# CONVERGENCE ANALYSIS OF PROBABILITY FLOW ODE FOR SCORE-BASED GENERATIVE MODELS

DANIEL ZHENGYU HUANG[1], JIAOYANG HUANG[2], AND ZHENGJIANG LIN[3]

Abstract. Score-based generative models have emerged as a powerful approach for sampling high-dimensional probability distributions. Despite their effectiveness, their theoretical underpinnings remain relatively underdeveloped. In this work, we study the convergence properties of deterministic samplers based on probability flow ODEs from both theoretical and numerical perspectives. Assuming access to $L^2$-accurate estimates of the score function, we prove the total variation between the target and the generated data distributions can be bounded above by $\mathcal{O}(d\sqrt{\delta})$ in the continuous time level, where $d$ denotes the data dimension and $\delta$ represents the $L^2$-score matching error. For practical implementations using a $p$-th order Runge-Kutta integrator with step size $h$, we establish error bounds of $\mathcal{O}(d(\sqrt{\delta} + (dh)^p))$ at the discrete level. Finally, we present numerical studies on problems up to 128 dimensions to verify our theory, which indicate a better score matching error and dimension dependence.

## 1. Introduction

In recent years, score-based generative models [13, 18, 35, 37, 38] have emerged as a powerful paradigm for sampling high-dimensional probability distributions. Unlike traditional generative models that directly parameterize the mapping from random noise to target distribution samples [17, 20, 30, 33], score-based generative models consist of two stochastic processes—the forward and reverse processes. The forward process transforms samples from the target data distribution $\mu_*$ with density $q_0$ into pure noise, a step commonly referred to as the diffusion process. The gradient of the log-density function, also known as the score function, is learned from these trajectories using score matching techniques [19, 37, 38, 42]. The reverse process, guided by the score function, transforms random noise back into samples from $q_0$. This methodology has been proven effective in synthesizing high-fidelity audio and image data [12–14, 31, 32].

The reverse process is commonly implemented either as stochastic dynamics or deterministic dynamics, the latter often formulated as probability flow ordinary differential equation (ODE). These probability flow ODEs can typically be discretized using numerical methods such as forward Euler, exponential integrator, Heun, and high-order Runge-Kutta methods. Recent advancements in methods like those proposed in [28, 29, 36, 38, 44, 45] have enabled denoising steps to be completed in just a few iterations (e.g., 50 steps), compared to the Euler-Maruyama scheme typically employed for stochastic dynamics, which often requires a significantly larger number of steps (e.g., 1000 steps). Consequently, these deterministic methods achieve better efficiency in generating samples with only moderate quality degradation. The deterministic dynamics depends on the score function, which is typically learned by a neural network through the score matching process involving non-convex optimization. Consequently, the score estimation is inherently imperfect. This imprecision, coupled with discretization error, poses a critical question: How does the interplay between score matching error and discretization error influence the convergence of the deterministic dynamics towards the true data distribution? Our work seeks to address this question by delving into the convergence analysis of probability flow ODEs within the context of score-based generative models.

For the stochastic dynamics, convergence analyses have been explored in various works such as [5, 7, 9, 11, 21–23, 39, 40, 43], with notable contributions from [7, 9, 23, 40], offering convergence guarantees with polynomial complexity, without relying on any structural assumptions on the data distribution like log-concavity. The stochastic nature of these dynamics plays a crucial role in mitigating error accumulation. However, the deterministic counterpart warrants further exploration. Related works include [10], which assumes no score matching error and provides a discretization analysis for the probability flow ODE in KL divergence. However, their bounds exhibit a large dependence on dimensionality and are exponential in the Lipschitz constant of the score integrated over time. In contrast, [8] assumes $L^2$ bounds on the score estimation offers polynomial-time convergence guarantees for the probability flow ODE combined with a stochastic Langevin corrector, without relying on any structural assumptions on the data distribution. Similarly, [24,25] provide polynomial-time convergence guarantees for the probability flow ODE by requiring control of the difference between the derivatives of the true and approximate scores. Additionally, [2,4] analyze the convergence of the deterministic dynamics at continuous time level, exhibiting exponential dependence on the Lipschitz constant, stemming from a more general stochastic interpolant or flow matching setup [1,6,26,27]. Finally, [16] offers convergence analysis for the general probability flow ODEs with log-concavity data assumption, where the error bounds grow exponentially with time $T$ in the presence of the score matching error.

1.1. **Our Contributions.** We analyze the convergence of the probability flow ODE from both theoretical and numerical perspectives. Our detailed contributions are as follows:

- We provide convergence guarantees of the probability flow ODE at the continuous time level under three mild assumptions. These assumptions are as follows: Assumption 3.1 asserts that the target density has a compact support, Assumption 3.2 asserts the $L^2$ score matching error over time is bounded by $\delta$, and Assumption 3.3 asserts the first and second derivatives of the estimated score are bounded. Under these assumptions, we prove in Theorem 3.4 that the total variation distance between the target and the generated data distributions can be bounded above by $\mathcal{O}(d\sqrt{\delta})$, where $d$ is the data dimension.
- We provide convergence guarantees of the probability flow ODE at the disretized level. To accommodate a $p$-th order time integrator, we further require Assumption 3.7, that the estimated score function's first $(p+1)$-th derivatives are bounded. We establish in Theorem 3.9 that the total variation distance between the target and the generated data distributions can be bounded above by $\mathcal{O}(d(\sqrt{\delta}+(dh)^p))$ .
- We verify our theoretical discoveries through numerical studies on problems with Gaussian mixture target densities up to 128 dimensions. By intentionally introducing artificial score matching errors and employing the widely used second-order Heun's time integrator, our numerical results demonstrate a total variation error of $\mathcal{O}(\delta+h^2)$ (for the marginal distributions), with improved dependencies on dimension and score matching error.

In our theoretical proof at the continuous time level, we combine the method of characteristic lines and calculus of variations to estimate the total variation between the generated data distribution $\widehat{q}_t$ and the target distribution $q_t$ along the diffusion process. Compared to using Grönwall's inequality directly, our error estimate in Theorem A.1 does not include an exponential term in time. We provide two mathematically rigorous yet simple proofs of Theorem A.1, and also illustrate our intuition in Section A. Furthermore, our methods imply a more general Theorem A.3 for the $L^1$-norms of solutions of general transport equations. In Remark 3.10 and Remark A.4, we highlight that our methods can also estimate the $L^1(\mathbb{R}^d)$-norms of derivatives of $\widehat{q}_t - q_t$. Consequently, we can conclude that the $L^r(\mathbb{R}^d)$-norm of $\widehat{q}_t - q_t$ is also small when $r > 1$. See Remark 3.10 and Remark A.4 for further details. Additionally, our method extends to estimating the pointwise difference between $\widehat{q}_t$ and $q_t$, although we defer this investigation to future work to maintain the manuscript's conciseness. In our proof of Theorem 3.4, we leverage the Gagliardo-Nirenberg interpolation inequality with a universal constant, meaning the constant is independent of the dimension. To provide a comprehensive literature review, we include the proof of this dimension-free interpolation inequality as Lemma C.1.

For our convergence analysis of the probability ODE flow at the discretized level, a first step is to reformulate the discrete solution obtained by the $p$-th order Runge-Kutta method as a continuous-time ODE flow using interpolation. We derive an interpolation in Proposition D.1, and crucially, the score function associated with the interpolated ODE flow and the original approximated score function (and their derivatives) are close up to a $p$-th order error, i.e., $\mathcal{O}(h^p)$. Employing the characteristic method described in Appendix A again, the error at the discrete level decomposes into two parts: the score matching error between the generated data distribution $\widehat{q}_t$ and the target distribution $q_t$ along the diffusion process, and the discretization error between the interpolated ODE flow solution and the generated data distribution $\widehat{q}_t$. Consequently, the score matching error and time discretization error do not interact to magnify, thus preserving the time discretization error at the $p$-th order.

Our assumptions on the true data distribution $\mu_*$ are quite general. In Assumption 3.1, we assume that $\mu_*$ has a compact support. In Appendix B, we extend our main result to the case where $\mu_*$ is a Gaussian mixture. We emphasize that under Assumption 3.1, $\mu_*$ may not have a density, and the compact support $K_*$ of $\mu_*$ can be a submanifold of a much lower dimension in $\mathbb{R}^d$, particularly point masses. Refer to our Example B.1 and Example B.2, where we observe that assuming $\nabla \log q_t$ is uniformly Lipschitz with a constant independent of $t$ is unreasonable, as it actually tends to $\infty$ as $t \to 0^+$. In Lemma B.3 and Lemma B.5, we compute and estimate the high-order derivatives of $q_t$ (and $\log q_t$), when $\mu_*$ has a compact support and when $\mu_*$ is a Gaussian mixture. For Gaussian mixtures, the error estimates are better than the case when $K_*$ is a submanifold of a much lower dimension. We also mention in Remark C.4 that our methods also apply to other reasonable assumptions on $\mu_*$ once some simple estimates are satisfied.

### 1.2. **Natations.**

- Diacritics: $\widehat{\square}$ denotes quantities involve score error, $\widetilde{\square}$ denotes quantities involve time discretization error.
- Time steps: $0 = t_0 < t_1 < \cdots < t_N = T - \tau$, where $\tau > 0$ is a small parameter.
- Distributions on $\mathbb{R}^d$: $q_t, \widehat{q}_t, \widetilde{q}_t$ denote forward process, $\varrho_t = q_{T-t}, \widehat{\varrho}_t = \widehat{q}_{T-t}, \widetilde{\varrho}_t = \widetilde{q}_{T-t}$ reverse process. We also define $\widehat{\varepsilon}_t(x) := \widehat{q}_t(x) - q_t(x), \widetilde{\varepsilon}_t(x) := \widetilde{q}_t(x) - q_t(x)$.
- Vector fields from $\mathbb{R}^d$ to $\mathbb{R}^d$: Forward process: $U_t(x) := x + \nabla \log q_t(x)$, $\widehat{U}_t := x + s_{T-t}(x)$, $\widetilde{U}_t := x + \widetilde{s}_{T-t}(x)$, $\delta_t(x) := \widehat{U}_t - U_t = s_{T-t}(x) - \nabla \log q_t(x)$, $\widetilde{\delta}_t(x) := \widetilde{U}_t(x) - U_t(x)$; Reversal process: $V_t := U_{T-t}$, $\widehat{V}_t := \widehat{U}_{T-t}$, $\widetilde{V}_t := \widetilde{U}_{T-t}$; Other vector fields: $Z_t(x)$.
- $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index with nonnegative integers $\alpha_i$'s, $|\alpha| := \alpha_1 + \alpha_2 + \ldots + \alpha_d$, and we define $\partial_x^\alpha := \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \cdots \partial_{x_d}^{\alpha_d}$. We also use $\partial_i := \partial_{x_i}$ for simplicity.
- Constants: We use $C_u$ to denote universal constants like $10, 50, 100, 200$, i.e., $C_u$ is independent of the dimension $d$ and other parameters in this paper. Also, $C_u$ may vary by lines.
- Norms: For a vector $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, we use $\|x\| = \|x\|_2 := (x_1^2 + x_2^2 + \ldots + x_d^2)^{\frac{1}{2}}$, $\|x\|_\infty := \sup_{1 \le i \le d} |x_i|$, $\|x\|_1 := |x_1| + |x_2| + \ldots + |x_d|$, $\|x\|_p := (|x_1|^p + |x_2|^p + \ldots + |x_d|^p)^{\frac{1}{p}}$. We similarly define $\|\cdot\|_2$, $\|\cdot\|_\infty$, $\|\cdot\|_1$, $\|\cdot\|_p$, for matrices or even more general tensors, because we can view them as vectors and forget their tensor structures. For a vector-valued function $F(x) = (f_1(x), f_2(x), \ldots, f_m(x)) : \mathbb{R}^d \mapsto \mathbb{R}^m$, where $m$ is a positive integer, we usually regard $F(x)$ as a vector in $\mathbb{R}^m$ and similarly use the notations $\|F(x)\|$, $\|F(x)\|_1$, $\|F(x)\|_\infty$.
- Function class: We say a vector-valued function $F(x) = (f_1(x), f_2(x), \ldots, f_m(x)) : \mathbb{R}^d \mapsto \mathbb{R}^m$ as being in $C^r$, if each of its components $f_i(x)$ has continuous first $r$-th derivatives. We say $F(x)$ is in the $L^s$-space $L^s(\mathbb{R}^d)$ if for each of its components $f_i(x)$, its $L^s(\mathbb{R}^d)$-norm defined as $\|f_i\|_{L^s(\mathbb{R}^d)} := (\int_{\mathbb{R}^d} |f_i(x)|^s \mathrm{d}x)^{\frac{1}{s}}$ is finite. We say $F(x)$ is in the Sobolev space $W^{r,s}(\mathbb{R}^d)$

if for each of its components $f_i(x)$, $\partial_x^\alpha f_i \in L^s(\mathbb{R}^d)$ for each $\alpha$ with $|\alpha| \leq r$. We define the $W^{r,s}(\mathbb{R}^d)$-norm of $f_i$ as $\|f_i\|_{W^{r,s}(\mathbb{R}^d)} \coloneqq (\sum_{|\alpha|\leq r} \|\partial_x^\alpha f_i\|_{L^s(\mathbb{R}^d)}^s)^{\frac{1}{s}}$.

## 2. Preliminaries

2.1. **Score-based Generative Model.** Score-based generative models begin with $d$ dimensional true data samples $\{X_0\}$ following an unknown target distribution $\mu_*$ with density $q_0$. The objective is to sample new data from the target distribution. Typically, the score-based generative models usually involve two processes—the forward and reverse processes.

In the forward process, we start with data samples from $q_0$, and progressively transform the data into noise. This process is often based on the canonical Ornstein-Uhlenbeck (OU) process given by

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \qquad X_0 \sim q_0, \qquad 0 \leq t \leq T, \tag{1}$$

where $(B_t)_{t\in[0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$. The OU process has an analytical solution

$$X_t \stackrel{d}{=} \lambda_t X_0 + \sigma_t W, \qquad W \sim \mathcal{N}(0, \mathbb{I}_d), \tag{2}$$

with $\lambda_t = e^{-t}$ and $\sigma_t = \sqrt{1-e^{-2t}}$. The OU process exponentially converges to its stationary distribution, the standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$. Let $q_t$ denote the density of $X_t$, which evolves according to the following Fokker–Planck equation:

$$\partial_t q_t = \nabla \cdot \big((x + \nabla \log q_t(x))q_t\big) = \nabla \cdot (U_t q_t), \qquad 0 \leq t \leq T,$$

with $U_t(x) \coloneqq x + \nabla \log q_t(x)$.

By denoting $\varrho_t = q_{T-t}$, the time reversal process from time $T$ to 0, satisfies the following partial differential equation (PDE):

$$\partial_t \varrho_t = -\nabla \cdot \big((x + \nabla \log q_{T-t})\varrho_t\big) = -\nabla \cdot (V_t \varrho_t), \qquad 0 \leq t \leq T, \tag{3}$$

with $V_t(x) \coloneqq x + \nabla \log q_{T-t}(x)$. The score function $\nabla \log q_{T-t}(x)$ is typically learned by a neural network trained using score matching techniques with progressively corrupted trajectories $\{X_t\}$ from (1). Subsequently, the reverse PDE can be solved from $\varrho_0 = q_T$ to sample new data from $q_0$.

The reverse PDE (3) is often reformulated into a mean field equation for sampling instead of being directly solved. This mean field equation can manifest as stochastic dynamics

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log q_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t, \qquad Y_0 \sim q_T, \qquad 0 \leq t \leq T, \tag{4}$$

where $(B'_t)_{0\leq t\leq T}$ is a Brownian motion in $\mathbb{R}^d$. This formulation is commonly referred to as the denoising diffusion probabilistic model (DDPM). Alternatively, the mean-field equation can adopt a deterministic dynamics framework in terms of an ordinary differential equation with velocity field $V_t$:

$$\partial_t Y_t = Y_t + \nabla \log q_{T-t}(Y_t) = V_t(Y_t), \qquad Y_0 \sim q_T, \qquad 0 \leq t \leq T, \tag{5}$$

known as the probability flow ODE. Additionally, when $\nabla \log q_{T-t}(x)$, is represented by the learned score function $s_t(x)$, the probability flow ODE (5) becomes

$$\partial_t \widehat{Y}_t = \widehat{Y}_t + s_t(\widehat{Y}_t) = \widehat{V}_t(\widehat{Y}_t), \qquad \widehat{Y}_0 \sim \widehat{\varrho}_0, \tag{6}$$

here the velocity field becomes $\widehat{V}_t(x) \coloneqq x + s_t(x)$. And $\widehat{Y}_0$ is sampled from $\widehat{\varrho}_0$, since the density $q_T$ is unknown. $\widehat{\varrho}_0$ is commonly approximated by the standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$, which serves as a reliable approximation of $q_T$ for sufficiently large $T$. The associated density of $\widehat{Y}_t$ is denoted as $\widehat{\varrho}_t$, which differs from $\varrho_t$ that describes the density of $Y_t$, due to the score matching error.

2.2. **Time Integrator.** To numerically solve the probability flow ODE (6), a time integrator is essential. Fix a small $\tau > 0$, we discretize the time interval $[0, T-\tau]$ into $N$ time steps $0 = t_0 < t_1 < \cdots < t_N = T-\tau$, typically using a uniform step size $h = (T-\tau)/N$, Starting from an initial condition $\widetilde{Y}_{t_0}$ sampled from $q_T$, the time integrator iteratively estimates $\widetilde{Y}_{t_i}$ at time $t_i$. One commonly used time integrator is the Runge-Kutta method, the family of explicit $s$-stage $p$-th order Runge-Kutta methods updates $\{\widetilde{Y}_{t_i}\}$ as follows:

$$\widetilde{Y}_{t_{i+1}} = \widetilde{Y}_{t_i} + h\sum_{j=1}^{s} b_j k_j, \tag{7}$$

where

$$\begin{aligned}
k_1 &= \widehat{V}_{t_i+c_1h}\big(\widetilde{Y}_{t_i}\big),\\
k_2 &= \widehat{V}_{t_i+c_2h}\big(\widetilde{Y}_{t_i} + (a_{21}k_1)h\big),\\
k_3 &= \widehat{V}_{t_i+c_3h}\big(\widetilde{Y}_{t_i} + (a_{31}k_1 + a_{32}k_2)h\big),\\
&\ \ \vdots\\
k_s &= \widehat{V}_{t_i+c_sh}\big(\widetilde{Y}_{t_i} + (a_{s1}k_1 + a_{s2}k_2 + \cdots + a_{s,s-1}k_{s-1})h\big).
\end{aligned} \tag{8}$$

The lower triangular matrix $[a_{jk}]$ is called the Runge–Kutta matrix, while the $b_j$ and $c_j$ are known as the weights and the nodes. The stage number $s$ and the parameters are chosen such that the local truncation error of (7) is $\mathcal{O}(h^{p+1})$. In general, $s \geq p$ and if $p \geq 5$, then $s \geq p+1$.

For example, forward Euler scheme is the 1-stage first order Runge-Kutta method:

$$\widetilde{Y}_{t_{i+1}} = \widetilde{Y}_{t_i} + hk_1 \qquad k_1 = \widehat{V}_{t_i}\big(\widetilde{Y}_{t_i}\big)$$

Heun's method is the 2-stage second order Runge-Kutta method:

$$\widetilde{Y}_{t_{i+1}} = \widetilde{Y}_{t_i} + \frac{h}{2}(k_1 + k_2) \qquad k_1 = \widehat{V}_{t_i}\big(\widetilde{Y}_{t_i}\big) \qquad k_2 = \widehat{V}_{t_{i+1}}\big(\widetilde{Y}_{t_i} + hk_1\big)$$

**Remark 2.1.** *The time discretization error between Runge-Kutta solution $\widetilde{Y}_{t_i}$ and the true solution $\widehat{Y}_{t_i}$ is typically analyzed through the concept of local truncation error, which is interpreted as follows. Consider any time interval $[t_i, t_{i+1}]$, solve (6) in the time interval with $\widetilde{Y}_{t_i} = \widehat{Y}_{t_i}$ analytically, gives*

$$\widehat{Y}_{t_{i+1}} = \widetilde{Y}_{t_i} + h\widehat{V}_{t_i}(\widetilde{Y}_{t_i}) + \frac{h^2}{2}\frac{\mathrm{d}\widehat{V}_{t_i}(\widetilde{Y}_{t_i})}{\mathrm{d}t} + \cdots + \frac{h^p}{p!}\frac{\mathrm{d}^{p-1}\widehat{V}_{t_i}(\widetilde{Y}_{t_i})}{\mathrm{d}t^{p-1}} + \frac{1}{(p+1)!}\int_{t_i}^{t_{i+1}}(t_{i+1}-t)^p\frac{\mathrm{d}^p\widehat{V}_t(\widetilde{Y}_t)}{\mathrm{d}t^p}\mathrm{d}t. \tag{9}$$

*Similarly for the $s$-stage $p$-th order Runge-Kutta methods (7), we can view $\widetilde{Y}_{t_i+r} = \widetilde{F}_r(\widetilde{Y}_{t_i})$ as a function of $r \in [0, h]$ (by replacing $h$ to $r$ in (7) amd (8)) and $\widetilde{Y}_{t_i}$, and perform a Taylor expansion around $r = 0$*

$$\widetilde{Y}_{t_{i+1}} = \widetilde{Y}_{t_i} + h\frac{\mathrm{d}F_0(\widetilde{Y}_{t_i})}{\mathrm{d}r} + \cdots + \frac{h^p}{p!}\frac{\mathrm{d}^pF_0(\widetilde{Y}_{t_i})}{\mathrm{d}r^p} + \frac{1}{(p+1)!}\int_0^{t_{i+1}-t_i}(h-r)^p\frac{\mathrm{d}^{p+1}\widetilde{F}_r(\widetilde{Y}_{t_i})}{\mathrm{d}r^{p+1}}\mathrm{d}r. \tag{10}$$

*The Runge–Kutta matrix $[a_{jk}]$, weights $b_j$ and nodes $c_j$ are carefully chosen such that the coefficients in front of $h, h^2, \cdots, h^p$ (as a function of $\widetilde{Y}_{t_i}$) in (9) and (10) cancel perfectly. Thus the Runge-Kutta estimation $\widetilde{Y}_{t_{i+1}}$ satisfies*

$$\widetilde{Y}_{t_{i+1}} = \widehat{Y}_{t_{i+1}} + R_{p+1}(\widehat{V}_t, \widetilde{Y}_i, t_i, h), \quad R_{p+1}(\widehat{V}_t, \widetilde{Y}_{t_i}, t_i, h) = \mathcal{O}(h^{p+1}).$$

## 3. Main Results

The probability flow ODE (5) describes the evolution of $Y_t$; while its counterpart with the estimated score function (6) describes the evolution of $\widehat{Y}_t$. We denote the density of $Y_t$ and $\widehat{Y}_t$ as $\varrho_t$ and $\widehat{\varrho}_t$ respectively. Then they satisfy the following PDEs

$$\begin{aligned} \partial_t \varrho_t &= -\nabla \cdot (V_t \varrho_t), \quad V_t = x + \nabla \log \varrho_t, \\ \partial_t \widehat{\varrho}_t &= -\nabla \cdot (\widehat{V}_t \widehat{\varrho}_t), \quad \widehat{V}_t = x + s_t(x). \end{aligned} \tag{11}$$

One major focus of our work is to understand propagation of the score matching error by analyzing the difference between $\widehat{\varrho}_t$ and $\varrho_t$.

We make the following assumption on the data distribution $\mu_*$.

**Assumption 3.1.** *The data distribution $\mu_*$ is positive and compactly supported on a compact set $K_*$, and we also define $D := 1 + \max_{x \in K_*} \|x\|_\infty$.*

We assume that the errors incurred during score matching are bounded in an $L^2$-sense.

**Assumption 3.2.** *Fix small $\tau > 0$. There exists a small $\delta > 0$, such that the score matching error is bounded by $\delta$ in the sense that*

$$\int_0^{T-\tau} \mathbb{E}_{\varrho_t} \left[ \|\nabla \log \varrho_t(x) - s_t(x)\|_2^2 \right] \mathrm{d}t \le \delta^2.$$

We assume that score estimates $s_t(x)$ are in $C^2$, and the first two derivatives are bounded by $L_t$, which may depend on time.

**Assumption 3.3.** *Fix small $\tau > 0$. We assume that the score estimate $s_t(x)$ are $C^2$ for any $0 \le t \le T - \tau$, and there exists a function $L_t > 0$ in $t$, such that*

$$\max_{1 \le j \le d} |s_t^{(j)}(0)| \le L_t, \quad \sup_{x \in \mathbb{R}^d} \max_{1 \le |\alpha| \le 2} \max_{1 \le j \le d} |\partial_x^\alpha s_t^{(j)}(x)| \le L_t.$$

*Here, we write $s_t(x) = (s_t^{(1)}(x), s_t^{(2)}(x), \ldots, s_t^{(d)}(x))$, and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index with nonnegative integers $\alpha_i$'s, $|\alpha| := \alpha_1 + \alpha_2 + \ldots + \alpha_d$, and we define $\partial_x^\alpha := \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \cdots \partial_{x_d}^{\alpha_d}$. We also define that $\mathcal{L} := \int_0^{T-\tau} L_t \, \mathrm{d}t$.*

**Theorem 3.4.** *Adopt Assumption 3.1,Assumption 3.2 and Assumption 3.3, there is a universal constant $C_u > 0$, such that the total variational distance between $\varrho_{T-\tau}$ and $\widehat{\varrho}_{T-\tau}$ is small in the sense that*

$$\mathsf{TV}(\varrho_{T-\tau}, \widehat{\varrho}_{T-\tau}) \le C_u \cdot d \cdot T^{\frac{1}{4}} \cdot (\mathcal{L} + T \cdot \tau^{-2} \cdot D^3)^{\frac{1}{2}} \cdot \delta^{\frac{1}{2}} + \mathsf{TV}(\varrho_0, \widehat{\varrho}_0). \tag{12}$$

*If we take the initialization to be the standard normal distribution $\widehat{\varrho}_0 = \mathcal{N}(0, \mathbb{I}_d)$, then $\mathsf{TV}(\varrho_0, \widehat{\varrho}_0) \le C_u e^{-T} \sqrt{d} D$, which is exponentially small in $T$.*

**Remark 3.5.** *We remark that the right hand side of (12) is only an upper bound. There are several ways to modify it:*

*(1) The term $\tau^{-1}$ appears because when $K_*$ is a submanifold of a dimension much lower than d, for example, several points, then $\nabla \log \varrho_{T-\tau} \approx \frac{1}{\sigma_\tau^2} \approx \tau^{-1}$ near $K_*$ and it becomes much more singular as $\tau \to 0^+$. See Example B.1 and Example B.2. One way to modify this is to modify the Assumption 3.2 by a time-weighted score matching error, which will be disscussed in Remark C.2; another way is to assume that our data distribution $\mu_*$ has a sufficiently regular density, e.g. a Gaussian mixture, then there will be uniform upper bounds (depending on parameters of $\mu_*$) on $\nabla \log \varrho_t$ together with its higher order derivatives, which are independent of the time $t$. So, one can let $\tau \to 0^+$ and the term $\tau^{-1}$ will not appear in the error estimate. Our proof of Theorem 3.4 works, essentially verbatim, after using*

*those bounds for $\nabla \log \varrho_t$ together with its higher order derivatives. See Lemma B.5 and Remark C.4 for more details if $\mu_*$ does not have a compact support and is possibly a Gaussian mixture.*

(2) *In our numerical simulation, the total variation distance is linear in $\delta$ (See Section 4). In Theorem A.1, the upper bound in (17) is of order $\delta^{1/2}$ because we use (36) in Lemma C.1 to prove Theorem 3.4 under the Assumption 3.3 first. If we can add Assumption 3.7, that is, the score estimate $s_t(x)$ has higher derivatives and we can control them, then we can use (37) in Lemma C.1 to replace the exponent $\frac{1}{2}$ of $\delta$ in (12) with $1-\frac{1}{k}$ for $k \geq 2$, as discussed in Remark (C.3). In this case, the $(\mathcal{L}+T\cdot\tau^{-2}\cdot D^3)^{\frac{1}{2}}$ on the right hand side of (12) will then be replaced by $C_k(\mathcal{L}+T\cdot\tau^{-k}\cdot D^{k+1})^{\frac{1}{k}}$ for some positive constant $C_k$ depending on the $k$.*

**Remark 3.6.** *Although in Theorem 3.4 we only estimate the error up to the time $T-\tau$ instead of the true data $\mu_* = q_0$, the Wasserstein 2-distance between $q_0 = \varrho_T$ and $q_\tau = \varrho_{T-\tau}$ is actually small if $\tau > 0$ is small enough. This is because if we let $\gamma$ be the distribution of $(X_0, X_\tau)$ on $\mathbb{R}^{2d}$ for $X_t$ defined in (2), then*

$$W_2(\varrho_T, \varrho_{T-\tau})^2 \leq \mathbb{E}_\gamma \|X_0 - X_\tau\|^2 = \mathbb{E}_\gamma \|(1-\lambda_\tau)X_0 - \sigma_\tau W\|^2 \leq 2(1-\lambda_\tau)^2 \mathbb{E}_{\mu_*}\|X_0\|^2 + 2\sigma_\tau^2,$$

*where $\lambda_\tau = e^{-\tau}$ and $\sigma_\tau = \sqrt{1-\lambda_\tau^2}$. We notice that $1-\lambda_\tau \leq \tau$, and by Assumption 3.1, $\mathbb{E}_{\mu_*}\|X_0\|^2 \leq dD^2$, so*

$$W_2(\varrho_T, \varrho_{T-\tau})^2 \leq 2\tau^2 dD^2 + 4\tau,$$

*which goes to zero of order $\tau$ as $\tau \to 0^+$.*

Furthermore, to numerically solve the probability flow ODE (6), we discretize the time interval $[0, T-\tau]$ into $N$ time steps $0 = t_0 < t_1 < \cdots < t_N = T-\tau$, and employ time integration until $t_N = T-\tau$ to circumvent the potential singularity at $T$. Let $\widetilde{\varrho}_{t_i}$ denote the distribution of $\widetilde{Y}_{t_i}$ obtained by using the Runge-Kutta method described in Section 2.2. Another focus of our work is to further understand the impact of the time discretization error by analyzing the difference between $\widetilde{\varrho}_t$ and $\varrho_t$.

To use the $p$-th order Runge-Kutta method, we assume that $s_t(x)$ is in $C^{p+1}$ in the following assumption.

**Assumption 3.7.** *Fix $p \geq 1$ and a small $\tau > 0$. Assume there exits a large number $L = L(p,\tau) \geq 1$ the following holds. The approximate score function $s_t(x)$ satisfies $\|s_t(x)\|_2 \leq L(\sqrt{d} + \|x\|_2)$ and is $C^{p+1}$, such that the following holds*

$$\sup_{x\in\mathbb{R}^d} \max_{1\leq k+|\alpha|\leq p+1} \max_{1\leq j\leq d} \|\partial_t^k \partial_x^\alpha s_t^{(j)}(x)\| \leq L.$$

*for any $0 \leq t \leq T-\tau$.*

**Remark 3.8.** *In Assumption 3.7, besides the upper bounds of the derivatives of $s_t(x)$, we also assumed that $\|s_t(x)\|_2 \leq L(\sqrt{d} + \|x\|_2)$. This is an easy consequence, if $s_t(x)$ is $L$-Lipschitz.*

**Theorem 3.9.** *Adopt Assumption 3.1,Assumption 3.2, and Assumption 3.7, there is a universal constant $C_u > 0$ and constant $C(p,s) > 0$ (depending on the stage and the order of the Runge-Kutta method), such that the total variation between $\varrho_{T-\tau}$ and $\widetilde{\varrho}_{T-\tau}$ is small in the sense that*

$$\begin{aligned}\mathsf{TV}(\varrho_{T-\tau}, \widetilde{\varrho}_{T-\tau}) \leq C_u \Big( &\underbrace{dT^{\frac{3}{4}}(L+\tau^{-2}\cdot D^3)^{\frac{1}{2}}\delta^{\frac{1}{2}}}_{\textit{score matching error}} + \underbrace{C(p,s)d(hd)^p(LD)^{p+1}\log(T/\tau)}_{\textit{discretization error}}\Big) \\ &+ \mathsf{TV}(\varrho_0, \widetilde{\varrho}_0).\end{aligned}$$

*If we take the initialization to be the standard normal distribution $\widetilde{\varrho}_0 = \mathcal{N}(0, \mathbb{I}_d)$, then $\mathsf{TV}(\varrho_0, \widetilde{\varrho}_0) \le C'_u e^{-T}\sqrt{d}D$ for another positive universal constant $C'_u$. So, $\mathsf{TV}(\varrho_0, \widetilde{\varrho}_0)$ is exponentially small in $T$.*

**Remark 3.10.** *Under the assumptions of Theorem 3.9, one can even estimate the $L^1$-norms of higher order derivatives of $\varrho_{T-\tau} - \widehat{\varrho}_{T-\tau}$ by the Gagliardo-Nirenberg inequality in Lemma C.1, as illustrated in Remark A.4. By the same arguments in Section D, we can also get error estimates for the $L^1$-norms of higher order derivatives of $\varrho_{T-\tau} - \widetilde{\varrho}_{T-\tau}$. This means we can obtain the $W^{p,1}$-norms for $\varrho_{T-\tau} - \widetilde{\varrho}_{T-\tau}$, where $W^{p,1}$ means Sobolev spaces. By Sobolev inequalities, one can also obtain the corresponding $W^{p-1,r}$-norms for $r > 1$, in particular, the $L^r$-norm of $\varrho_{T-\tau} - \widetilde{\varrho}_{T-\tau}$ with $r > 1$.*

3.1. **Proof Outline.** To prove Theorem 3.4, we first consider these two first-order PDEs describing the forward processes

$$\partial_t q_t = \nabla \cdot (U_t q_t) \qquad \text{and} \qquad \partial_t \widehat{q}_t = \nabla \cdot (\widehat{U}_t \widehat{q}_t). \tag{13}$$

Here $\widehat{q}_t = \widehat{\varrho}_{T-t}$ and $\widehat{U}_t = x + s_{T-t}(x)$, and the second equation describes the density evolution of $\widehat{Y}_{T-t} \sim \widehat{q}_t$ denoted in (6). We denote $\delta_t(x) := \widehat{U}_t - U_t = s_{T-t}(x) - \nabla \log q_t(x)$ as the score matching error, and $\widehat{\varepsilon}_t(x) := \widehat{q}_t(x) - q_t(x)$ as the error in generated data distribution. Our goal is to use $\int_\tau^T \int_{\mathbb{R}^d} q_t(x)\|\delta_t(x)\|^2 \, \mathrm{d}x\mathrm{d}t$ to bound the $L^1$ error between $\widehat{q}_t$ and $q_t$ at time $\tau$, i.e., $\int_{\mathbb{R}^d} |\widehat{\varepsilon}_t(x)| \, \mathrm{d}x$ for $t = \tau$.

By employing the characteristic method for first-order PDEs (13), we can derive a bound for the time derivative of the $L^1$ error as follows:

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} |\widehat{\varepsilon}_t(x)| \, \mathrm{d}x \right| \le \int_{\mathbb{R}^d} \left| \big(\nabla \cdot (q_t \delta_t)\big)(x) \right| \mathrm{d}x.$$

The proof can be found in Appendix A. Integrating from $\tau > 0$ to $T$, the $L^1$ error is controlled by the gradient of the score error:

$$\int_{\mathbb{R}^d} |\widehat{\varepsilon}_\tau(x)| \, \mathrm{d}x \le \int_\tau^T \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t \delta_t))(x) \right| \mathrm{d}x + \int_{\mathbb{R}^d} |\widehat{\varepsilon}_T(x)| \, \mathrm{d}x.$$

Then we use Gagliardo-Nirenberg Lemma C.1 and estimations on derivatives of density $q_t$ as presented in Appendix B to control each component of the gradient $\nabla \cdot (q_t \delta_t)$ in the right-hand side in terms of the $L^2$ score error, leading to

$$\int_\tau^T \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t \delta_t))(x) \right| \mathrm{d}x\mathrm{d}t \le C_u \cdot d \cdot T^{\frac{1}{4}} \cdot (\mathcal{L} + T \cdot \tau^{-2} \cdot D^3)^{\frac{1}{2}} \left( \int_\tau^T \int_{\mathbb{R}^d} q_t(x)(\delta_t(x))^2 \, \mathrm{d}x \, \mathrm{d}t \right)^{\frac{1}{4}}.$$

We observe a 1/2-th order dependence on the $L^2$ score error and linear dependence on the dimensionality $d$. The presence of $\tau^{-2}$ is attributed to the possibility that true data lies on a submanifold of lower dimensionality than $d$. For a detailed proof and improved bounds concerning the Gaussian mixture true data distribution, please refer to Appendix C.

To prove Theorem 3.9, we first interpolate the discrete solution $\{\widetilde{Y}_{t_i}\}_{i=0}^N$ obtained by the $p$-th order Runge-Kutta method using interpolation. Then, we obtain a continuous time process on each time interval $[t_i, t_{i+1}]$, which can then be treated as an ODE flow

$$\partial_t \widetilde{Y}_t = \widetilde{V}_t(\widetilde{Y}_t),$$

where $\widetilde{V}_t$ is continuous on the $t$-direction when $t \in [t_i, t_{i+1}]$, but it may not be continuous crossing each $t_i$. The discrepancy between $\widetilde{V}_t$ and $\widehat{V}_t$ is studied in Proposition D.1. Specifically, we analyze

$$\|\widetilde{V}_t(x) - \widehat{V}_t(x)\|_\infty, \quad \|\nabla(\widetilde{V}_t(x) - \widehat{V}_t(x))\|_\infty \le C(p, s) \cdot L\big((\sqrt{d} + \|x\|_2)h\sqrt{d}L\big)^p, \tag{14}$$

which essentially represents the Runge-Kutta local truncation error with detailed dimension and Lipschitz constants. Let $\widetilde{q}_t$ denote the density of $\widetilde{Y}_{T-t}$, which satisfies the forward process

$$\partial_t \widetilde{q}_t = \nabla \cdot (\widetilde{U}_t \widetilde{q}_t),$$

with $\widetilde{U}_t = \widetilde{V}_{T-t}$. We will quantify the total variation between $\widetilde{q}_t$ and $q_t$, the density of $Y_{T-t}$. We define $\widetilde{\delta}_t(x) := \widetilde{U}_t(x) - U_t(x)$, $\widetilde{\varepsilon}_t(x) := \widetilde{q}_t(x) - q_t(x)$. By using the characteristic method described in Appendix A again, the error at the discrete level boils down into the score matching error and time discretization error:

$$\int_{\mathbb{R}^d} |\widetilde{\varepsilon}_\tau(x)| \, \mathrm{d}x - \int_{\mathbb{R}^d} |\widetilde{\varepsilon}_T(x)| \, \mathrm{d}x$$
$$\leq \sum_{i=0}^{N-1} \int_{T-t_{i+1}}^{T-t_i} \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t \delta_t))(x) \right| \mathrm{d}x\mathrm{d}t + \sum_{i=0}^{N-1} \int_{T-t_{i+1}}^{T-t_i} \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t(\widetilde{\delta}_t - \delta_t)))(x) \right| \mathrm{d}x\mathrm{d}t.$$

By using the fact that $\widetilde{\delta}_t - \delta_t = \widetilde{U}_t - \widehat{U}_t = \widetilde{V}_{T-t} - \widehat{V}_{T-t}$, the discretization error becomes

$$\int_{T-t_{i+1}}^{T-t_i} \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t(\widetilde{\delta}_t - \delta_t)))(x) \right| \mathrm{d}x\mathrm{d}t = \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_{T-t}(\widetilde{V}_t - V_t))(x) \right| \mathrm{d}x\mathrm{d}t.$$

Using the Runge-Kutta local truncation error estimations from (14), the discretization error can be bounded as

$$\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_{T-t}(\widetilde{V}_t - V_t))(x) \right| \mathrm{d}x\mathrm{d}t \leq C(p,s)d \cdot (dh)^p (LD)^{p+1} \int_{t_i}^{t_{i+1}} \frac{1}{\sigma_t^2} \mathrm{d}t.$$

As a result, the score matching error and time discretization error do not interact to magnify, thereby preserving the time discretization error at $p$-th order, albeit with significant dimensionality dependence. The detailed proof is in Appendix D.

## 4. Numerical Study

In this section, we numerically analyze the convergence rate of the probability flow ODE, specifically focusing on a $K$-mode Gaussian mixture target distribution

$$q_0(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x; m_k, C_k). \tag{15}$$

The forward process, as denoted in (2) with $\lambda_t = e^{-t}$ and $\sigma_t = \sqrt{1 - e^{-2t}}$, yields

$$q_t(y) = \int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot \exp\left( -\frac{\|y - \lambda_t x\|_2^2}{2\sigma_t^2} \right) q_0(x) \mathrm{d}x = \sum_{k=1}^{K} w_k \mathcal{N}(y; \lambda_t m_k, \lambda_t^2 C_k + \sigma_t^2 I).$$

The score function takes the following analytical format

$$\nabla_x \log q_t(x) = -\sum_{k=1}^{K} \frac{w_k \mathcal{N}(x; \lambda_t m_k, \lambda_t^2 C_k + \sigma_t^2 I)}{q_t(x)} (\lambda_t^2 C_k + \sigma_t^2 I)^{-1} (x - \lambda_t m_k). \tag{16}$$

In general, the score function is represented by a neural network with inputs $t$ and $x$, trained through score matching with sequentially corrupted training data [19, 37, 38, 42]. However, in our present work, we circumvent the score matching step. Instead we assume that we have access to an imperfect score function characterized by the following three types of artificial score errors $\delta(t,x) = s_{T-t}(x) - \nabla \log q_t(x)$

- constant error : $\delta(t,x) = \delta \frac{1}{\sqrt{d}}$;
- linear error: $\delta(t,x) = \delta \frac{x-m}{\sqrt{d}}$;
- sinusoidal error: $\delta(t,x) = \delta \sin(x) \frac{x-m}{\sqrt{d}}$.

Here $m$ is the mean of the target Gaussian mixture distribution. The sin function is used for pointwise evaluation, and its product with the following term also represents pointwise multiplication. In the subsequent numerical investigation, we evaluate the convergence rate of the probability flow ODE for estimating $q_0$ using an analytical score function (16) with various magnitudes of artificial score errors parameterized by a scalar $\delta$. Specifically, we consider $\delta$ values of 0.005, 0.01, 0.02, 0.04, 0.08, and 0.16. For the probability flow ODE, we integrate the deterministic reverse process (6) using Heun's second-order time integrator until $T = 8$. Because the Gaussian mixture target density has no singularities, we integrate the reverse process until the final time $T$ with $\tau = 0$. Based on our theoretical analysis outlined in Theorem 3.9, to balance the score error and time discretization error, we choose $h^2$ to be approximately equal to $\delta$. Consequently, as we vary the magnitude of the score error $\delta$ from 0.005 to 0.16, we adjust the corresponding number of time steps as follows: $N_t = 96$, 64, 48, 32, 24, and 16. It is worth noticing the step size must be within the stability regime of the explicit time integrator. To initialize the ODE flow, we sample $J = 4 \times 10^4$ particles from the standard Gaussian distribution $\mathcal{N}(x; 0, \mathbb{I}_d)$. All code used to produce the numerical results and figures in this paper are available at `https://github.com/Zhengyu-Huang/InverseProblems.jl/tree/master/Diffusion/Gaussian-mixture-density.ipynb`.

4.1. **One Dimensional Test.** We first consider a one dimensional 3-mode Gaussian mixture (15) with

$$w = [0.1;\, 0.4;\, 0.5] \quad m = [-6.0;\, 4.0;\, 6.0] \quad \text{and} \quad C = [0.25;\, 0.25;\, 0.25].$$

Initially, we explore the convergence behavior of the probability flow ODE in the mean field limit by numerically solving the Fokker-Planck PDE (11). We employ an initial distribution $\mathcal{N}(x; 0, 1)$ and discretize the computational domain $[-10, 10]$ into 1000 cells using a second-order finite volume method [41]. Integration is performed with Heun's second-order time integrator using a time step of $h = 10^{-3}$ until $T = 8$. To ensure accurate understanding at the continuous level, we choose $\Delta x$ and $h$ to be sufficiently small, such that discretization errors are negligible compared to the score error. The corresponding score function, reference density $q_t$, and estimated density $\widehat{\varrho}_{T-t}$ under various imperfect score estimations are illustrated in Fig. 1. While the solution error increases with larger $\delta$, all modes are captured, including the left-side mode with relatively small density. The convergence rate, assessed in terms of the total variation $\mathrm{TV}(q_0, \widehat{\varrho}_T)$, relative mean error, and relative covariance error, is depicted in Fig. 2. The linear relationship between these errors and $\delta$ is clearly demonstrated.

Then, we investigate the convergence of the probability ODE flow by integrating the deterministic reverse process (6). The corresponding score function, reference density $q_t$, and estimated density $\widehat{\varrho}_{T-t}$ with various imperfect score estimations are depicted in Fig. 3. Kernel density estimates are computed with bandwidth determined by Silverman's rule [34] $\left(\frac{4}{J(d+2)}\right)^{1/(d+4)}(1 - \frac{t}{2T})$ (interpolating from 0.5 to 1). Notably, the estimated densities closely resemble the PDE solution (See Fig. 1), highlighting the significant efficiency of the time integrator with such small numbers of time steps. The convergence rate, evaluated in terms of the total variation $\mathrm{TV}(q_0, \widehat{\varrho}_T)$, relative mean error, and relative covariance error, is depicted in Fig. 4. The linear relationship between these errors and $\delta$ (or $h^2$) is clearly demonstrated.

4.2. **High Dimensional Test.** Finally, we consider $d$ dimensional 5-mode Gaussian mixtures (15). The weights are sampled uniformly $w_k \sim \mathrm{Uniform}[0, 1]$ are then normalized to sum to 1. The means are generated from a Gaussian distribution, $m_k \sim \mathcal{N}(0, 3^2\mathbb{I}_d)$, and the covariance matrices $C_k$ are generated as $C_k = \frac{1}{8}(W_k^T W_k/d + \mathbb{I}_d)$ with $(W_k)_{ij} \sim \mathcal{N}(0, 1)$ for $i, j = 1, \cdots, d$.

We investigate the convergence of the ODE flow by integrating the deterministic reverse process (6) with the same setup as before. For $d = 128$, we visualize the results in terms of the marginal density for the first dimension, including its score, reference density $q_t^1$, and estimated density $\widehat{\varrho}_{T-t}^1$ with various artificial score errors, as depicted in Fig. 5. We compute kernel density estimates

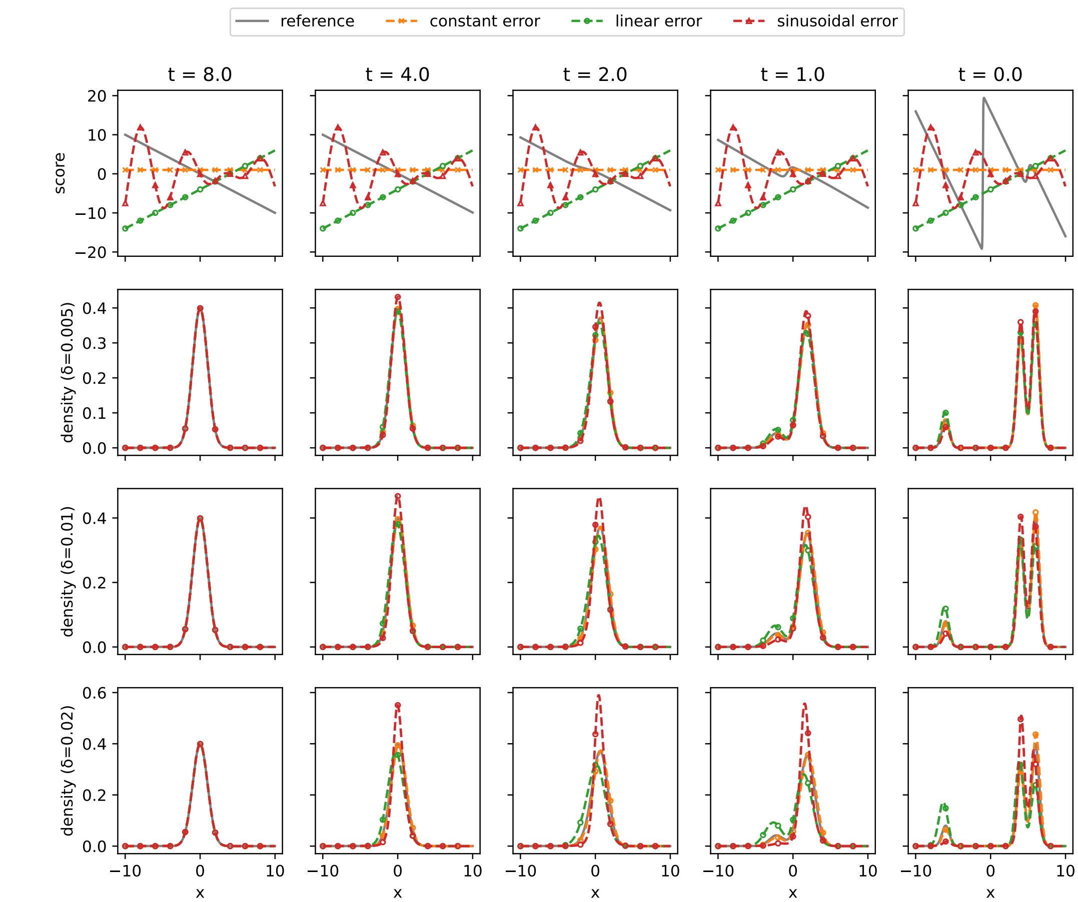


FIGURE 1. One dimensional test: Density estimations obtained by solving the Fokker-Planck PDE (11) numerically with various artificial score errors. From top to bottom: score, estimated density with $\delta = 0.005$, 0.01, 0.02. From left to right estimated $q_t$ at $t = 8$, 4, 2, 1, 0.

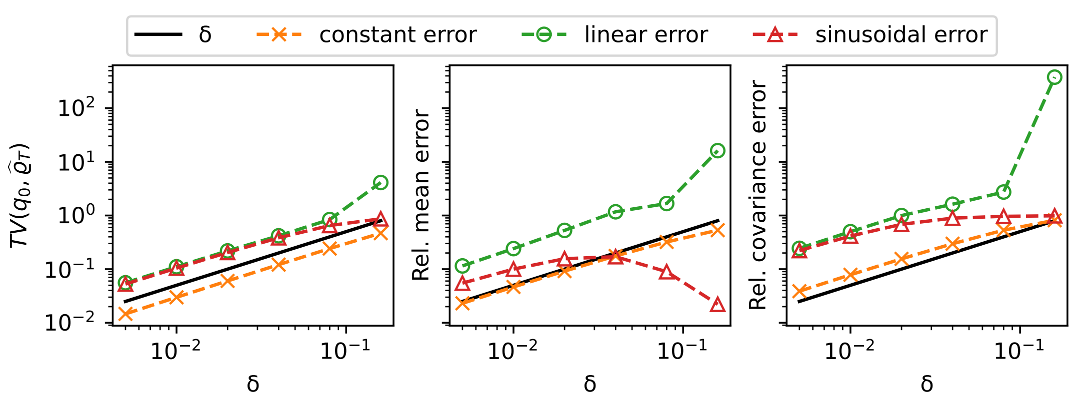


FIGURE 2. One dimensional test: convergence of the density estimations obtained by solving the Fokker-Planck PDE (11) numerically with various artificial score errors.

with the bandwidth determined by the same Silverman's rule [34] as in Section 4.1. Surprisingly, the estimated densities are as good as those of the one-dimensional test (See Fig. 3). We further

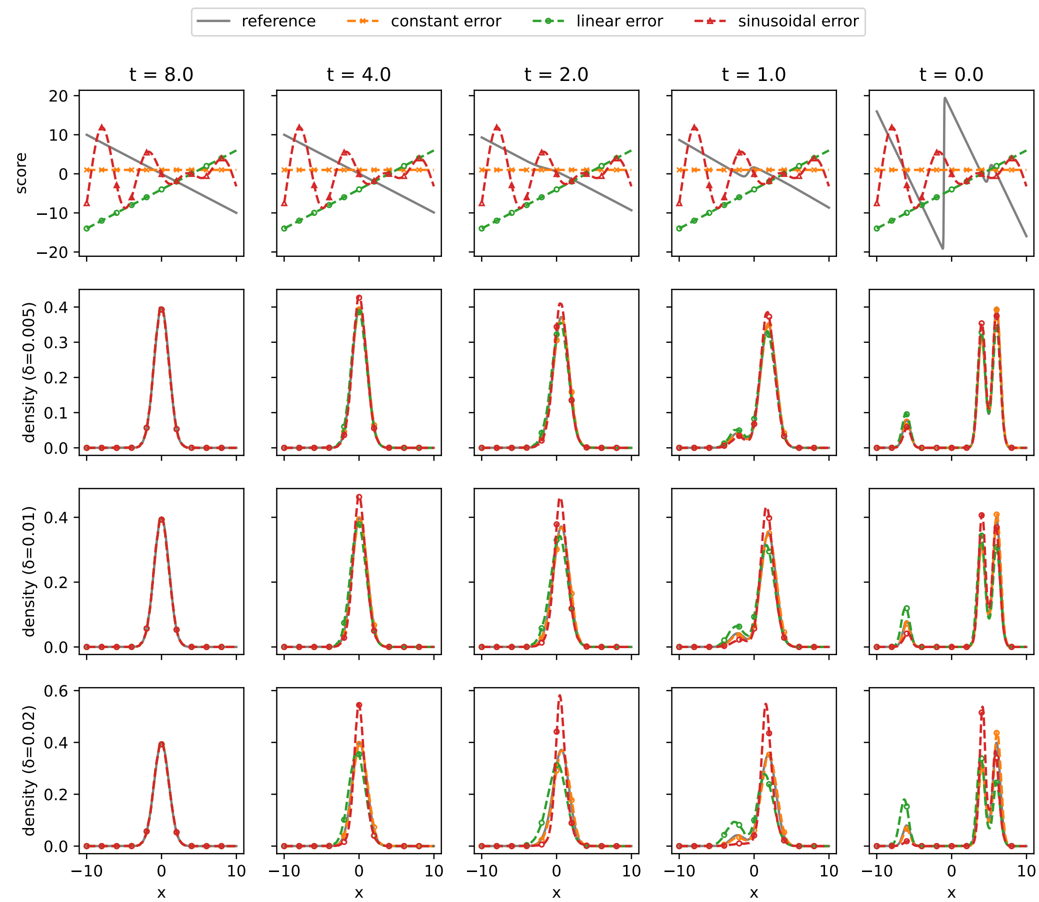


FIGURE 3. One dimensional test: density estimations obtained by solving the probability flow ODE with Heun's method with various artificial score errors. From top to bottom: score, estimated density with $\delta = 0.005$, 0.01, 0.02. From left to right estimated $q_t$ at $t = 8$, 4, 2, 1, 0.

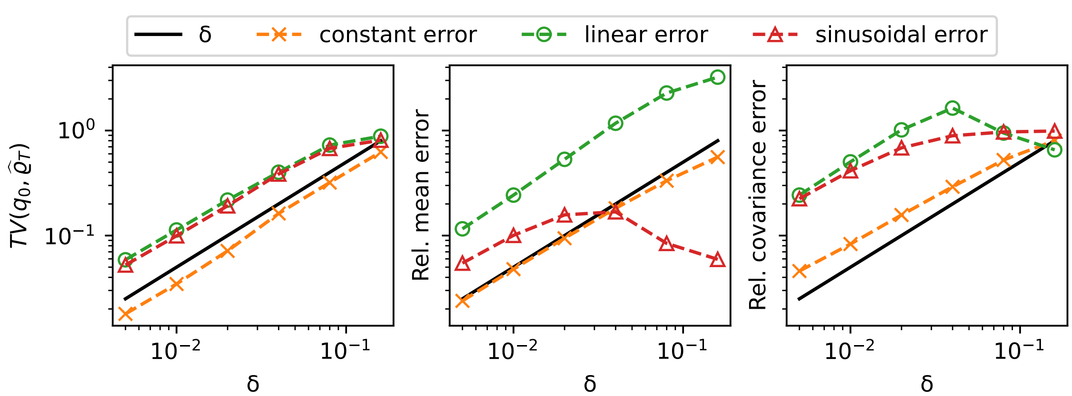


FIGURE 4. One dimensional test: convergence of the density estimations obtained by solving the probability flow ODE with Heun's method, where $h^2 \approx \delta$ with various artificial score errors.

consider $d = 8$ and 32 by generating the target density $q_0$ through marginalizing the 128-dimensional Gaussian mixture target density. The convergence rate, evaluated in terms of the total variation of the marginal density $\mathrm{TV}(q_0^1, \widehat{\varrho}_T^1)$, relative mean error, and relative covariance error, is depicted in Fig. 6. The linear relationship between these errors and $\delta$ (or $h^2$) is clearly demonstrated. Additionally, we also explore the scenario without score matching errors for comprehensive analysis. The convergence rate, evaluated using the same error indicators, is presented in Fig. 7, showing the quadratic relationship with $h^2$. Notably, in this high dimensional test study, we do not observe any dimension dependence.

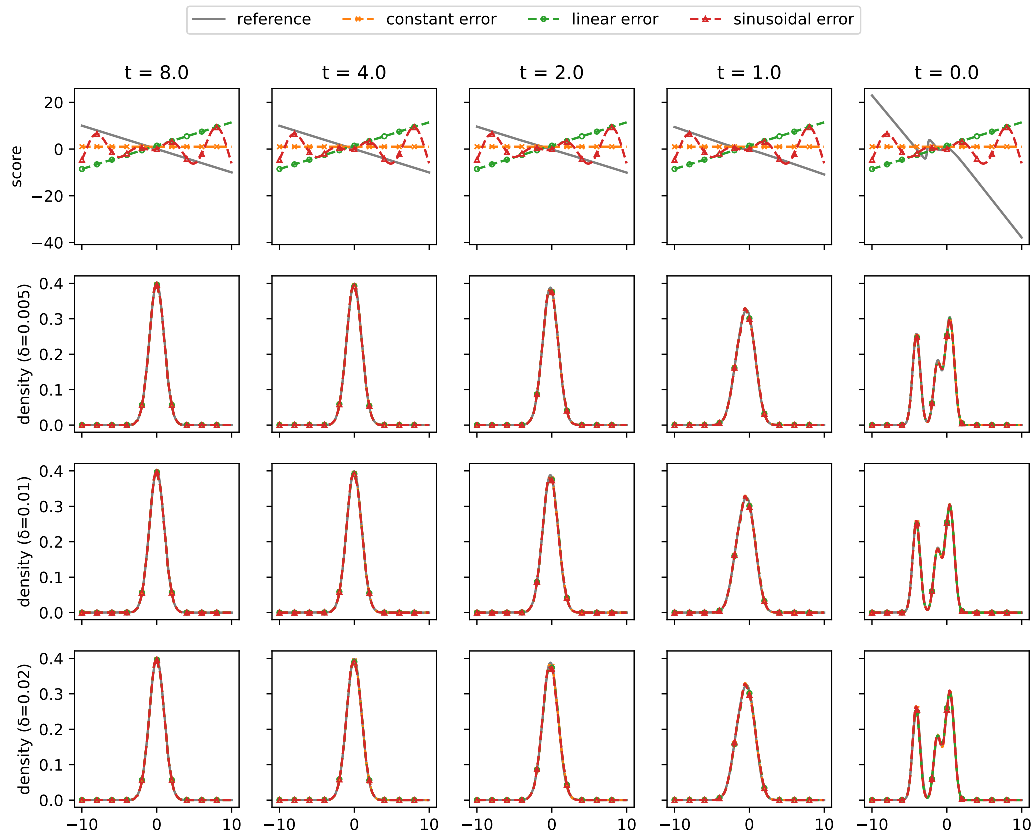


FIGURE 5. 128 dimensional test: marginal densities obtained by solving the probability flow ODE with Heun's method with various artificial score errors. From top to bottom: score, estimated densities with $\delta = 0.005$, 0.01, 0.02. From left to right estimated $q_t$ at $t = 8$, 4, 2, 1, 0.

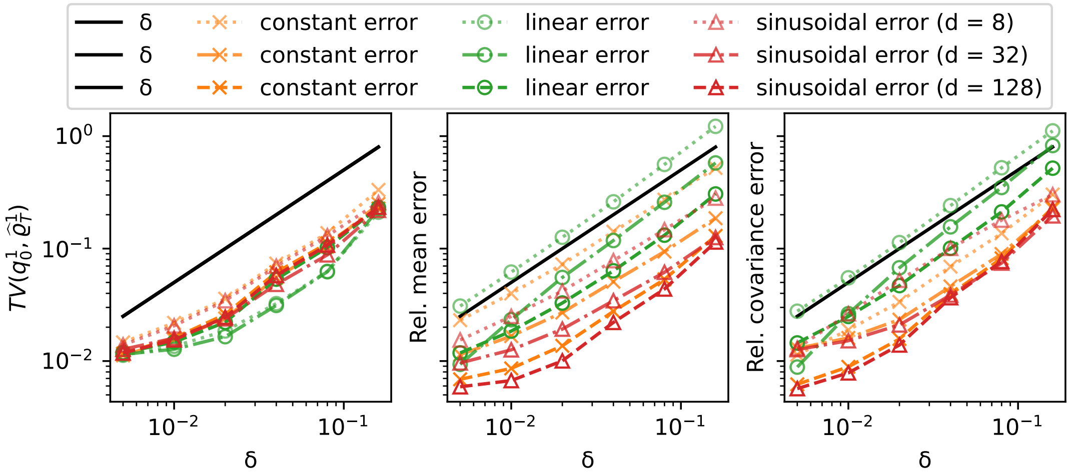


FIGURE 6. High dimensional test: convergence of the density estimations obtained by solving the probability flow ODE with Heun's method, where $h^2 \approx \delta$ with various artificial score errors. The dotted lines, dash dot lines and dashed lines indicate $d = 8, 32$ and, 128 dimension tests, respectively.

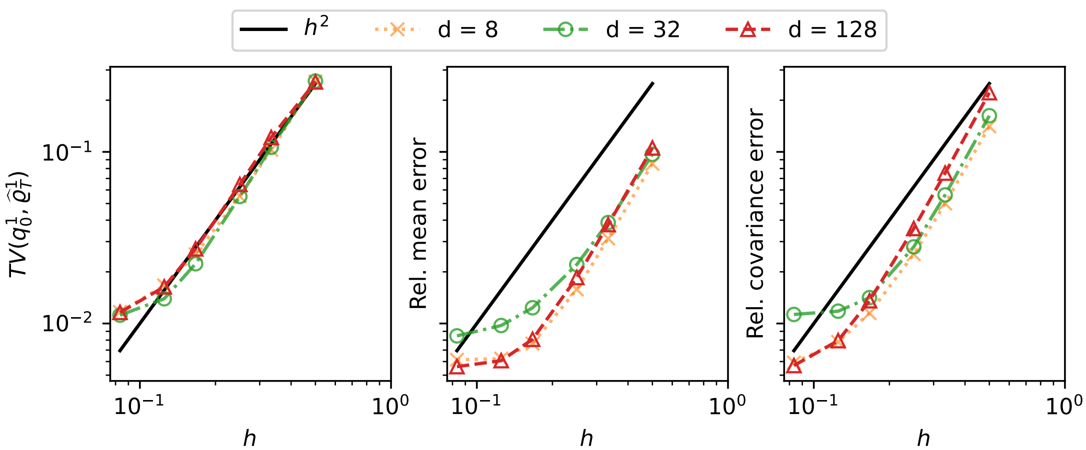


FIGURE 7. High dimensional test: convergence of the density estimations with no score errors obtained by solving the probability flow ODE with Heun's method, across different time step sizes $h$. The dotted cross lines, dash dot circle lines and dashed triangle lines indicate $d = 8, 32$ and, 128 dimension tests, respectively.

## 5. CONCLUSION

In this study, we have investigated the convergence of score-based generative model based on the probability flow ODE, both theoretically and numerically. Our analysis provided theoretical convergence guarantees at both continuous and discrete levels. Additionally, our numerical studies, conducted on problems up to 128 dimensions, provided empirical verification of our theoretical findings. One notable observation is the superior error bound of $\mathcal{O}(\delta + h^p)$, indicating potential sharper estimations with improved dimension and score matching error dependence. Moreover, conducting rigorous numerical-based analyses with neural network-based score estimation errors is also a focus for future research.

**Acknowledgments** DZH is supported by the high-performance computing platform of Peking University. JH is supported by NSF grant DMS-2331096, and the Sloan research fellowship. ZJL would like to thank Professor Guido De Philippis for the insightful discussions.

## References

[1] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[2] M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[3] L. Ambrosio. Transport equation and Cauchy problem for BV vector fields. *Invent. Math.*, 158(2):227–260, 2004.

[4] J. Benton, G. Deligiannidis, and A. Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

[5] A. Block, Y. Mroueh, and A. Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

[6] N. Boffi and E. Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *arXiv preprint arXiv:2206.04642*, 2022.

[7] H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[8] S. Chen, S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[10] S. Chen, G. Daras, and A. Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023.

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

[12] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[13] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[14] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[15] A. Fiorenza, M. R. Formica, T. G. Roskovec, and F. Soudský. Detailed proof of classical Gagliardo-Nirenberg interpolation inequality with historical remarks. *Z. Anal. Anwend.*, 40(2):217–236, 2021.

[16] X. Gao and L. Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] D. Kwon, Y. Fan, and K. Lee. Score-based generative modeling secretly minimizes the wasserstein distance. *Advances in Neural Information Processing Systems*, 35:20205–20217, 2022.

[22] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

[23] H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[24] G. Li, Y. Huang, T. Efimov, Y. Wei, Y. Chi, and Y. Chen. Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024.

[25] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

[26] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[27] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[28] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.
[29] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.
[30] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.
[31] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.
[32] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
[33] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
[34] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.
[35] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
[36] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
[37] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
[38] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
[39] W. Tang and H. Zhao. Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*, 2024.
[40] W. Tang and H. Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. *arXiv preprint arXiv:2402.07487*, 2024.
[41] B. Van Leer. Towards the ultimate conservative difference scheme. ii. monotonicity and conservation combined in a second-order scheme. *Journal of computational physics*, 14(4):361–370, 1974.
[42] P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.
[43] K. Y. Yang and A. Wibisono. Convergence of the inexact langevin algorithm and score-based generative models in kl divergence. *arXiv preprint arXiv:2211.01512*, 2022.
[44] Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.
[45] W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.


## Appendix A. Total Variation Estimates along Probability Flow

In this section, we prove the following general theorem for continuity equations, which is independent of the particular choice of $U_t, \widehat{U}_t$ (equivalently, $V_{T-t}, \widehat{V}_{T-t}$) in (11).

**Theorem A.1.** *Fix any $0 < \tau < T$. Let $\widehat{q}_t(x), q_t(x) \in C^1([\tau, T] \times \mathbb{R}^d) \cap L^1([\tau, T] \times \mathbb{R}^d)$ solve the following two continuity equations on $\mathbb{R}^d$ respectively,*

$$\partial_t q_t = \nabla \cdot (U_t q_t), \quad \partial_t \widehat{q}_t = \nabla \cdot (\widehat{U}_t \widehat{q}_t).$$

*We also assume that for $t \in [\tau, T]$, $U_t, \widehat{U}_t$ are locally Lipschitz on $\mathbb{R}^d$. Then, if we denote $\delta_t(x) \coloneqq \widehat{U}_t(x) - U_t(x)$, $\widehat{\varepsilon}_t(x) \coloneqq \widehat{q}_t(x) - q_t(x)$, we have that,*

$$\left| \int_{\mathbb{R}^d} |\widehat{\varepsilon}_\tau(x)| \, \mathrm{d}x - \int_{\mathbb{R}^d} |\widehat{\varepsilon}_T(x)| \, \mathrm{d}x \right| \le \int_\tau^T E(t) \, \mathrm{d}t, \quad \text{with } E(t) \coloneqq \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t \delta_t))(x) \right| \mathrm{d}x.$$

*First Proof.* Fix this given $\tau > 0$. For $t > 0$, we define the total variation between $\widehat{q}_t(x)$ and $q_t(x)$ as

$$f(t) \coloneqq \int_{\mathbb{R}^d} |\widehat{\varepsilon}_{t+\tau}(x)| \, \mathrm{d}x. \tag{17}$$

We use $x_t \coloneqq T_t(x)$ to denote the solution of the ODE (also called the characteristic line)

$$\frac{\mathrm{d}}{\mathrm{d}t} T_t(x) = -\widehat{U}_{\tau+t}(T_t(x)),$$

with the initial data $x_0 = T_0(x) = x$. We notice that, by change of variables (we can do this because $\widehat{U}_t$ is locally Lipschitz on $\mathbb{R}^d$. So, $T_t$ is a diffeomorphism of $\mathbb{R}^d$),

$$f(t) = \int_{\mathbb{R}^d} |\widehat{\varepsilon}_{t+\tau}(x_t)| JT_t(x)\ \mathrm{d}x,$$

where $JT_t(x) = |\det(\nabla T_t(x))|$. A direct computation shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\varepsilon}_{t+\tau}(x) = (\nabla\cdot(\widehat{U}_{t+\tau}\widehat{\varepsilon}_{t+\tau}))(x) + (\nabla\cdot(q_{t+\tau}\delta_{t+\tau}))(x).$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\varepsilon}_{t+\tau}(x_t) = (\nabla\cdot\widehat{U}_{t+\tau})(x_t)\cdot\widehat{\varepsilon}_{t+\tau}(x_t) + (\nabla\cdot(q_{t+\tau}\delta_{t+\tau}))(x_t).$$

Jacobi's formula gives that

$$\frac{\mathrm{d}}{\mathrm{d}t}\det(\nabla T_t(x)) = \mathrm{trace}\bigg(\Big(\frac{\mathrm{d}}{\mathrm{d}t}\nabla T_t(x)\Big)\cdot\Big(\nabla T_t(x)\Big)^{-1}\bigg)\cdot\det(\nabla T_t(x)).$$

We also notice that

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla T_t(x) = \nabla\frac{\mathrm{d}}{\mathrm{d}t}T_t(x) = -\nabla(\widehat{U}_{\tau+t}(T_t(x))) = -(\nabla\widehat{U}_{\tau+t})(T_t(x))\cdot\nabla T_t(x).$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\det(\nabla T_t(x)) = -\mathrm{trace}\Big((\nabla\widehat{U}_{\tau+t})(T_t(x))\Big)\cdot\det(\nabla T_t(x)) = -(\nabla\cdot\widehat{U}_{\tau+t})(T_t(x))\cdot\det(\nabla T_t(x)).$$

So,

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}f(t)\right| &\le \int_{\mathbb{R}^d}\Big|\big[(\nabla\cdot\widehat{U}_{t+\tau})(x_t)\cdot\widehat{\varepsilon}_{t+\tau}(x_t) + (\nabla\cdot(q_{t+\tau}\delta_{t+\tau}))(x_t)\big]\cdot\det(\nabla T_t(x))\\ &\qquad + \widehat{\varepsilon}_{t+\tau}(x_t)\cdot(-(\nabla\cdot\widehat{U}_{\tau+t})(T_t(x)))\cdot\det(\nabla T_t(x))\Big|\ \mathrm{d}x\\ &= \int_{\mathbb{R}^d}\Big|(\nabla\cdot(q_{t+\tau}\delta_{t+\tau}))(x_t)\Big|\cdot JT_t(x)\ \mathrm{d}x\\ &= \int_{\mathbb{R}^d}\Big|(\nabla\cdot(q_{t+\tau}\delta_{t+\tau}))(x)\Big|\ \mathrm{d}x = E(t+\tau),\end{aligned}$$

where the second last equality is by change of variables again. Hence, for any $t' > 0$,

$$f(0) \le \int_0^{t'} E(t+\tau)\ \mathrm{d}t + f(t'). \tag{18}$$

One can also use the characteristic line method starting from $t = T$ to $t = \tau$, and then obtain an inverse inequality. Hence, we finish the proof of the theorem. ☐

**Remark A.2.** *Notice that in this first proof, we exchange the order of integrals and derivatives. We can do this in our problem setting (11), because by Assumption 3.1, $\mu_*$ is compactly supported, so $q_t(x)$ always has a exponential tail as $\|x\| \to +\infty$. $\widehat{q}_t(x)$ also has such a property by our Lemma B.6.*

Indeed, one can prove the following more general theorem with reasonable assumptions, essentially verbatim, by the same method.

**Theorem A.3.** *Fix any $0 < \tau < T$. Let $p_t(x) \in C^1([\tau, T]\times\mathbb{R}^d)\cap L^1([\tau, T]\times\mathbb{R}^d)$ solve the following continuity equation on $\mathbb{R}^d$ with $h_t(x) \in L^1([\tau, T]\times\mathbb{R}^d)$,*

$$\partial_t p_t(x) = (\nabla\cdot(Z_t p_t))(x) + h_t(x).$$

*We also assume that for $t \in [\tau, T]$, $Z_t$ is locally Lipschitz on $\mathbb{R}^d$. Then, for almost all $t \in [\tau, T]$, we have that,*

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} |p_t(x)|\ \mathrm{d}x\right| \leq \int_{\mathbb{R}^d} |h_t(x)|\ \mathrm{d}x.$$

*Hence,*

$$\left|\int_{\mathbb{R}^d} |p_\tau(x)|\mathrm{d}x - \int_{\mathbb{R}^d} |p_T(x)|\mathrm{d}x\right| \leq \int_\tau^T \int_{\mathbb{R}^d} |h_t(x)|\ \mathrm{d}x.$$

**Remark A.4.** *We will prove a Gagliardo-Nirenberg interpolation inequality with dimension free constants in Lemma C.1. With this inequality, if one can know that $\int_{\mathbb{R}^d} |\partial^2_{ii} p_\tau(x)|\mathrm{d}x$ is bounded for some $i \in [\![1, d]\!]$, then one can use Lemma C.1 to see that $\int_{\mathbb{R}^d} |\partial_i p_\tau(x)|\mathrm{d}x$ is also small if $\int_{\mathbb{R}^d} |p_\tau(x)|\mathrm{d}x$ is small. Similar arguments work for higher order derivatives of $p_\tau(x)$. Moreover, if we can conclude that the $W^{k,1}(\mathbb{R}^d)$-norms of $p_\tau(x)$ is small, where $W^{k,1}$ means Sobolev spaces, we can use Sobolev inequalities to conclude that the $W^{k-1,r}(\mathbb{R}^d)$-norms of $p_\tau(x)$ for a corresponding $r > 1$ is also mall. In particular, we can conclude that the $L^r(\mathbb{R}^d)$-norm of $p_\tau(x)$ is small.*

Before we give the second proof of Theorem A.1 and Theorem A.3, we need to explain our intuition a little bit. With those notations in Theore A.3, we notice that for

$$g(t) := \int_{\mathbb{R}^d} |p_t(x)|\ \mathrm{d}x,$$

we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t) = \int_{\mathbb{R}^d} \mathrm{sign}(p_t(x))\frac{\mathrm{d}}{\mathrm{d}t}p_t(x)\ \mathrm{d}x = \int_{\{p_t \geq 0\}} \frac{\mathrm{d}}{\mathrm{d}t}p_t(x)\ \mathrm{d}x - \int_{\{p_t < 0\}} \frac{\mathrm{d}}{\mathrm{d}t}p_t(x)\ \mathrm{d}x.$$

If the boundary $\partial\{p_t \geq 0\}$ consists of $(d-1)$-dimensional piecewise smooth submanifolds (or at least rectifiable sets), then by the divergence theorem, we know that

$$\int_{\{p_t \geq 0\}} (\nabla \cdot (Z_t p_t))(x)\ \mathrm{d}x = \int_{\partial\{p_t \geq 0\}} \mathbf{n}(x) \cdot (Z_t p_t)(x)\ d\mathcal{H}^{d-1}(x) = 0,$$

where $\mathbf{n}(x)$ is the unit outer normal vector and $\mathcal{H}^{d-1}(\cdot)$ is the $(d-1)$-dimensional surface measure. Similarly,

$$\int_{\{p_t < 0\}} (\nabla \cdot (Z_t p_t))(x)\ \mathrm{d}x = 0.$$

So,

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t) = \int_{\mathbb{R}^d} \mathrm{sign}(p_t(x))h_t(x)\ \mathrm{d}x, \text{ and } \left|\frac{\mathrm{d}}{\mathrm{d}t}g(t)\right| \leq \int_{\mathbb{R}^d} |h_t(x)|\ \mathrm{d}x.$$

However, in general, we cannot know whether the boundary set $\partial\{p_t \geq 0\}$ is always of $(d-1)$-dimension. For an arbitrarily given smooth function, its zero level set can also be arbitrarily strange and does not necessarily need to be of $(d-1)$-dimension.

The following second proof of Theorem A.3 is inspired by [3] and the communication with Professor Guido De Philippis. We also assume that $(\|Z_t(x)\| \cdot |p_t(x)|) \in L^1(\mathbb{R}^d \times [\tau, T])$.

*Second Proof.* Let $a > 0$. We let $\beta_a(s) = \sqrt{a^2 + s^2}$ be a function on $\mathbb{R}$ which approximates the function $|s|$ as $a \to 0^+$. Then, the function $p_{a,t}(x) := \beta_a(p_t(x))$ on $\mathbb{R}^d$ solves the equation

$$\begin{aligned}\partial_t p_{a,t}(x) &= ((\nabla \cdot Z_t)(x))p_t(x)\beta_a'(p_t(x)) + Z_t(x) \cdot \nabla p_{a,t}(x) + h_t(x)\beta_a'(p_t(x)) \\ &= \nabla \cdot (Z_t(x)p_{a,t}(x)) + ((\nabla \cdot Z_t)(x))[p_t(x)\beta_a'(p_t(x)) - p_{a,t}(x)] + h_t(x)\beta_a'(p_t(x)).\end{aligned}$$

For any $R > 0$, consider the integral of the above equation on the ball $B_R \coloneqq B_R(0) \subset \mathbb{R}^d$, we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{B_R} p_{a,t}(x)\mathrm{d}x = &\int_{\partial B_R} (Z_t(x)p_{a,t}(x))\cdot \frac{x}{\|x\|} d\mathcal{H}^{d-1}(x) + \int_{B_R} ((\nabla\cdot Z_t)(x))[p_t(x)\beta_a'(p_t(x)) - p_{a,t}(x)]\,\mathrm{d}x\\ &+ \int_{B_R} h_t(x)\beta_a'(p_t(x))\,\mathrm{d}x,\end{aligned}$$

and then for any $t', t'' \in [\tau, T]$,

$$\begin{aligned}\left|\int_{B_R} p_{a,t''}(x)\mathrm{d}x - \int_{B_R} p_{a,t'}(x)\mathrm{d}x\right| = \Bigg|&\int_{t''}^{t'}\int_{\partial B_R} (Z_t(x)p_{a,t}(x))\cdot\frac{x}{\|x\|}d\mathcal{H}^{d-1}(x)\mathrm{d}t\\ &+ \int_{t''}^{t'}\int_{B_R} ((\nabla\cdot Z_t)(x))[p_t(x)\beta_a'(p_t(x)) - p_{a,t}(x)]\,\mathrm{d}x\mathrm{d}t\\ &+ \int_{t''}^{t'}\int_{B_R} h_t(x)\beta_a'(p_t(x))\,\mathrm{d}x\mathrm{d}t\Bigg|.\end{aligned}$$

Notice that $|\beta_a'(s)| = \frac{|s|}{\sqrt{a^2+s^2}} \le 1$, $|p_t(x)\beta_a'(p_t(x)) - p_{a,t}(x)| = \frac{a^2}{\sqrt{a^2+p_t(x)^2}} \le 1$, and for any given $x \in B_R$, $\lim_{a\to 0^+} \frac{a^2}{\sqrt{a^2+p_t(x)^2}} = 0$. By the dominated convergence theorem, we can let $a \to 0^+$ on both sides and obtain that

$$\begin{aligned}\left|\int_{B_R} |p_{t''}(x)|\mathrm{d}x - \int_{B_R} |p_{t'}(x)|\mathrm{d}x\right| = \Bigg|&\int_{t''}^{t'}\int_{\partial B_R} (Z_t(x)|p_t(x)|)\cdot\frac{x}{\|x\|}d\mathcal{H}^{d-1}(x)\mathrm{d}t\\ &+ \int_{t''}^{t'}\int_{B_R} ((\nabla\cdot Z_t)(x))\cdot 0\,\mathrm{d}x\mathrm{d}t\\ &+ \int_{t''}^{t'}\int_{B_R} h_t(x)\,\mathrm{sign}(p_t(x))\,\mathrm{d}x\mathrm{d}t\Bigg|\\ \le &\int_{t''}^{t'}\int_{\partial B_R} \|Z_t(x)\|\cdot|p_t(x)|d\mathcal{H}^{d-1}(x)\mathrm{d}t\\ &+ \int_{t''}^{t'}\int_{B_R} |h_t(x)|\,\mathrm{d}x\mathrm{d}t.\end{aligned}$$

Because we assumed that $(\|Z_t(x)\|\cdot|p_t(x)|) \in L^1(\mathbb{R}^d\times[\tau,T])$, we can choose a sequence $\{R_j\}$ with $R_j \to +\infty$, such that the first term on the right hand side goes to 0. So, by the monotone convergence theorem, after passing $R_j \to +\infty$, we obtain that for any $t', t'' \in [\tau, T]$,

$$\left|\int_{\mathbb{R}^d} |p_{t''}(x)|\mathrm{d}x - \int_{\mathbb{R}^d} |p_{t'}(x)|\mathrm{d}x\right| \le \int_{t''}^{t'}\int_{\mathbb{R}^d} |h_t(x)|\,\mathrm{d}x\mathrm{d}t.$$

□

## Appendix B. Preliminary Estimates on Forward and Backward Density

In this section, $U_t, \widehat{U}_t$ are defined as in (11). Under Assumption 3.1, we take $\lambda_t = e^{-t}$ and $\sigma_t = \sqrt{1-\lambda_t^2}$ $(t > 0)$, and we also assume that

$$q_t(y) \coloneqq \int_{\mathbb{R}^d} \frac{1}{\left(\sqrt{2\pi}\sigma_t\right)^d}\cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}}\cdot\mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x, \tag{19}$$

which satisfies the first-order PDE: $\partial_t q_t = \nabla\cdot(U_t q_t)$ for $U_t = x + \nabla\log q_t(x)$. We remark that although we use the notation $\mu_*(x/\lambda_t^d)/\lambda_t^d$ in the definition (19) of $q_t$, because the data distribution

$\mu_*$ can be supported on a submanifold $K_*$, or more general lower dimensional rectifiable sets, the meaning of it is actually the push-forward measure defined by $(\lambda_t)_\# \mu_*(A) := \mu_*(\lambda_t^{-1} A)$ for any measurable set $A \subset \mathbb{R}^d$. So, the rescaling factor $\lambda_t^d$ is actually $\lambda_t^k$ for $k \le d$ when $K_*$ is a $k-$dimensional submanifold in $\mathbb{R}^d$. But this notation doesn't affect our computations. As readers will see, the only property we will use is that in the integrand of (19), $x/\lambda_t \in K_*$ and hence $\|x\|_\infty \le \lambda_t D$.

**Example B.1.** *When $\mu_*$ is the delta mass at a point $y_0 \in \mathbb{R}^d$, then for $t > 0$,*

$$q_t(y) = \frac{1}{\left(\sqrt{2\pi}\sigma_t\right)^d} \cdot e^{-\frac{\|y-y_0\|^2}{2\sigma_t{}^2}}, \quad \nabla q_t(y) = -\frac{y-y_0}{\sigma_t^2} q_t(y), \quad \nabla^2 \log q_t(y) = -\frac{1}{\sigma_t^2} \cdot \mathbb{I}_d.$$

*We notice that as $t \to 0^+$, the derivatives of $\log q_t(y)$ blow up.*

**Example B.2.** *When $\mu_*$ is the unit 2-sphere $\mathbb{S}^2 \subset \mathbb{R}^3 \subset \mathbb{R}^d$, if we write $y = (y', y'') \in \mathbb{R}^3 \times \mathbb{R}^{d-3}$ and $y' = (y_1, y_2, y_3)$, a direct computation shows that*

$$q_t(y) = \frac{2\pi}{\left(\sqrt{2\pi}\sigma_t\right)^d} \cdot e^{-\frac{\|y\|^2 + \lambda_t^2}{2\sigma_t{}^2}} \cdot \frac{\sigma_t^2}{\lambda_t \|y'\|} \cdot \left( e^{\frac{\lambda_t \|y'\|}{\sigma_t^2}} - e^{-\frac{\lambda_t \|y'\|}{\sigma_t^2}} \right).$$

*The derivatives of $\log q_t(y)$ also blow up as $t \to 0^+$ by a direct computation.*

In general, we have the following estimates for space directions derivatives of $q_t(y)$ and $\log q_t(y)$. An interesting fact is that the upper bound in the statement of Lemma B.3 is uniform for $y \in \mathbb{R}^d$.

**Lemma B.3.** *For any $p \ge 3$, any $y \in \mathbb{R}^d$ and any $t > 0$,*

$$\|\nabla^p \log q_t(y)\|_\infty \le (4p!) \frac{\lambda_t{}^p D^p}{\sigma_t{}^{2p}}. \tag{20}$$

*Also, for any $i, j$,*

$$|\partial_{y_i} \log q_t(y)| \le \frac{|y_i| + \lambda_t D}{\sigma_t^2}, \text{ and } |\partial^2_{y_i y_j} \log q_t(y)| \le \frac{\delta_{ij}}{\sigma_t^2} + 2\frac{\lambda_t^2 D^2}{\sigma_t^4}.$$

*Proof.* For simplicity, for a function $f(x)$ defined on $\mathbb{R}^d$, we denote

$$\langle f(x) \rangle := \int_{\mathbb{R}^d} f(x) \, \mathrm{d}(\lambda_t)_\# \mu_*(x) = \int_{\mathbb{R}^d} f(x) \mu_* \left( \frac{x}{\lambda_t} \right) \frac{1}{\lambda_t{}^d} \, \mathrm{d}x. \tag{21}$$

Hence, $q_t(y) = \langle \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y-x\|^2}{2\sigma_t{}^2}} \rangle$. For simplicity, we discuss the derivatives of the logarithm of the $y$-function $h(y) := \langle e^{a\|y+x\|^2} \rangle$, where $a \in \mathbb{R}$ is a constant. We will choose $a = -\frac{1}{2\sigma_t^2}$ finally, and the difference between $\|y-x\|^2$ and $\|y+x\|^2 = \|-y-x\|^2$ doesn't influence the $\|\cdot\|_\infty$-norms of derivatives.

We first compute the first and second derivatives of $\log h(y)$. We notice that

$$\partial_i h(y) = 2a \langle (y_i + x_i) e^{a\|y+x\|^2} \rangle = 2a y_i h(y) + 2a \langle x_i e^{a\|y+x\|^2} \rangle. \tag{22}$$

So,

(23)
$$\partial^2_{ij} h(y) = 2a\delta_{ij} h(y) + 2a y_j (2a y_i h(y) + 2a \langle x_i e^{a\|y+x\|^2} \rangle) + 4a^2 \langle (x_i + y_i) x_j e^{a\|y+x\|^2} \rangle$$

$$= 2a\delta_{ij} h(y) + 4a^2 y_i y_j h(y) + 4a^2 y_i \langle (x_j e^{a\|y+x\|^2} \rangle + 4a^2 y_j \langle x_i e^{a\|y+x\|^2} \rangle + 4a^2 \langle x_i x_j e^{a\|y+x\|^2} \rangle. \tag{24}$$

Hence,

$$
\begin{aligned}
\partial^2_{ij} \log h(y) &= \frac{\partial^2_{ij} h(y) h(y) - \partial_i h(y) \partial_j h(y)}{h(y)^2} \\
&= 2a\delta_{ij} + 4a^2 \frac{\langle x_i x_j e^{a\|y+x\|^2}\rangle h(y) - \langle x_i e^{a\|y+x\|^2}\rangle \langle x_j e^{a\|y+x\|^2}\rangle}{h(y)^2}.
\end{aligned}
\tag{25}
$$

Before we compute higher order derivatives of $\log h(y)$, we first illustrate how we obtain an upper bound for $\partial^2_{ij} \log h(y)$. Notice that in the definition of $\langle f(x)\rangle$ in (21), because $\mu_*$ has a compact support $K_*$, the $x$ in the integrand satisfies that $x \in \lambda_t K_*$. Hence, in the integrand of (21), $\|x\|_\infty \le \lambda_t D$ by Assumption 3.1. So, $|\langle x_i e^{a\|y+x\|^2}\rangle| \le \lambda_t D h(y)$, $|\langle x_i x_j e^{a\|y+x\|^2}\rangle| \le \lambda_t^2 D^2 h(y)$. So, $|\partial^2_{ij} \log h(y)| \le 2a\delta_{ij} + 8a^2 \lambda_t^2 D^2$.

Then, we compute $\nabla^p \log q_t(y)$ for $p \ge 3$. We use induction to show that in the expression of $\nabla^p \log h(y)$, there is no polynomial term of $y$ like we have seen in (25). Assume that for an $m \in \mathbb{Z}_+$, and for any multi-index $\alpha$ with $|\alpha| = m$, the derivative $\partial_y^\alpha \log h(y)$ has a form

$$\partial_y^\alpha \log h(y) = (2a)^m \frac{P_\alpha(y)}{h(y)^m},$$

where $P_\alpha(y)$ is the summation of at most $4(m-1)! - 2$ terms, where each term has a form

$$\pm \langle x^{\beta_1} e^{a\|y+x\|^2}\rangle \cdot \langle x^{\beta_2} e^{a\|y+x\|^2}\rangle \cdots \langle x^{\beta_m} e^{a\|y+x\|^2}\rangle,$$

and each $\beta_i$ is a multi-index and $|\beta_1| + |\beta_2| + \ldots + |\beta_m| = m$. This assumption is satisfied when $m = 2$, because the derivative of the term $2a\delta_{ij}$ in (25) is zero so that we can omit this term. Then,

$$\partial_{y_1}\partial_y^\alpha \log h(y) = (2a)^m \partial_{y_1} \frac{P_\alpha(y)}{h(y)^m} = (2a)^{m+1} \cdot \frac{\frac{1}{2a}\partial_{y_1} P_\alpha(y) h(y) - m P_\alpha(y) \langle (y_1 + x_1) e^{a\|y+x\|^2}\rangle}{h(y)^{m+1}}. \tag{26}$$

We notice that

$$
\begin{aligned}
&\frac{1}{2a}\partial_{y_1}\left(\langle x^{\beta_1} e^{a\|y+x\|^2}\rangle \cdot \langle x^{\beta_2} e^{a\|y+x\|^2}\rangle \cdots \langle x^{\beta_m} e^{a\|y+x\|^2}\rangle\right) \\
&= m y_1 \langle x^{\beta_1} e^{a\|y+x\|^2}\rangle \cdot \langle x^{\beta_2} e^{a\|y+x\|^2}\rangle \cdots \langle x^{\beta_m} e^{a\|y+x\|^2}\rangle \\
&\quad + \langle x_1 x^{\beta_1} e^{a\|y+x\|^2}\rangle \cdot \langle x^{\beta_2} e^{a\|y+x\|^2}\rangle \cdots \langle x^{\beta_m} e^{a\|y+x\|^2}\rangle \\
&\quad + \langle x^{\beta_1} e^{a\|y+x\|^2}\rangle \cdot \langle x_1 x^{\beta_2} e^{a\|y+x\|^2}\rangle \cdots \langle x^{\beta_m} e^{a\|y+x\|^2}\rangle \\
&\quad + \ldots + \langle x^{\beta_1} e^{a\|y+x\|^2}\rangle \cdot \langle x^{\beta_2} e^{a\|y+x\|^2}\rangle \cdots \langle x_1 x^{\beta_m} e^{a\|y+x\|^2}\rangle.
\end{aligned}
$$

Hence, $\frac{1}{2a}\partial_{y_1} P_\alpha(y)$ is of the form $m y_1 P_\alpha(y) + Q_\alpha(y)$, where $Q_\alpha(y)$ is the summation of at most $4(m-1)! - 2$ terms without $y_1$ showing in the polynomial terms. We notice that those $y_1$ terms in the numerator of (26) will then cancel, and the remaining terms all look like

$$\pm \langle x^{\beta'_1} e^{a\|y+x\|^2}\rangle \cdot \langle x^{\beta'_2} e^{a\|y+x\|^2}\rangle \cdots \langle x^{\beta'_{m+1}} e^{a\|y+x\|^2}\rangle,$$

where each $\beta'_i$ is a multi-index and $|\beta'_1| + |\beta'_2| + \ldots + |\beta'_{m+1}| = m+1$. The number of them is at most $m(4(m-1)! - 2) + m = 4m! - m \le 4m! - 2$ when $m \ge 2$. We then conclude the proof of Lemma B.3. □

The following lemma describes the situation when $t > 0$ is very large.

**Lemma B.4.** *Let $q_t(y)$ be defined in (19), and let*

$$g(y) := \frac{1}{(\sqrt{2\pi})^d} \cdot e^{-\frac{\|y\|^2}{2}}.$$

*Then, we have that for $t > 0$,*

$$\int_{\mathbb{R}^d} |q_t(y) - g(y)| \mathrm{d}y \le \frac{2d(1-\sigma_t)}{{\sigma_t}^{d+3}} + \lambda_t \sqrt{d} D \frac{4 + \lambda_t \sqrt{d} D}{\sigma_t^2} \cdot e^{\frac{\lambda_t^2 d D^2}{2{\sigma_t}^2}}. \tag{27}$$

*In particular, there exists a universal constant $C_u > 0$, such that*

$$\int_{\mathbb{R}^d} |q_t(y) - g(y)| \mathrm{d}y \le C_u e^{-t} \sqrt{d} D, \tag{28}$$

*which goes to $0$ exponentially as $t \to +\infty$.*

*Proof.* Because one can write

$$q_t(y) = \int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y - \lambda_t x\|^2}{2{\sigma_t}^2}} \, d\mu_*(x),$$

and $\int_{\mathbb{R}^d} d\mu_*(x) = 1$, we see that

$$\begin{aligned} \int_{\mathbb{R}^d} |q_t(y) - g(y)| \mathrm{d}y &= \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y - \lambda_t x\|^2}{2{\sigma_t}^2}} - \frac{1}{(\sqrt{2\pi})^d} \cdot e^{-\frac{\|y\|^2}{2}} \, d\mu_*(x) \right| \mathrm{d}y \\ &\le \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left| \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y - \lambda_t x\|^2}{2{\sigma_t}^2}} - \frac{1}{(\sqrt{2\pi})^d} \cdot e^{-\frac{\|y\|^2}{2}} \right| \mathrm{d}y \, d\mu_*(x). \end{aligned} \tag{29}$$

We denote $y = (y_1, y') \in \mathbb{R}^d$, and we define the function

$$f(y, r) := \left| \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{|y_1 - \lambda_t r|^2 + \|y'\|^2}{2{\sigma_t}^2}} - \frac{1}{(\sqrt{2\pi})^d} \cdot e^{-\frac{\|y\|^2}{2}} \right|.$$

By the rotation symmetry, we see that the right hand side of (29) equals to

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(y, \|x\|) \mathrm{d}y \, d\mu_*(x). \tag{30}$$

Let's first see what is $|\frac{\mathrm{d}}{\mathrm{d}r} f(y, r)|$ for $r \le \sqrt{d} D$. A direct computation shows that

$$\begin{aligned} \left| \frac{\mathrm{d}}{\mathrm{d}r} f(y, r) \right| &= \frac{\lambda_t}{(\sqrt{2\pi}\sigma_t)^d} \cdot \frac{|y_1 - \lambda_t r|}{\sigma_t^2} \cdot e^{-\frac{|y_1 - \lambda_t r|^2 + \|y'\|^2}{2{\sigma_t}^2}} \\ &\le \frac{\lambda_t}{(\sqrt{2\pi}\sigma_t)^d} \cdot \frac{|y_1| + \lambda_t \sqrt{d} D}{\sigma_t^2} \cdot e^{-\frac{\frac{1}{2}|y_1|^2 - \lambda_t^2 d D^2 + \|y'\|^2}{2{\sigma_t}^2}}. \end{aligned}$$

So, (30) is bounded by

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(y, 0) \mathrm{d}y \, d\mu_*(x) + \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{\lambda_t \|x\|}{(\sqrt{2\pi}\sigma_t)^d} \cdot \frac{|y_1| + \lambda_t \sqrt{d} D}{\sigma_t^2} \cdot e^{-\frac{\frac{1}{2}|y_1|^2 - \lambda_t^2 d D^2 + \|y'\|^2}{2{\sigma_t}^2}} \mathrm{d}y \, d\mu_*(x). \tag{31}$$

The second term in (31) is bounded by

$$\int_{\mathbb{R}^d} \frac{\lambda_t \sqrt{d} D}{(\sqrt{2\pi}\sigma_t)^d} \cdot \frac{|y_1| + \lambda_t \sqrt{d} D}{\sigma_t^2} \cdot e^{-\frac{\frac{1}{2}|y_1|^2 - \lambda_t^2 d D^2 + \|y'\|^2}{2{\sigma_t}^2}} \mathrm{d}y \le \lambda_t \sqrt{d} D \frac{4 + \lambda_t \sqrt{d} D}{\sigma_t^2} \cdot e^{\frac{\lambda_t^2 d D^2}{2{\sigma_t}^2}}, \tag{32}$$

which goes to 0 exponentially as $t \to +\infty$ because $\lambda_t = e^{-t}$ and $\sigma_t^2 = 1 - \lambda_t^2$.

The first term in (31) is

$$\int_{\mathbb{R}^d} \left| \frac{1}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y\|^2}{2{\sigma_t}^2}} - \frac{1}{(\sqrt{2\pi})^d} \cdot e^{-\frac{\|y\|^2}{2}} \right| \mathrm{d}y, \tag{33}$$

because $\int_{\mathbb{R}^d} d\mu_*(x) = 1$. We define the function

$$h(y,\alpha) = \left| \frac{1}{(\sqrt{2\pi}\alpha)^d} \cdot e^{-\frac{\|y\|^2}{2\alpha^2}} - \frac{1}{(\sqrt{2\pi})^d} \cdot e^{-\frac{\|y\|^2}{2}} \right|,$$

and then for $\alpha \in [0,1]$,

$$\left| \frac{\mathrm{d}}{\mathrm{d}\alpha} h(y,\alpha) \right| = \frac{1}{(\sqrt{2\pi})^d} \left| -d \frac{1}{\alpha^{d+1}} e^{-\frac{\|y\|^2}{2\alpha^2}} + \frac{1}{\alpha^d} e^{-\frac{\|y\|^2}{2\alpha^2}} \cdot \frac{\|y\|^2}{\alpha^3} \right| \le \frac{e^{-\frac{\|y\|^2}{2\alpha^2}}}{(\sqrt{2\pi})^d \alpha^{d+3}} \cdot |d + \|y\|^2|.$$

Because $h(y,1) = 0$, we see that (33) is bounded by

$$\frac{(1-\sigma_t)}{{\sigma_t}^{d+3}} \int_{\mathbb{R}^d} \frac{e^{-\frac{\|y\|^2}{2}}}{(\sqrt{2\pi})^d} \cdot |d + \|y\|^2| \mathrm{d}y = \frac{2d(1-\sigma_t)}{{\sigma_t}^{d+3}}, \tag{34}$$

which goes to 0 exponentially as $t \to +\infty$ because $\sigma_t = \sqrt{1-e^{-2t}}$. The estimates (32) and (34) together give (27). Next we show (28). There are two cases: if $e^{-t}\sqrt{d}D \ge 1$, then (28) holds trivially, since the lefthand side is at most 2. Otherwise $e^{-t}\sqrt{d}D \le 1$, we have

$$\begin{aligned}
&\frac{2d(1-\sigma_t)}{{\sigma_t}^{d+3}} + \lambda_t \sqrt{d} D \frac{4 + \lambda_t \sqrt{d} D}{\sigma_t^2} \cdot e^{\frac{\lambda_t^2 d D^2}{2{\sigma_t}^2}} \le \frac{2de^{-2t}}{\sigma_t^{2d}} + \frac{4e^{-t}\sqrt{d}D + (e^{-t}\sqrt{d}D)^2}{\sigma_t^2} \cdot e^{\frac{1}{2{\sigma_t}^2}} \\
&\le \frac{2de^{-2t}}{\sigma_t^{2d}} + \frac{5e^{-t}\sqrt{d}D}{\sigma_t^2} e^{\frac{2}{3}} = \frac{2de^{-2t}}{(1-e^{-2t})^d} + \frac{5e^{-t}\sqrt{d}D}{1-e^{-2t}} e^{\frac{2}{3}} \\
&\le \frac{2\sqrt{d}e^{-t}}{1-de^{-2t}} \cdot \sqrt{d}e^{-t} + \frac{5e^{-t}\sqrt{d}D}{1-e^{-2t}} e^{\frac{2}{3}} \le C_u e^{-t} \sqrt{d} D,
\end{aligned}$$

for a universal constant $C_u > 0$. In the last inequality, we assume that $e^{-t} \le \frac{1}{4\sqrt{d}} \le \frac{1}{4}$. Otherwise, if $e^{-t} \ge \frac{1}{4\sqrt{d}}$, (28) also holds trivially. This finishes the proof of (28). $\square$

When we prove our main theorem, Theorem 3.4, in the following Section C, we will assume that $\mu_*$ has a compact support $K_*$ and use Lemma B.3 to proceed the proof. On the other hand, as we have mentioned in Remark 3.5, our methods work under other assumptions on the initial data $\mu_*$, as long as under those assumptions, we can reasonably obtain the properties in Remark C.4. We next assume that $\mu_*$ is a Gaussian mixture and obtain an estimate similar to Lemma B.3.

**Lemma B.5.** *Assume that $\mu_*$ is a Gaussian mixture, i.e., we assume that*

$$\mu_*(x) \coloneqq \sum_{k=1}^M c_k \frac{1}{(\sqrt{2\pi})^d a_k} \cdot \exp\left( -\frac{1}{2}(x-b_k) \cdot A_k^{-1}(x-b_k) \right),$$

*where $A_k$'s are positive definite matrices, $a_k \coloneqq \sqrt{\det A_k}$, $b_k$'s are constant vectors in $\mathbb{R}^d$, $c_k > 0$ and $\sum_{k=1}^M c_k = 1$. Then, for any $\ell \in \mathbb{Z}_+$ and any multi-index $\alpha$ with $|\alpha| \le \ell$, there is a constant $C(\ell,\mu_*)$, depending on $\ell$ and $A_k, b_k$ in the formula of $\mu_*$, such that for any $x \in \mathbb{R}^d$ and any $t \ge 0$,*

$$\frac{|\partial_x^\alpha q_t(x)|}{q_t(x)} \le C(\ell,\mu_*) \cdot \|x\|^\ell + C(\ell,\mu_*). \tag{35}$$

*Proof.* A standard computation, by (19), shows that the density function of $q_t(x)$ is

$$q_t(x) \coloneqq \sum_{k=1}^M c_k \frac{1}{(\sqrt{2\pi})^d \cdot a_k(t)} \cdot \exp\left( -\frac{1}{2}(x-b_k(t)) \cdot A_k(t)^{-1}(x-b_k(t)) \right),$$

where $A_k(t) = \lambda_t^2 A_k + \sigma_t^2 \mathbb{I}_d$, $b_k(t) = \lambda_t b_k$. Hence,

$$\nabla q_t(x) = \sum_{k=1}^{M} c_k \frac{-A_k(t)^{-1}(x - b_k(t))}{(\sqrt{2\pi})^d \cdot a_k(t)} \cdot \exp\Big( -\frac{1}{2}(x - b_k(t)) \cdot A_k(t)^{-1}(x - b_k(t)) \Big).$$

So,

$$|\partial_1 q_t(x)| \le \max_{1 \le k \le M} \| - A_k(t)^{-1}(x - b_k(t))\| \cdot q_t(x).$$

Notice that because when $t = 0$, $A_k$'s are positive definite, and when $t \to +\infty$, $A_k(t) \to \mathbb{I}_d$, there is an upper bound $C(\ell, \mu_*)$, such that $\sup_{t\in[0,+\infty)} \|A_k(t)^{-1}\| \le C(\ell, \mu_*)$. Hence, we can obtain (35) for $\ell = 1$. For general $\ell \ge 1$, one can either use induction or compute them directly, exactly using the same way as the case $\ell = 1$ we have shown. For the purpose of proving Theorem 3.4, the estimates for $\ell \le 3$ will be enough.

□

Next, we also point out that those $\widehat{q}_t$'s also have exponential tails, as long as one of them has an exponential tail at a given time. For example, if $\widehat{q}_T = q_T$.

**Lemma B.6.** *If for a $t' > 0$, $\widehat{q}_{t'}(x)$ has an exponential tail as $\|x\| \to +\infty$, then for any $t > 0$, $\widehat{q}_t(x)$ also has an exponential tail as $\|x\| \to +\infty$.*

*Proof.* Notice that $\partial_t \widehat{q}_t = \nabla \cdot (\widehat{U}_t \widehat{q}_t)$. Hence, along the characteristic line $\frac{\mathrm{d}}{\mathrm{d}t} T_t(x) = -\widehat{U}_{t'+t}(T_t(x))$, one can solve that

$$\widehat{q}_{t'+t}(T_t(x)) = \widehat{q}_{t'}(x) \cdot \exp\Big( \int_0^t (\nabla \cdot \widehat{U}_{t'+s})(T_s(x))\ \mathrm{d}s \Big).$$

Because $\widehat{U}_t(x) = x + s_{T-t}(x)$, by Assumption 3.3, we have that $|(\nabla \cdot \widehat{U}_t)(x)| \le d(1 + L_{T-t})$ for any $t > 0$. The remaining estimate is on the norm of $T_t(x)$. We notice that, by Assumption 3.3 again,

$$\|s_{T-t}(x)\| \le \|s_{T-t}(0)\| + (\|x\| \cdot d \cdot L_{T-t}) \le d \cdot L_{T-t}(1 + \|x\|).$$

Hence,

$$\begin{aligned} |\partial_t \|T_t(x)\|^2| = |\widehat{U}_{t'+t}(T_t(x)) \cdot T_t(x)| &\le (1 + dL_{T-t'-t})\|T_t(x)\|^2 + dL_{T-t'-t}\|T_t(x)\| \\ &\le (2 + dL_{T-t'-t})\|T_t(x)\|^2 + d^2 L_{T-t'-t}^2. \end{aligned}$$

Denote $\mathcal{L}_t = \int_0^t (2 + d \cdot L_{T-t'-s})\ \mathrm{d}s$. Hence, for $t > 0$,

$$e^{-\mathcal{L}_t} \cdot \Big( \|x\|^2 - \int_0^t e^{\mathcal{L}_s} \cdot d^2 \cdot L_{T-t'-s}^2\ \mathrm{d}s \Big) \le \|T_t(x)\|^2 \le e^{\mathcal{L}_t} \cdot \Big( \|x\|^2 + \int_0^t e^{-\mathcal{L}_s} \cdot d^2 \cdot L_{T-t'-s}^2\ \mathrm{d}s \Big).$$

Because we can compare the norms of $T_t(x)$ and $x$ by a factor only depending on $t$, and $T_t$ is also a diffeomorphism since $\widehat{U}_t(x)$ is locally Lipschitz, then we know that $\widehat{q}_t(x)$ also has an exponential tail as $\|x\| \to +\infty$. One can obtain a similar result for $t < 0$. □

## Appendix C. Score Estimation Error

In this section, we assume that $\mu_*$ has a compact support $K_*$ as in Assumption 3.1 to proceed the proof first. At the end of this section, we will point out some possible ways to use other assumptions on this initial data $\mu_*$ in Remark C.4. We first need the following Gagliardo-Nirenberg interpolation inequality to estimate $E(t)$ defined in Theorem A.1, and finally use $\int_{\mathbb{R}^d} q_t(x)\delta_t^2(x)\ \mathrm{d}x$ to control $E(t)$ when $t \in [\tau, T]$. For the convenience of readers, let us also sketch the proof of this inequality here.

**Lemma C.1** (Gagliardo-Nirenberg)**.** *There is a positive universal constant $C_u$, such that for any $d \in \mathbb{Z}_+$, any $w \in L^1$, any $i \in [\![1, d]\!]$, if $\partial_{ii}^2 w \in L^1$, then*

$$\left(\int_{\mathbb{R}^d} |\partial_i w(x)|\ \mathrm{d}x\right)^2 \le C_u \left(\int_{\mathbb{R}^d} |\partial_{ii}^2 w(x)|\ \mathrm{d}x\right)\left(\int_{\mathbb{R}^d} |w(x)|\ \mathrm{d}x\right). \tag{36}$$

*In general, if $\partial_i^k w \in L^1$ with $k \ge 2$, then*

$$\int_{\mathbb{R}^d} |\partial_i w(x)|\ \mathrm{d}x \le C_u^{\frac{k-1}{2}} \left(\int_{\mathbb{R}^d} |\partial_i^k w(x)|\ \mathrm{d}x\right)^{\frac{1}{k}}\left(\int_{\mathbb{R}^d} |w(x)|\ \mathrm{d}x\right)^{\frac{k-1}{k}}. \tag{37}$$

*Proof.* Without loss of generality, we assume that $i = 1$. For any $u \in C_c^2(\mathbb{R}^d)$, we fix its remaining coordinates $x' = (x_2, \dots, x_d)$, then according to Lemma 3.4 of [15], there is a universal constant $C_u > 0$, such that

$$\left(\int_{\mathbb{R}} |\partial_1 w(x_1, x')|\ \mathrm{d}x_1\right)^2 \le C_u \left(\int_{\mathbb{R}} |\partial_{11}^2 w(x_1, x')|\ \mathrm{d}x_1\right)\left(\int_{\mathbb{R}} |w(x_1, x')|\ \mathrm{d}x_1\right).$$

Hence,

$$\begin{aligned}
\left(\int_{\mathbb{R}^d} |\partial_i w(x)|\ \mathrm{d}x\right) &= \left(\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} |\partial_1 w(x_1, x')|\ \mathrm{d}x_1 \mathrm{d}x'\right) \\
&\le C_u^{\frac{1}{2}} \int_{\mathbb{R}^{d-1}} \left(\int_{\mathbb{R}} |\partial_{11}^2 w(x_1, x')|\ \mathrm{d}x_1\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}} |w(x_1, x')|\ \mathrm{d}x_1\right)^{\frac{1}{2}} \mathrm{d}x' \\
&\le C_u^{\frac{1}{2}} \left(\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} |\partial_{11}^2 w(x_1, x')|\ \mathrm{d}x_1\ \mathrm{d}x'\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} |w(x_1, x')|\ \mathrm{d}x_1\ \mathrm{d}x'\right)^{\frac{1}{2}}.
\end{aligned}$$

In general, assume that we already know that for some $k \ge 2$,

$$\int_{\mathbb{R}^d} |\partial_i w(x)|\ \mathrm{d}x \le C_u^{\frac{k-1}{2}} \left(\int_{\mathbb{R}^d} |\partial_i^k w(x)|\ \mathrm{d}x\right)^{\frac{1}{k}}\left(\int_{\mathbb{R}^d} |w(x)|\ \mathrm{d}x\right)^{\frac{k-1}{k}},$$

then we replace $w(x)$ with $\partial_i w(x)$, and obtain that

$$\int_{\mathbb{R}^d} |\partial_{ii}^2 w(x)|\ \mathrm{d}x \le C_u^{\frac{k-1}{2}} \left(\int_{\mathbb{R}^d} |\partial_i^{k+1} w(x)|\ \mathrm{d}x\right)^{\frac{1}{k}}\left(\int_{\mathbb{R}^d} |\partial_i w(x)|\ \mathrm{d}x\right)^{\frac{k-1}{k}}.$$

Combine this inequality and the inequality (36), we can obtain (37) for $k + 1$. □

Now, let's estimate the integral of $E(t)$ in $t$. By (36) in Lemma C.1 and Hölder inequality,

(38)

$$\begin{aligned}
\int_\tau^T \left(\int_{\mathbb{R}^d} |\partial_1 (q_t \delta_t^1)(x)|\ \mathrm{d}x\right) \mathrm{d}t &\le C_u^{\frac{1}{2}} \int_\tau^T \left(\int_{\mathbb{R}^d} |\partial_{11}^2 (q_t \delta_t^1)(x)|\ \mathrm{d}x\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^d} |(q_t \delta_t^1)(x)|\ \mathrm{d}x\right)^{\frac{1}{2}} \mathrm{d}t \\
&\le C_u^{\frac{1}{2}} \left(\int_\tau^T \int_{\mathbb{R}^d} |\partial_{11}^2 (q_t \delta_t^1)(x)|\ \mathrm{d}x\ \mathrm{d}t\right)^{\frac{1}{2}}\left(\int_\tau^T \int_{\mathbb{R}^d} |(q_t \delta_t^1)(x)|\ \mathrm{d}x\ \mathrm{d}t\right)^{\frac{1}{2}},
\end{aligned}$$

where we use the notation $\delta_t = (\delta_t^1, \dots, \delta_t^d)$. For the term

$$\int_\tau^T \int_{\mathbb{R}^d} |(q_t \delta_t^1)(x)|\ \mathrm{d}x\ \mathrm{d}t,$$

we use Hölder inequality twice and the fact that $\int_{\mathbb{R}^d} q_t(x)\,\mathrm{d}x = 1$, and see that

$$\int_\tau^T \int_{\mathbb{R}^d} |(q_t\delta_t^1)(x)|\,\mathrm{d}x\,\mathrm{d}t \le \int_\tau^T \left(\int_{\mathbb{R}^d} q_t(x)\big(\delta_t^1(x)\big)^2\,\mathrm{d}x\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} q_t(x)\,\mathrm{d}x\right)^{\frac{1}{2}}\mathrm{d}t$$
$$\le \left(\int_\tau^T \int_{\mathbb{R}^d} q_t(x)\big(\delta_t^1(x)\big)^2\,\mathrm{d}x\,\mathrm{d}t\right)^{\frac{1}{2}} \cdot (T-\tau)^{\frac{1}{2}}.$$

Notice that our assumption is that $\int_\tau^T \int_{\mathbb{R}^d} q_t(x)\big(\delta_t^1(x)\big)^2\,\mathrm{d}x\mathrm{d}t$ can be made very small. Next, we are going to show that the term

$$\int_\tau^T \int_{\mathbb{R}^d} |\partial_{11}^2(q_t\delta_t^1)(x)|\,\mathrm{d}x\,\mathrm{d}t \tag{39}$$

is bounded by a positive constant depending on $\tau, T, \mathcal{L}, D$. For the terms in $\partial_{11}^2(q_t\delta_t^1)(x) = (\partial_{11}^2 q_t)\delta_t^1 + 2\partial_1 q_t\partial_1\delta_t^1 + q_t(\partial_{11}^2\delta_t^1)$, we notice that, because $\delta_t(x) = s_{T-t}(x) - \nabla\log q_t(x)$, by the proof of Lemma B.3

$$|\partial_1\delta_t^1(x)| \le L_{T-t} + \frac{1}{{\sigma_t}^4}(\sigma_t^2 + 2\lambda_t^2 D^2),$$

$$|\partial_{11}^2\delta_t^1(x)| \le L_{T-t} + \frac{24{\lambda_t}^3}{{\sigma_t}^6}D^3,$$

and we see that for any $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$, by the proof of Lemma B.3 again,

$$\left|\frac{\partial_1 q_t(x)}{q_t(x)}\right| \le \frac{(|x_1| + \lambda_t D)}{\sigma_t^2}, \tag{40}$$

$$\left|\frac{\partial_{11}^2 q_t(x)}{q_t(x)}\right| \le \frac{2(|x_1|^2 + \lambda_t^2 D^2) + \sigma_t^2}{\sigma_t^4}.$$

Hence,

$$\left(\int_{\mathbb{R}^d} |(\partial_{11}^2 q_t)(x)\delta_t^1(x)|\,\mathrm{d}x\right)^2 = \left(\int_{\mathbb{R}^d} \left|\frac{(\partial_{11}^2 q_t)(x)}{q_t(x)}\delta_t^1(x)q_t(x)\right|\mathrm{d}x\right)^2$$
$$\le \left(\int_{\mathbb{R}^d} \left(\frac{2(|x_1|^2 + \lambda_t^2 D^2) + \sigma_t^2}{\sigma_t^4}\right)^2 q_t(x)\,\mathrm{d}x\right) \cdot \left(\int_{\mathbb{R}^d} \big(\delta_t^1(x)\big)^2 q_t(x)\,\mathrm{d}x\right),$$

where the first term is a bounded term by similarly analyzing $q_t(x)$. For example, let us show that the $x_1$-fourth moment of $q_t$, i.e., $\int_{\mathbb{R}^d} |x_1|^4 q_t(x)\,\mathrm{d}x$, is bounded. By (19), we know that

$$\begin{aligned}\int_{\mathbb{R}^d} |y_1|^4 q_t(y)\,\mathrm{d}y &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{|y_1|^4}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}} \cdot \mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x\mathrm{d}y \\ &\le 8\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{|y_1 - x_1|^4 + |x_1|^4}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}} \cdot \mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x\mathrm{d}y \\ &\le 8\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{|y_1 - x_1|^4 + \lambda_t^4 D^4}{(\sqrt{2\pi}\sigma_t)^d} \cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}} \cdot \mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x\mathrm{d}y \\ &= C_u(\sigma_t^4 + \lambda_t^4 D^4) \le C_u(1 + D^4),\end{aligned} \tag{41}$$

for a universal constant $C_u > 0$ coming from the fourth moment of the standard Gaussian. We can similarly estimate the remaining two terms in the expansion of $\partial_{11}^2(q_t\delta_t^1)(x)$ and get an upper bound for (39). Also, we remark that after taking the time integral from $\tau$ to $T$, the main order of $\sigma_t$ involved in (39) is at most

$$\int_\tau^T \frac{1}{\sigma_t^6}\,\mathrm{d}t = \int_\tau^T \frac{e^{6t}}{(e^{2t}-1)^3}\,\mathrm{d}t,$$

which blows up of order $T$ as $T \to +\infty$ and blows up of order $\tau^{-2}$ as $\tau \to 0^+$. Hence, there is a universal constant $C_u > 0$, such that

$$\int_\tau^T \int_{\mathbb{R}^d} |\partial_{11}^2(q_t\delta_t^1)(x)|\,\mathrm{d}x\,\mathrm{d}t \le C_u \int_\tau^T \left(L_{T-t} + \frac{D^3}{\sigma_t^6}\right)\mathrm{d}t \le C_u(\mathcal{L} + T\cdot\tau^{-2}\cdot D^3).$$

Then, by (38) and Assumption 3.2,

$$\int_\tau^T \left(\int_{\mathbb{R}^d} |\partial_1(q_t\delta_t^1)(x)|\,\mathrm{d}x\right)\mathrm{d}t \le C_u \cdot \delta^{\frac{1}{2}} \cdot T^{\frac{1}{4}} \cdot (\mathcal{L} + T\cdot\tau^{-2}\cdot D^3)^{\frac{1}{2}}.$$

**Remark C.2.** *We notice that one can also modify the inequality (38) by*

$$\int_\tau^T \left(\int_{\mathbb{R}^d} |\partial_1(q_t\delta_t^1)(x)|\,\mathrm{d}x\right)\mathrm{d}t \le C_u^{\frac{1}{2}}\left(\int_\tau^T \varphi(t)\int_{\mathbb{R}^d} |\partial_{11}^2(q_t\delta_t^1)(x)|\,\mathrm{d}x\mathrm{d}t\right)^{\frac{1}{2}}$$
$$\times\left(\int_\tau^T \frac{1}{\varphi(t)}\int_{\mathbb{R}^d} |(q_t\delta_t^1)(x)|\,\mathrm{d}x\mathrm{d}t\right)^{\frac{1}{2}},$$

*with a suitably chosen positive function $\varphi(t)$ when we use the Hölder inequality. For example, we can let $\varphi(t) \to 0$ of order $t^2$ as $t \to 0^+$, and let $\varphi(t) \to 0$ of order $\frac{1}{t}$ as $t \to +\infty$. In this way, the first term in the above inequality is uniformly bounded so that we can pass $\tau \to 0^+$ and $T \to +\infty$. So, we only need to control the second term so that it is small enough.*

**Remark C.3.** *In our settings, if we know that $s_t(x)$ has higher order derivatives up to $k$ for $k \ge 2$, we can replace (36) with (37) when we estimate $E(t)$ in (38). Then, in order to estimate the derivatives of $q_t$ in the expansion of $\partial_1^k(q_t\delta_t)$, a similar proof of Lemma B.3 should work because $q_t(x)$ has an exponential tail as $\|x\| \to +\infty$.*

**Remark C.4.** *As readers have seen these proofs under the assumption that $\mu_*$ has a compact support $K_*$, the main reason we need this compact support assumption is to estimate the term*

$$\int_\tau^T \int_{\mathbb{R}^d} |\partial_{11}^2(q_t\delta_t^1)(x)|\,\mathrm{d}x\,\mathrm{d}t.$$

*Notice that $q_t\delta_t = q_t(s_t - \nabla \log q_t) = q_t s_t - \nabla q_t$. If we replace the assumption with $\mu_*$ being a Gaussian mixture, according to Lemma B.5, we can do a similar estimate on $|\partial_{11}^2(q_t\delta_t^1)(x)|$, and hence we can similarly obtain an upper bound for $\int_\tau^T \int_{\mathbb{R}^d} |\partial_{11}^2(q_t\delta_t^1)(x)|\,\mathrm{d}x\,\mathrm{d}t$. Such an upper bound will then depend on those parameters in the initial Gaussian mixture $\mu_*$, but don't blow up as $\tau \to 0^+$. See Lemma B.5. One can make other reasonable assumptions on $\mu_*$ as long as one can reasonably estimate this second derivative integral.*

## Appendix D. Discretization Error

As discussed in Section 2.2, we solve the ODE flow $\partial_t\widehat{Y}_t = \widehat{Y}_t + s_t(\widehat{Y}_t) = \widehat{V}_t(\widehat{Y}_t)$ using the Runge-Kutta method. Although the Runge-Kutta updating rule as in (7) is a discrete time process, we can interpolate it as a continuous time process on $t_i \le t \le t_{i+1}$ as

$$\widetilde{Y}_t = F_r(\widetilde{Y}_{t_i}), \quad t = t_i + r,$$

where

$$F_r(x) = x + r\sum_{j=1}^{s} b_j k_j, \quad 0 \le r \le t_{i+1} - t_i,$$

and

$$\begin{aligned} k_1 &= V_{t_i + rc_1}(x), \\ k_2 &= V_{t_i + rc_2}\big(x + r(a_{21}k_1)\big), \\ k_3 &= V_{t_i + rc_3}\big(x + r(a_{31}k_1 + a_{32}k_2)\big), \\ &\ \ \vdots \\ k_s &= V_{t_i + rc_s}\big(x + r(a_{s1}k_1 + a_{s2}k_2 + \cdots + a_{s,s-1}k_{s-1})\big). \end{aligned}$$

We denote the density of $\widetilde{Y}_t$ as $\widetilde{\varrho}_t$, then at times $t_i$, $\widetilde{\varrho}_{t_i}$ is the density of $Y_i$ as given by the $i$-th step of the Runge-Kutta methods. If $V_t(\cdot)$ is differentiable in $t$ on $[t_i, t_{i+1}]$, one can see from the above contruction $\widetilde{Y}_t$ is differentiable in $t$, and

$$\partial_t \widetilde{Y}_t = \partial_r F_r(\widetilde{Y}_{t_i}). \tag{42}$$

The following proposition states that for $t_{i+1} - t_i = h$ small enough, we can rewrite (42) as an ODE flow

$$\partial_t \widetilde{Y}_t = \widetilde{V}_t(\widetilde{Y}_t), \tag{43}$$

for $t_i \le t \le t_{i+1}$, and $\widetilde{V}_t(\cdot)$ is close to $V_t(\cdot)$ up to an error of size $\mathcal{O}(h^p)$.

**Proposition D.1.** *Adopt Assumption 3.7. There exists a large constant $C(p,s)$ (depending on the stage and order of the Runge-Kutta methods), if $C(p,s)hdL \le 1$, then the following holds. For any $0 \le r \le h$, $F_r$ is a diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. We denote its functional inverse as $\Phi_r(x)$, then*

$$\widetilde{V}_{t_i + r}(x) = \partial_r F_r(\Phi_r(x)). \tag{44}$$

*Moreover, for $t_i \le t \le t_{i+1}$, $\partial_t \widetilde{Y}_t = \widetilde{V}_t(\widetilde{Y}_t)$, and*

$$\|\widetilde{V}_t(x) - V_t(x)\|_\infty, \quad \|\nabla(\widetilde{V}_t(x) - V_t(x))\|_\infty \le C(p,s)\cdot L((\sqrt{d} + \|x\|)\sqrt{d}hL)^p, \tag{45}$$

*Proof of Theorem 3.9.* Then we need to analyze the density evolution under (43). We let $\widetilde{q}_t$ piece-wisely solve the transport equation

$$\partial_t \widetilde{q}_t = \nabla\cdot(\widetilde{U}_t \widetilde{q}_t) \quad \text{with} \quad \widetilde{U}_t = \widetilde{V}_{T-t},$$

on each interval $[t_i, t_{i+1}]$ for $i \ge 1$, where $0 = t_0 < t_1 < \cdots < t_N = T - \tau$. Then, we define $\widetilde{\delta}_t(x) \coloneqq \widetilde{U}_t(x) - U_t(x)$, $\widetilde{\varepsilon}_t(x) \coloneqq \widetilde{q}_t(x) - q_t(x)$. We remark that $\widetilde{U}_t$ is continuous on the $t$-direction when $t \in [t_i, t_{i+1}]$ but it may not be continuous crossing each $t_i$.

$$\begin{aligned} \left| \int_{\mathbb{R}^d} |\widetilde{\varepsilon}_\tau(x)|\, \mathrm{d}x - \int_{\mathbb{R}^d} |\widetilde{\varepsilon}_T(x)|\, \mathrm{d}x \right| &\le \sum_{i=0}^{N-1} \left| \int_{\mathbb{R}^d} |\widetilde{\varepsilon}_{T-t_i}(x)|\, \mathrm{d}x - \int_{\mathbb{R}^d} |\widetilde{\varepsilon}_{T-t_{i+1}}(x)|\, \mathrm{d}x \right| \\ &\le \sum_{i=0}^{N-1} \int_{T-t_{i+1}}^{T-t_i} \int_{\mathbb{R}^d} \left| (\nabla\cdot(q_t \widetilde{\delta}_t))(x) \right| \mathrm{d}x\, \mathrm{d}t, \end{aligned}$$

where we used Theorem A.1 on each interval $[t_i, t_{i+1}]$. We also notice that

$$\begin{aligned}\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left|(\nabla\cdot(q_t\widetilde{\delta}_t))(x)\right|\,\mathrm{d}x\mathrm{d}t \le& \int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left|(\nabla\cdot(q_t\delta_t))(x)\right|\,\mathrm{d}x\mathrm{d}t\\ &+\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left|(\nabla\cdot(q_t(\widetilde{\delta}_t-\delta_t)))(x)\right|\,\mathrm{d}x\mathrm{d}t,\end{aligned}\tag{46}$$

where the summation of the first term on the righthand side from $i=0$ to $i=N-1$ is $\int_\tau^T E(t)\mathrm{d}t$ and we have estimated this error term in Seciton C and also obtained Theorem 3.4,

$$\sum_{i=0}^{N-1}\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left|(\nabla\cdot(q_t\delta_t))(x)\right|\,\mathrm{d}x\mathrm{d}t \le C_u\cdot d\cdot T^{\frac14}\cdot(TL+T\cdot\tau^{-2}\cdot D^3)^{\frac12}\cdot\delta^{\frac12},\tag{47}$$

where under Assumption 3.7, $\mathcal{L}$ in Theorem 3.4 is bounded by $TL$. This gives the score matching error in Theorem 3.9.

The second term on the righthand side of (46) is

$$\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left|(\nabla\cdot(q_t(\widetilde{U}_t-\widehat{U}_t)))(x)\right|\,\mathrm{d}x\mathrm{d}t,$$

which can be further bounded as (the term corresponding to the derivative $\partial_1$)

$$\begin{aligned}\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left|(\partial_1(q_t(\widetilde{U}_t^1-\widehat{U}_t^1)))(x)\right|\,\mathrm{d}x\mathrm{d}t \le& \int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\|\widetilde{U}_t(x)-\widehat{U}_t(x)\|_\infty\cdot|\partial_1 q_t(x)|\,\mathrm{d}x\mathrm{d}t\\ &+\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\|\nabla\widetilde{U}_t(x)-\nabla\widehat{U}_t(x)\|_\infty\cdot q_t(x)\,\mathrm{d}x\mathrm{d}t,\end{aligned}\tag{48}$$

where we use the notation $\widetilde{U}_t(x)=(\widetilde{U}_t^1(x),\widetilde{U}_t^2(x),\ldots,\widetilde{U}_t^d(x))$. By (45), we obtain that

$$\|\widetilde{V}_t(x)-V_t(x)\|_\infty,\quad \|\nabla(\widetilde{V}_t(x)-V_t(x))\|_\infty\le C(p,s)\cdot L((\sqrt{d}+\|x\|)\sqrt{d}hL)^p,$$

and the definition that $\widetilde{U}_t=\widetilde{V}_{T-t}$, $\widehat{U}_t=\widehat{V}_{T-t}$, we know that the right hand side of (48) can be bounded by

$$C(p,s)L^{p+1}(h\sqrt{d})^p\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}\left(\sqrt{d}+\|x\|\right)^p\cdot(|\partial_1 q_t(x)|+q_t(x))\,\mathrm{d}x\mathrm{d}t.\tag{49}$$

According to Lemma B.3, the integral can be bounded by

$$\int_{T-t_{i+1}}^{T-t_i}\int_{\mathbb{R}^d}(\sqrt{d}+\|x\|^p)\cdot\left(\frac{(|x_1|+\lambda_t D)}{\sigma_t^2}+1\right)\cdot q_t(x)\,\mathrm{d}x\mathrm{d}t.\tag{50}$$

The above integral can be bounded by using the following two relations.

$$\begin{aligned}&\int_{\mathbb{R}^d}|y_1|(\sqrt{d}+\|y\|)^p\cdot q_t(x)\,\mathrm{d}x=\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{|y_1|(\sqrt{d}+\|y\|)^p}{(\sqrt{2\pi}\sigma_t)^d}\cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}}\cdot\mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x\mathrm{d}y\\ &\le\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{(|y_1-x_1|+|x_1|)(\sqrt{d}+\|y-x\|+\|x\|)^p}{(\sqrt{2\pi}\sigma_t)^d}\cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}}\cdot\mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x\mathrm{d}y\\ &\le\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{(|y_1-x_1|+\lambda_t D)(\sqrt{d}(1+\lambda_t D)+\|y-x\|)^p}{(\sqrt{2\pi}\sigma_t)^d}\cdot e^{-\frac{\|y-x\|^2}{2{\sigma_t}^2}}\cdot\mu_*\left(\frac{x}{\lambda_t}\right)\frac{1}{{\lambda_t}^d}\,\mathrm{d}x\mathrm{d}y\\ &\le\int_{\mathbb{R}^d}\frac{(|z_1|+\lambda_t D)(\sqrt{d}(1+\lambda_t D)+\|z\|)^p}{(\sqrt{2\pi}\sigma_t)^d}\cdot e^{-\frac{\|z\|^2}{2{\sigma_t}^2}}\mathrm{d}z\\ &\le C(p)d^{p/2}(\sigma_t+\lambda_t D)((1+\lambda_t D+\sigma_t)^p)\le 4^pC(p)d^{p/2}D^{p+1},\end{aligned}\tag{51}$$

where we used that $D \geq 1 \geq \sigma_t$ and $\lambda_t \leq 1$; and similarly

$$\int_{\mathbb{R}^d} (\sqrt{d} + \|y\|)^p q_t(x)\, \mathrm{d}x \leq 4^p C(p) d^{p/2} D^p, \tag{52}$$

where $C(p)$ is constant depending only on $p$. Combine these above estimates (51) and (52), we see that

$$(50) \leq 4^{p+1} \int_{T-t_{i+1}}^{T-t_i} \frac{C(p) d^{p/2} D^{p+1}}{\sigma_t^2} \mathrm{d}t. \tag{53}$$

Finally by pluggin (50) and (53) back into (49), we conclude

$$\begin{aligned}\sum_{i=0}^{N-1} \int_{T-t_{i+1}}^{T-t_i} \int_{\mathbb{R}^d} \left| (\nabla \cdot (q_t(\widetilde{U}_t - \widehat{U}_t)))(x) \right| \mathrm{d}x\mathrm{d}t &\leq C(p,s) L^{p+1} (hd)^p D^{p+1} \cdot d \int_\tau^T \frac{1}{\sigma_t^2}\, \mathrm{d}t \\ &\leq C(p,s) d (hd)^p (LD)^{p+1} \log(T/\tau).\end{aligned} \tag{54}$$

This gives the discretization error in Theorem 3.9. Theorem 3.9 follows from combining (47) and (54). ☐

*Proof of Proposition D.1.* For simplicity of notations, we will first prove the statement for $p = 2$ for the Heun's method, which is a 2-stage second order Runge-Kutta method. For general $s$-stage $p$-th Runge-Kutta method, the proof is similar, and we will point out the necessary changes at the end of the proof.

For the Heun's method, $F_r(x)$ is given by

$$F_r(x) = x + \frac{r}{2}(\widehat{V}_{t_i}(x) + \widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))). \tag{55}$$

In the following, we prove Proposition D.1 for the Heun's method, under the assumption that the step size satisfies $8dhL \leq 1$.

For the vector $\widehat{V}_{t_i}(x)$, we first notice that, by Assumption 3.7 that

$$\|s_t(x)\| \leq L(\sqrt{d} + \|x\|).$$

It follows that

$$\|\widehat{V}_{t_i}(x)\|, \|\widehat{V}_{t_i+r}(x)\| \leq \|x\| + L(\sqrt{d} + \|x\|) \leq 2L(\sqrt{d} + \|x\|), \tag{56}$$

and

$$\begin{aligned}\|F_r(x) - x\| &\leq \frac{r}{2}\left(\|\widehat{V}_{t_i}(x)\| + \|\widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))\|\right) \\ &\leq \frac{r}{2}\left(2L(\sqrt{d} + \|x\|) + 2L(\sqrt{d} + \|x + r\widehat{V}_{t_i}(x)\|)\right) \\ &\leq (2rL + (2rL)^2)(\sqrt{d} + \|x\|) \leq 4rL(\sqrt{d} + \|x\|) \leq \frac{1}{2}(\sqrt{d} + \|x\|),\end{aligned} \tag{57}$$

where we used that $8rL \leq 8hdL \leq 1$. It follows that

$$\sqrt{d} + \|F_r(x)\| \geq \sqrt{d} + \|x\| - \|F_r(x) - x\| \geq \frac{1}{2}(\sqrt{d} + \|x\|). \tag{58}$$

Next we show the map $F_r(x)$ from (55) is a local diffeomorphism, we check its Jacobian matrix

$$DF_r(x) = \mathbb{I}_d + A, \quad A := \frac{r}{2} D\widehat{V}_{t_i}(x) + \frac{r}{2} D\widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))(\mathbb{I}_d + rD\widehat{V}_{t_i}(x)). \tag{59}$$

We notice that $D\widehat{V}_{t_i}(x) = \mathbb{I}_d + Ds_{t_i}(x)$, and each entry of $Ds_{t_i}(x)$ is bounded by $L$. The same bound holds for the entries of $D\widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))$. Thus the $(i,j)$-th entry of $A$ is bounded by

$$|A_{ij}| \leq \frac{r}{2}(1 + L) + \frac{r^2 d}{2}(1 + L)^2 \leq rL + 2dr^2L^2 \leq 2rL, \quad 1 \leq i, j \leq d, \tag{60}$$

where we used that $L \geq 1$ and provided $2rdL \leq 8hdL \leq 1$. The spectral norm $\|A\|$ of the matrix $A$ is bounded by its Frobenius norm as

$$\|A\|_{\text{norm}} \leq \|A\|_{\text{F}} \leq \sqrt{\sum_{ij} A_{ij}^2} \leq 2rdL \leq 1/2,$$

where again we used that $2rdL \leq 8hdL \leq 1$.

It follows that $DF_r(x) = \mathbb{I}_d + A$ is invertible, and $F_r$ is a local diffeomorphism, and then Hadamard-Caccioopoli theorem implies that $F_r$ is also a bijection from $\mathbb{R}^d$ to itself. Therefore, $F_r$ is a diffeomorphism from $\mathbb{R}^d$ to itself. Moreover, thanks to (60), we have the following entrywise bound for the inverse matrix $(DF_r(x))^{-1}$:

$$\begin{aligned} |((DF_r(x))^{-1} - \mathbb{I}_d)_{ij}| &= |(\mathbb{I}_d + A)^{-1}_{ij} - \delta_{ij}| \leq |(\mathbb{I}_d - A + A^2 - A^3 + \cdots)_{ij} - \delta_{ij}| \\ &\leq \sum_{k\geq 1} (2rL)^k d^{k-1} \leq 4rL, \quad 1 \leq i,j \leq d, \end{aligned} \tag{61}$$

where we used that $4rdL \leq 8hdL \leq 1$.

We denote the functional inverse of $F_r$ as $\Phi_r(x)$, then (44) follows from (42).

Next we show

$$\begin{aligned} &\|\widetilde{V}_{t_i+r}(F_r(x)) - \widehat{V}_{t_i+r}(F_r(x))\|_\infty \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}rL)^2, \\ &\|\nabla(\widetilde{V}_{t_i+r}(F_r(x)) - \widehat{V}_{t_i+r}(F_r(x)))\|_\infty \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}rL)^2. \end{aligned} \tag{62}$$

and the claim (45) follows. In fact, if we denote $y = F_r(x)$, then

$$\|\widetilde{V}_{t_i+r}(y) - \widehat{V}_{t_i+r}(y)\|_\infty \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2 \leq 4C_u L((\sqrt{d} + \|y\|)\sqrt{d}rL)^2,$$

where in the last inequality we used (58).

For the gradient, by the chain rule we have

$$(\nabla\widetilde{V}_{t_i+r})(y) - (\nabla\widehat{V}_{t_i+r})(y) = \nabla(\widetilde{V}_{t_i+r}(F_r(x)) - \widehat{V}_{t_i+r}(F_r(x)))DF_r(x)^{-1}. \tag{63}$$

By plugging (61) into (63), we conclude that

$$\begin{aligned} \|(\nabla\widetilde{V}_{t_i+r})(y) - (\nabla\widehat{V}_{t_i+r})(y)\|_\infty &\leq (1 + 4rdL)\|\nabla(\widetilde{V}_{t_i+r}(F_r(x)) - \widehat{V}_{t_i+r}(F_r(x)))\|_\infty \\ &\leq 2C_u L((\sqrt{d} + \|x\|)\sqrt{d}rL)^2 \leq 8C_u L((\sqrt{d} + \|y\|)\sqrt{d}rL)^2, \end{aligned} \tag{64}$$

where we used that $4rdL \leq 8hdL \leq 1$, and in the last inequality we used (58).

In the rest we prove (62). Explicitly $\widetilde{V}_{t_i+r}(F_r(x))$ is given by

$$\widetilde{V}_{t_i+r}(F_r(x)) = \partial_r F_r(x) = \frac{1}{2}(\widehat{V}_{t_i}(x) + \widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))) + \frac{r}{2}\partial_r(\widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i(x)})). \tag{65}$$

We can perform Taylor expansions for the two terms on the righthand side of (65)

$$\begin{aligned} &\frac{1}{2}(\widehat{V}_{t_i}(x) + \widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))) = \widehat{V}_{t_i}(x) + \frac{r}{2}\partial_t\widehat{V}_{t_i}(x) + \frac{r}{2}\nabla\widehat{V}_{t_i}(x)\widehat{V}_{t_i}(x) + R_1'(r,x), \\ &R_1'(r,x) := \frac{1}{4}\int_0^r s\partial_s^2(\widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x)))\mathrm{d}s. \end{aligned} \tag{66}$$

For the second term on the righthand side of (65), we can rewrite it as

$$\begin{aligned} &\frac{r}{2}\partial_r(\widehat{V}_{t_i+r}(x + r\widehat{V}_{t_i}(x))) = \frac{r}{2}(\partial_r\widehat{V}_{t_i}(x) + \nabla\widehat{V}_{t_i}\widehat{V}_{t_i}(x)) + R_1''(r,x), \\ &R''(r,x) = \frac{r}{2}\int_0^r \partial_s^2((\widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))))\mathrm{d}s, \end{aligned} \tag{67}$$

and

$$\begin{aligned}\widehat{V}_{t_i+r}(F_r(x)) &= \widehat{V}_{t_i+r}(x + \frac{r}{2}\left(\widehat{V}_{t_i}(x) + \widehat{V}_{t_{i+r}}\left(x + r\widehat{V}_{t_i}(x)\right)\right) \\ &= \widehat{V}_{t_i}(x) + r\partial_t \widehat{V}_{t_i}(x) + r\nabla \widehat{V}_{t_i}(x)\widehat{V}_{t_i}(x) + R_2(r,x), \\ R_2(r,x) :&= \frac{1}{2}\int_0^r s\partial_s^2(\widehat{V}_{t_i+s}(F_s(x)))\mathrm{d}s.\end{aligned} \tag{68}$$

Comparing (66), (67) and (68), all the leading terms cancel out, and we get

$$\widetilde{V}_{t_i+r}(F_r(x)) - \widehat{V}_{t_i+r}(F_r(x)) = R_1'(r,x) + R_1''(r,x) - R_2(r,x). \tag{69}$$

Next we show that under Assumption 3.7 with $p=2$, the error terms satisfy

$$\begin{aligned}\|R_1'(r,x)\|_\infty, \|R_1''(r,x)\|_\infty, \|R_2(r,x)\|_\infty &\leq C_u L((\sqrt{d} + \|x\|)\sqrt{d} rL)^2, \\ \|\nabla R_1'(r,x)\|_\infty, \|\nabla R_1''(r,x)\|_\infty, \|\nabla R_2(r,x)\|_\infty &\leq C_u L((\sqrt{d} + \|x\|)\sqrt{d} rL)^2.\end{aligned} \tag{70}$$

The claim (62) then follows (69) and its gradient given in (70).

The three error terms $R_1', R_1'', R_2$ involves the two terms $\partial_s^2(\widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))), \partial_s^2(\widehat{V}_{t_i+s}(F_s(x)))$. We first exam the time derivatives of $\widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))$. Its first derivative gives

$$\partial_s \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x)) + \nabla \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x)],$$

and its second derivative is

$$\partial_s^2 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x)) + 2\partial_s \nabla \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x)] + \nabla^2 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x), \widehat{V}_{t_i}(x)]. \tag{71}$$

By Assumption 3.7, the entries of the vector $\partial_s^2 \widehat{V}_{t_i+s}$, the matrices $\nabla \widehat{V}_{t_i+s}, \partial_s \nabla \widehat{V}_{t_i+s}$ and the tensor $\nabla^2 \widehat{V}_{t_i+s}$ are all bounded by $2L$. From (56), and the relation $\|y\|_1 \leq \sqrt{d}\|y\|$, we have

$$\|\widehat{V}_{t_i}(x)\|_1 \leq \sqrt{d}\|\widehat{V}_{t_i}(x)\| \leq 2(\sqrt{d} + \|x\|)\sqrt{d}L.$$

Thus the entries of the vectors in (71) are bounded by $C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2$:

$$\begin{aligned}&\|\partial_s^2 \widehat{V}_{t_i+s}\|_\infty \leq L, \\ &\|\partial_s \nabla \widehat{V}_{t_i+s}[\widehat{V}_{t_i}]\|_\infty \leq \|\partial_s \nabla \widehat{V}_{t_i+s}\|_\infty \|\widehat{V}_{t_i}\|_1 \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}L), \\ &\|\nabla^2 \widehat{V}_{t_i+s}[\widehat{V}_{t_i}, \widehat{V}_{t_i}]\|_\infty \leq \|\nabla^2 \widehat{V}_{t_i+s}\|_\infty \|\widehat{V}_{t_i}\|_1^2 \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2,\end{aligned} \tag{72}$$

where we used Assumption 3.7 and (56).

For $\partial_s^2(\widehat{V}_{t_i+s}(F_s(x)))$, by the same argument as above we also have $\|\partial_s^2(\widehat{V}_{t_i+s}(F_s(x)))\|_\infty \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2$. We conclude from (66), (67) and (68) that

$$\begin{aligned}\|R_1'(r,x)\|_\infty, \|R_1''(r,x)\|_\infty, \|R_2(r,x)\|_\infty &\leq r\int_0^r C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2 \mathrm{d}r \\ &\leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}rL)^2.\end{aligned} \tag{73}$$

To get the $C^1$ bound, we need to estimate $\|\nabla R_1'(r,x)\|_\infty, \|\nabla R_1''(r,x)\|_\infty, \|\nabla R_2(r,x)\|_\infty$. By the explicit formulas in (66), (67) and (68) bound, it boils down to show that

$$\|\partial_s^2 \nabla(\widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x)))\|_\infty, \|\partial_s^2 \nabla(\widehat{V}_{t_i+s}(F_s(x)))\|_\infty \leq C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2. \tag{74}$$

From the expression (71), we take one more gradient on $x$,

$$\begin{aligned}&\partial_s^2 \nabla(\widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))) = \partial_s^2 \nabla \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\mathbb{I}_d + s\nabla \widehat{V}_{t_i}(x)] \\ &+ 2\partial_s \nabla^2 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x), \mathbb{I}_d + s\nabla \widehat{V}_{t_i}(x)] + 2\partial_s \nabla \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\nabla \widehat{V}_{t_i}(x)] \\ &+ \nabla^3 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x), \widehat{V}_{t_i}(x), \mathbb{I}_d + s\nabla \widehat{V}_{t_i}(x)] + 2\nabla^2 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x), \nabla \widehat{V}_{t_i}(x)].\end{aligned} \tag{75}$$

Similarly to the argument as for (72), we have

$$
\begin{aligned}
&\|\partial_s^2 \nabla \widehat{V}_{t_i+s}[\mathbb{I}_d + s\nabla \widehat{V}_{t_i}]\|_\infty \le \|\partial_s^2 \nabla \widehat{V}_{t_i+s}\|_\infty (1 + rd\|\nabla \widehat{V}_{t_i}\|_\infty),\\
&\|\partial_s \nabla^2 \widehat{V}_{t_i+s}[\widehat{V}_{t_i}, \mathbb{I}_d + s\nabla \widehat{V}_{t_i}]\|_\infty \le \|\partial_s \nabla^2 \widehat{V}_{t_i+s}\|_\infty \|\widehat{V}_{t_i+s}\|_1 (1 + rd\|\nabla \widehat{V}_{t_i}\|_\infty),\\
&\|\partial_s \nabla \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\nabla \widehat{V}_{t_i}(x)]\|_\infty \le \|\partial_s \nabla \widehat{V}_{t_i+s}\|_\infty d \|\nabla \widehat{V}_{t_i}\|_\infty,\\
&\|\nabla^3 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x), \widehat{V}_{t_i}(x), \mathbb{I}_d + s\nabla \widehat{V}_{t_i}(x)]\|_\infty \le \|\nabla^3 \widehat{V}_{t_i+s}\|_\infty \|\widehat{V}_{t_i+s}\|_1^2 (1 + rd\|\nabla \widehat{V}_{t_i}\|_\infty),\\
&\|\nabla^2 \widehat{V}_{t_i+s}(x + s\widehat{V}_{t_i}(x))[\widehat{V}_{t_i}(x), \nabla \widehat{V}_{t_i}(x)]\|_\infty \le d\|\nabla^2 \widehat{V}_{t_i+s}\|_\infty \|\widehat{V}_{t_i+s}\|_1 \|\nabla \widehat{V}_{t_i}\|_\infty.
\end{aligned}
$$

Thanks to Assumption 3.7 and (56), all of the above expressions are bounded by

$$C_u (L((\sqrt{d} + \|x\|)\sqrt{d}L)^2.$$

For $\partial_s^2 \nabla(\widehat{V}_{t_i+s}(F_s(x)))$, by the same argument as above we also have $\|\partial_s^2(\widehat{V}_{t_i+s}(F_s(x)))\|_\infty \le C_u L((\sqrt{d} + \|x\|)\sqrt{d}L)^2$. We conclude from (66), (67) and (68) that

$$\|\nabla R_1'(r,x)\|_\infty, \|\nabla R_1''(r,x)\|_\infty, \|\nabla R_2(r,x)\|_\infty \le C_u L((\sqrt{d} + \|x\|)\sqrt{d} r L)^2. \tag{76}$$

The estimates (73) and (76) together give (70). This finishes the proof of Proposition D.1 for the Heun's method.

In general, for the $p$-th order Runge-Kutta methods, the proof is similar. We need to perform a Taylor expansion as in (66), (67) and (68), up to the $p$-th order, and the error involves the $p$-th time derivative. The main terms all cancel out thanks to the choice of Runge–Kutta matrix $[a_{jk}]$, weights $b_j$ and nodes $c_j$. For the error term, like the discussion above, each more time derivative gives an extra factor of $(\sqrt{d} + \|x\|)\sqrt{d}L$. So $p$-th derivative leads to an error $C(p,s)L((\sqrt{d} + \|x\|)h\sqrt{d}L)^p$, where the constant $C(p,s)$ depends only on the order and stages of the Runge-Kutta methods. This gives (45). □

[1]Beijing International Center for Mathematical Research, Center for Machine Learning Research, Peking University, Beijing, China
*Email address*: huangdz@bicmr.pku.edu.cn

[2]University of Pennsylvania, Philadelphia, PA, USA
*Email address*: huangjy@wharton.upenn.edu

[3]Courant Institute, New York University, NY, USA
*Email address*: malin@nyu.edu