldsymbol{x}} \log p_{t,\boldsymbol{\theta}}^{SDE}$. Then*

1. *(Lu et al., 2022) Eq. (4) associates with the following forward SDE whose marginal density is $\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}$:*

$$d\boldsymbol{x}_{\boldsymbol{\theta}}(t) = \left[\boldsymbol{f}(\boldsymbol{x}_{\boldsymbol{\theta}}(t),t) + g^2(t)\big(\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}(\boldsymbol{x}_{\boldsymbol{\theta}}(t),t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_{\boldsymbol{\theta}}(t),t)\big)\right]dt + g(t)\boldsymbol{w}_t$$

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

$$\partial_t \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}} - \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}\big(2\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}\big) + \frac{1}{2}g^2(t)\big(2\langle \boldsymbol{s}_{\boldsymbol{\theta}}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}\rangle - \big\|\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}\big\|_2^2\big) - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] = 0. \tag{45}$$

Proof. Lemma G.5

The proof of the first statement can be found in (Lu et al., 2022). We now prove the second statement.

Consider

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

in Eq. (25), and apply Proposition 3.1, the lemma is then established. □

Proof. Proposition 4.5

We recall Eq. (9), which indicates

$$\partial_t \boldsymbol{s}_{\boldsymbol{\theta}} - \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}}) + \frac{1}{2}g^2(t)\left\|\boldsymbol{s}_{\boldsymbol{\theta}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}}\rangle - \mathrm{div}_{\boldsymbol{x}}(\boldsymbol{f})\Big] - \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}] = 0. \tag{46}$$

First, we subtract Eq. (45) by the above equation and get

$$\partial_t(\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}} - \boldsymbol{s}_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{x}}\Big[\frac{1}{2}g^2(t)\mathrm{div}_{\boldsymbol{x}}(\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}) - \frac{1}{2}g^2(t)\left\|\boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}\right\|_2^2 - \langle \boldsymbol{f}, \boldsymbol{s}_{\boldsymbol{\theta}} - \boldsymbol{s}_{\boldsymbol{\theta}}^{\mathrm{SDE}}\rangle\Big] + \boldsymbol{\epsilon}[\boldsymbol{s}_{\boldsymbol{\theta}}] = 0. \tag{47}$$

Consider when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and let $\boldsymbol{u}_{\boldsymbol{\theta}_0} := \boldsymbol{s}_{\boldsymbol{\theta}_0}^{\mathrm{SDE}} - \boldsymbol{s}_{\boldsymbol{\theta}_0}$. 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 $\boldsymbol{s}_{\boldsymbol{\theta}_0}^{\mathrm{SDE}}$ 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, $\boldsymbol{s}_{\boldsymbol{\theta}_0}^{\mathrm{SDE}} \equiv \boldsymbol{s}_{\boldsymbol{\theta}_0}$.

We repeat the same trick to subtract Eq. (8) by Eq. (46) from which we can obtain $\boldsymbol{s}_{\boldsymbol{\theta}_0} \equiv \boldsymbol{s}$. Similarly, the same argument can be applied to Eq. (28) to prove $\boldsymbol{s}_{\boldsymbol{\theta}_0}^{\mathrm{ODE}} \equiv \boldsymbol{s}_{\boldsymbol{\theta}_0}$.

□

G.7. Proof of Proposition 4.6

Proof. By subtracting the following two equations

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

we obtain

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

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

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

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.

□

G.8. Proof of Proposition C.2

Proof. For any δ which is small enough, by (Skorski, 2021), we have

$$\mathbb{P}\big(\|\mathscr{H}[\boldsymbol{s}]\|_D \geq \delta\big) \leq \exp^{-\left(\frac{\frac{4M}{g^*}\delta^2}{2(1-\frac{16}{3g^*})\delta}\right)}. \tag{48}$$

On the other hand, Lemma C.1 implies that there is a $C > 0$ so that $\|\mathscr{F}[\boldsymbol{s}]\|_D < CDh$, where $h = \frac{h_s^2 h_d + h_d h_s^2}{h_s + h_d}$ and $\|\cdot\|_D$ indicates the ℓ_D-norm. Hence, we have

$$\mathbb{P}\big(\|\mathscr{F}[\boldsymbol{s}]\|_D \geq CDh\big) \leq 0. \tag{49}$$

Now rearranging, we have

$$\mathscr{E}[\boldsymbol{s}] = \big(\partial_t \boldsymbol{s} - \mathrm{FD}(\boldsymbol{s})\big) - \big(\frac{1}{2}g^2\mathrm{tr}(\nabla \boldsymbol{s}) - \frac{1}{2}g^2\mathrm{tr}_{H^{(M)}}(\nabla \boldsymbol{s})\big) = \mathscr{T}[\boldsymbol{s}] - \mathscr{H}[\boldsymbol{s}].$$

With the statistical bounds (48) and (49), we obtain

$$\mathbb{P}\big(\|\mathscr{E}[\boldsymbol{s}]\|_D \geq \delta + CDh\big) \leq \mathbb{P}\big(\|\mathscr{F}[\boldsymbol{s}]\|_D \geq \delta\big) + \mathbb{P}\big(\|\mathscr{H}[\boldsymbol{s}]\|_D \geq CDh\big) \leq \exp^{-\left(\frac{\frac{4M}{g^*}\delta^2}{2(1-\frac{16}{3g^*})\delta}\right)}$$

If h and ϵ is selected small enough (as the assumption), by taking $\delta = \epsilon - CDh$, we observe that $\epsilon > \delta > \frac{\epsilon}{2}$ and that

$$\exp^{-\left(\frac{4M\epsilon^2}{2(g^* - \frac{16}{3}\epsilon)}\right)} \leq \exp^{-\left(\frac{\frac{4M}{g^*}\delta^2}{2(1-\frac{16}{3g^*})\delta}\right)} \leq \exp^{-\left(\frac{M\epsilon^2}{2(g^* - \frac{8}{3}\epsilon)}\right)}.$$

Thus, the claimed error bound is established.

□