# CHARACTERISTIC LEARNING FOR PROVABLE ONE STEP GENERATION

ZHAO DING[1], CHENGUANG DUAN[1], YULING JIAO[1], RUOXUAN LI[1], JERRY ZHIJIAN YANG[1,*],
AND PINGWEN ZHANG[1,2]

ABSTRACT. We propose the characteristic generator, a novel one-step generative model that combines the efficiency of sampling in Generative Adversarial Networks (GANs) with the stable performance of flow-based models. Our model is driven by characteristics, along which the probability density transport can be described by ordinary differential equations (ODEs). Specifically, we first estimate the underlying velocity field and use the Euler method to solve the probability flow ODE, generating discrete approximations of the characteristics. A deep neural network is then trained to fit these characteristics, creating a one-step map that pushes a simple Gaussian distribution to the target distribution. In the theoretical aspect, we provide a comprehensive analysis of the errors arising from velocity matching, Euler discretization, and characteristic fitting to establish a non-asymptotic convergence rate in the 2-Wasserstein distance under mild data assumptions. Crucially, we demonstrate that under a standard manifold assumption, this convergence rate depends only on the intrinsic dimension of data rather than the much larger ambient dimension, proving our model's ability to mitigate the curse of dimensionality. To our knowledge, this is the first rigorous convergence analysis for a flow-based one-step generative model. Experiments on both synthetic and real-world datasets demonstrate that the characteristic generator achieves high-quality and high-resolution sample generation with the efficiency of just a single neural network evaluation.

## 1. INTRODUCTION

Generative models aim to learn and sample from an underlying target distribution, finding applications in diverse fields such as image and video generation (Radford et al., 2016, Meng et al., 2022, Ho et al., 2022), text-to-image generation (Ramesh et al., 2021, 2022, Kang et al., 2023), and speech synthesis (Kong et al., 2021, Chen et al., 2021). One of the most influential and widely-used approaches is GAN (Goodfellow et al., 2014) and its variants (Arjovsky et al., 2017). GANs offer the advantage of high sampling efficiency, as generating new samples merely entails a single evaluation of the trained generator. Despite the remarkable success in practical applications (Reed et al., 2016) and theoretical guarantee (Liang, 2021, Liu et al., 2021, Huang et al., 2022, Zhou et al., 2023), GANs have intrinsic limitations in terms of their stability (Salimans et al., 2016).

In recent years, diffusion models (Ho et al., 2020, Song et al., 2021b,c, Karras et al., 2022) and flow-based models (Liu et al., 2022, Lipman et al., 2023) have emerged as powerful generative models. These models have also laid the foundation for the development of generative AI models, such as DALL-E (Ramesh et al., 2021, 2022), Midjourney, Stable Diffusion (Esser et al., 2024), and Sora (Brooks et al., 2024). Theoretical analysis for these methods has been studied by Oko et al. (2023), Lee et al. (2022, 2023), Chen et al. (2023d,c), Benton et al. (2024a,b), Gao and Zhu (2024), Wu et al. (2024). Although diffusion or flow-based models outperform GANs in generation quality across various tasks (Dhariwal and Nichol, 2021), they require hundreds or even thousands of sequential steps involving large neural network evaluations for sampling. As a consequence, their sampling speed is much slower compared to one-step GANs.

The instability of GANs and the inefficiency of sampling in diffusion or flow-based models have emerged as significant bottlenecks in practical applications of generative models. This raises two crucial questions:

> *How can we develop a one-step generative model that combines the efficient sampling of GANs with the stable performance of diffusion or flow-based models? How can we establish a rigorous error analysis for this generative model?*

Several recent papers have made progress in addressing the first question using various techniques such as distillation (Luhman and Luhman, 2021, Salimans and Ho, 2022, Song et al., 2023, Zhou et al., 2024), operator learning (Zheng et al., 2023a), or trajectory models (Kim et al., 2024, Ren et al., 2024). For a more detailed discussion, please refer to Section 5.1. Despite these recent advancements, a unifying mathematical framework for designing and analyzing the one-step generative models remains largely limited (Li et al., 2024b). This paper aims to fill this gap and provide an answer to the aforementioned questions. Specifically, we introduce a comprehensive framework, known as the characteristic generator, aiming to streamline the sampling process using a single evaluation of the neural network. Our model merges the sampling efficiency of GANs with the promising performance of flow-based models. Furthermore, we present a rigorous error analysis for the characteristic generator. Through numerical experiments, we validate that our approach generates high-quality samples that is comparable to those generated through flow-based models, all while requiring just a single evaluation of the neural network.

## 1.1. **Contributions.** Our contributions are summarized as follows:

(i) We introduce the "characteristic generator," a one-step generative model that forms the basis of a unified mathematical framework for recent flow-based distillation methods. Our approach involves modeling the probability density transport equation with stochastic interpolants, which yields a probability flow ODE. By learning to approximate the characteristic curves of this ODE, the generator can directly transform a prior distribution to the target distribution without iterative simulation.

(ii) We provide a rigorous error analysis for the characteristic generator. Specifically, we derive a convergence rate $\mathcal{O}(n^{-\frac{1}{d+3}})$ for velocity matching (Theorem 3.10). Additionally, we propose an error bound for the 2-Wasserstein distance between the distribution of data generated by the Euler method and the target distribution (Theorem 3.12), which is of independent interest. Lastly, we present a non-asymptotic convergence rate for the characteristic generator in the 2-Wasserstein distance (Theorem 3.14). These findings also provide valuable theoretical understanding for distillation (Salimans and Ho, 2022, Song et al., 2023), operator learning (Zheng et al., 2023a), or trajectory model (Kim et al., 2024, Ren et al., 2024).

(iii) We prove that the characteristic generator mitigates the curse of dimensionality (Corollaries 3.19 and 3.20) under a standard manifold assumption. We show the convergence rate depends on the data's low intrinsic dimension rather than the high ambient dimension. To our knowledge, this is the first end-to-end analysis that formally explains the strong performance of flow-based one-step generative models on high-dimensional data.

(iv) We conduct extensive experiments on synthetic and real-world data (CIFAR-10, CelebA-HQ), demonstrating that the characteristic generator produces high-quality samples in a single network evaluation. Our model significantly outperforms prior non-adversarial one-step models (Salimans and Ho, 2022, Song et al., 2023, Kim et al., 2024) on CIFAR-10. Furthermore, it achieves performance competitive with state-of-the-art methods (Kim et al., 2024, CTM) in just a few iterations, without requiring additional adversarial training. The successful application to high-resolution image generation ($256 \times 256$ and $512 \times 512$) further validates the scalability of our approach.

1.2. **Main Results.** Let $\mu_0 \in P_{\mathrm{ac}}(\mathbb{R}^d)$ be a known prior distribution, and let $\mu_1 \in P_{\mathrm{ac}}(\mathbb{R}^d)$ be the target distribution with an unknown density function $\rho_1(x)$. Suppose one has access to finite data samples from $\mu_1$. In generative learning, we aim to learn a push-forward map $G^* : \mathbb{R}^d \to \mathbb{R}^d$ that transports the prior distribution $\mu_0$ onto the target distribution $\mu_1$. This relationship is defined by the normalizing equation (Rozen et al., 2021):

$$(1.1) \qquad\qquad G^*_\sharp \mu_0 = \mu_1.$$

Formally, the goal of the generative learning is to find an estimator $\widehat{G}$ of the push-forward operator $G^*$ based on finite samples drawn from $\mu_1$.

*Remark* 1.1 (Regularity of the push-forward map). The learnability of the push-forward map $G^*$ requires its regularity. Without crucial properties like boundedness and Lipschitz continuity, the map can be ill-behaved, making it practically impossible for function approximators like neural networks to learn. In this work, we define $G^*$ via a probability flow ODE (1.2), and establish the necessary regularity conditions in Section 3.2.

In this work, we construct the desired push-forward operator via a probability flow ODE:

$$(1.2) \qquad\qquad \mathrm{d}x(t) = b^*(t, x(t))\,\mathrm{d}t, \quad x(0) = x_0 \sim \mu_0,$$

where the velocity field $b^*$ is given as (2.3), and $x(t)$ is knwon as the characteristic curve of the transport equation (2.2). Denote by $\mu_t$ the distribution of $x(t)$ for each $t \in (0,1)$, and define the flow $g_{t,s}^*$ as $g_{t,s}^*(x(t)) = x(s)$ for each $0 \leq t \leq s \leq 1$, which transports particles along the characteristic curves. It is apparent that $(g_{t,s}^*)_\sharp \mu_t = \mu_s$. Then our target push-forward map is defined as $G^* := g_{0,1}^*$. Let $\widehat{b}$ be the estimated velocity field obtained by velocity matching (2.8), and let $\widehat{E}_{0,K}^\tau$ be the numerical approximation to the ODE solution by Euler method (2.9). Denote by $\widehat{g}_{s,t}$ be an estimation of $g_{t,s}^*$ defined as (2.11), which is referred to the characteristic generator.

Our theoretical results are established under the following assumptions on the prior and target distributions, which will be discussed in Section 3.1.

**Assumption 1** (Prior distribution). *The prior distribution $\mu_0 = N(0, I_d)$.*

**Assumption 2** (Target distribution). *There exists an unknown constant $\sigma > 0$, such that*

$$\mu_1(x) = N(0, \sigma^2 I_d) * \nu := \int \varphi_d\Big(\frac{x - x'}{\sigma}\Big) \,\mathrm{d}\nu(x'),$$

*where $\varphi_d$ represents the density of the $d$-dimensional standard Gaussian distribution, and $\mathrm{d}\nu(x) = p(x)\,\mathrm{d}x$ with $\mathrm{supp}(\nu) \subseteq [0,1]^d$.*

Assumption 2 requires the target distribution to be a Gaussian convolution. In other words, for each random variable $X \sim \mu_1$, there exist two independent random variables $Z \in [0,1]^d$ and $W \sim N(0, I_d)$, such that $X \overset{\mathrm{d}}{=} Z + \sigma W$. This assumption is essential as it ensures desirable properties of the probability flow ODE, such as bounded moments and the Lipschitz property of the velocity field. Further details can be found in Sections 3.1 and 3.2. It is noteworthy that this assumption can be considered relatively mild, given that the smoothed distribution $\mu_1$ is an approximation of the original distribution $\nu$, particularly when the variance $\sigma^2$ of the Gaussian distribution is small. In addition, it encompasses non-log-concave distributions, such as Gaussian mixtures (Grenioux et al., 2024, Appendix C). Similar assumptions have been considered by Saremi et al. (2024), Grenioux et al. (2024), Beyler and Bach (2025); however, Beyler and Bach (2025) requires prior knowledge of the variance $\sigma^2$.

Our first main result gives an error bound for the velocity matching.

**Theorem 1.2** (Informal version of Theorem 3.10). *Suppose that Assumptions 1 and 2 hold. Let $\mathbb{S}$ be a set of $n$ samples independently and identically drawn from the target distribution $\mu_1$. Suppose the depth and width of the velocity neural network are set properly. Then it follows that*

$$\mathbb{E}_{\mathbb{S}}\Big[\frac{1}{T}\int_0^T \mathbb{E}_{X_t \sim \mu_t}\big[\|b^*(t, X_t) - \widehat{b}(t, X_t)\|_2^2\big]\,\mathrm{d}t\Big] \leq C_T n^{-\frac{2}{d+3}}\log^2 n,$$

*where $C_T$ is a constant depending on $d$, $\sigma$ and $T$.*

We then present the 2-Wasserstein bound for generated data by Euler method.

**Theorem 1.3** (Informal version of Corollary 3.13). *Suppose that Assumptions 1 and 2 hold. Let $\mathbb{S}$ be a set of $n$ samples independently and identically drawn from the target distribution $\mu_1$. Suppose*

*the depth and width of the velocity neural network are set properly. Let $K$ be the number of time steps of Euler method. Then the following inequality holds*

$$\mathbb{E}_{\mathcal{S}}\Big[W_2^2\big((\widehat{E}_{0,K}^{\tau})_{\sharp}\mu_0, \mu_1\big)\Big] \leq C_T^1\Big\{n^{-\frac{2}{d+3}}\log^2 n + \frac{\log n}{K^2}\Big\} + C_T^2 W_2^2(\mu_0, \mu_1),$$

*where $C_T^1$ and $C_T^2$ are two constants depending on $d$, $\sigma$ and $T$. Further, as the stopping time $T \to 1$, the constant $C_T^1$ tends to infinity polynomially while $C_T^2$ decreases to zero polynomially.*

The averaged 2-Wasserstein bound for characteristic generator is stated as follows.

**Theorem 1.4** (Informal version of Theorem 3.14). *Suppose that Assumptions 1 and 2 hold. Let $\mathcal{S}$ be a set of $n$ samples independently and identically drawn from the target distribution $\mu_1$, and let $\mathcal{Z}$ be the set of numerical solutions by Euler method with a sufficiently large number. Suppose the depth and width of the velocity neural network and the characteristic neural network are set properly, respectively. Let $m$ be the number of time steps of Euler method, and $m$ be the number of samples for characteristic fitting. Then the following inequality holds*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{Z}}\Big[\frac{2}{T^2}\int_0^T\int_t^T W_2^2\big((\widehat{g}_{t,s})_{\sharp}\mu_t, \mu_s\big)\,\mathrm{d}s\mathrm{d}t\Big]$$
$$\leq C_T^1\Big\{n^{-\frac{2}{d+3}}\log^2 n + \frac{\log n}{K^2}\Big\} + C\Big\{m^{-\frac{2}{d+4}}\log^2 m + \frac{\log m}{K}\Big\},$$

*where $C_T^1$ is a constant depending on $d$, $\sigma$ and $T$, and $C$ is a constant depending on $d$ and $\sigma$. Further, as the stopping time $T \to 1$, the constant $C_T^1$ tends to infinity.*

As a consequence of Theorem 3.14, we propose in Table 1 the convergence rate of characteristic generators induced by two special probability flow ODEs.

TABLE 1. Convergence rates of characteristic generator.

| Probability flow ODE | Assumption | Convergence rate | |
|---|---|---|---|
| Linear interpolants | Assumptions 1 and 2 | $\mathcal{O}(n^{-\frac{2}{3(d+3)}})$ | Corollary 3.16 |
| Föllmer flow | Assumptions 1 and 2 | $\mathcal{O}(n^{-\frac{2}{5(d+3)}})$ | Corollary 3.17 |

The convergence rates in Table 1 suffer from the curse of dimensionality. However, in many practical applications, high-dimensional data – such as images or text documents – often lie close to a low-dimensional manifold embedded in the high-dimensional ambient space (Goodfellow et al., 2016, Bortoli, 2022, Jiao et al., 2023a, Oko et al., 2023). This observation motivates a refined assumption that captures the intrinsic low-dimensional structure of real-world data. We formalize this idea with the following assumption.

**Assumption 3.** There exists an unknown constant $\sigma > 0$ and $d^* \ll d$, such that

$$\mu_1 = N(0, \sigma^2 I_d) * (P_{\sharp}\tilde{\nu}),$$

where $P \in \mathbb{R}^{d \times d^*}$ is a matrix whose column vectors are orthonormal in $\mathbb{R}^d$, and $\tilde{\nu}$ is a distribution with $\mathrm{supp}(\tilde{\nu}) \subseteq [0, 1]^{d^*}$.

We note that the distribution $\tilde{\nu}$ in Assumption 3 is supported on a subset of a $d^*$-dimensional hypercube. Consequently, its push-forward $\nu := P_\sharp \tilde{\nu}$ lies on a $d^*$-dimensional manifold embedded in the ambient space $\mathbb{R}^d$. Although Gaussian convolution spreads $\nu$ throughout the full ambient space, the resulting distribution $\mu_1$ remains concentrated around the manifold, its density decaying exponentially with the distance from it. This manifold assumption is crucial, as it allows us to show that the convergence rate depends only on the *intrinsic dimension $d^*$* rather than the *ambient dimension $d$*, thereby effectively mitigating the curse of dimensionality. The convergence rates of characteristic generators under manifold assumption are summarized in Table 2.

TABLE 2. Convergence rates of characteristic generator under manifold assumption.

| Probability flow ODE | Assumption | Convergence rate | |
|---|---|---|---|
| Linear interpolants | Assumptions 1 and 3 | $\mathcal{O}(n^{-\frac{2}{3(d^*+3)}})$ | Corollary 3.19 |
| Föllmer flow | Assumptions 1 and 3 | $\mathcal{O}(n^{-\frac{2}{5(d^*+3)}})$ | Corollary 3.20 |

Experiment results and discussions can be found in Section 4. Our code is online available at https://github.com/burning489/CharacteristicGenrator.

### 1.3. **Preliminaries and Notations.**

1.3.1. *Wasserstein Distance.* Let $P_{\mathrm{ac}}(\mathbb{R}^d)$ be the space of probability measures on $\mathbb{R}^d$, which are absolutely continuous with respect to Lebesgue measure. Suppose $\mu_0, \mu_1 \in P_{\mathrm{ac}}(\mathbb{R}^d)$ with $\mathrm{d}\mu_0(x) = \rho_0(x)\,\mathrm{d}x$ and $\mathrm{d}\mu_1(x) = \rho_1(x)\,\mathrm{d}x$. The 2-Wasserstein distance (Villani, 2009, Definition 6.1) between $\mu_0$ and $\mu_1$ is defined by the formula

$$(1.3) \qquad W_2(\mu_0, \mu_1) = \inf\left\{ \mathbb{E}^{1/2}\big[\|X_0 - X_1\|_2^2\big] : \mathrm{law}(X_0) = \mu_0,\ \mathrm{law}(X_1) = \mu_1 \right\}.$$

The 2-Wasserstein distance satisfies the symmetry axiom and the triangle inequality. Further, the distance $W_2(\mu_0, \mu_1)$ is equal to zero if and only if $\mu_0 = \mu_1$.

1.3.2. *Deep Neural Networks.* A neural network $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ is a function defined by

$$f(x) = T_L(\varrho(T_{L-1}(\cdots \varrho(T_0(x)) \cdots))),$$

where the activation function $\varrho$ is applied component-wisely and $T_\ell(x) := A_\ell x + b_\ell$ is an affine transformation with $A_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell = 0, \ldots, L$. In this paper, we consider the case where $N_{L+1} = d$. The number $L$ is called the depth of neural networks. Additionally, $S := \sum_{\ell=0}^{L}(\|A_\ell\|_0 + \|b_\ell\|_0)$ represents the total number of non-zero weights within the neural network.

We denote by $N(L, S)$ the set of neural networks with depth at most $L$ and the number of non-zero weights at most $S$.

1.3.3. *Notations.* The set of positive integers is denoted by $\mathbb{N} = \{1, 2, \ldots\}$. We also denote $\mathbb{N}_0 = \{0\} \cup \mathbb{N}_+$ for convenience. For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we define its $\ell_p$-norms as $\|x\|_p^p = \sum_{i=1}^d |x_i|^p$ for $1 \le p < \infty$, with $\|x\|_\infty = \max_{1 \le i \le d} |x_i|$. Denote by $\langle \cdot, \cdot \rangle$ the inner product between vectors, that is, $\langle x, y \rangle = \sum_{k=1}^d x_k y_k$, where $y = (y_1, \ldots, y_d)$. For a matrix $A \in \mathbb{R}^{m \times n}$, the operator norm induced by the $\ell_2$ vector norm is defined as $\|A\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \|Ax\|_2$. Additionally, denote by $\mathbb{B}_R^\infty$ the $\ell_\infty$ ball in $\mathbb{R}^d$ with radius $R$, that is, $\mathbb{B}_R^\infty = \{x \in \mathbb{R}^d : \|x\|_\infty \le R\}$. For a matrix $M$, we say $M \succeq 0$ if and only if it is positive definite. Let $A$ and $B$ be two matrices, denote $A \succeq B$ if and only if $(A - B) \succeq 0$. For a function $u(t)$ of time $t$, the time derivative is denoted by $\dot{u}$ or $\partial_t u$. Further, let $\ddot{u}$ denote the second-order time derivative. Additionally, we use $\nabla$ and $\nabla\cdot$ to denote the spatial gradient and divergence operators, respectively.

1.4. **Organization.** The remainder of this article is organized as follows. Section 2 introduces the characteristic-driven generative learning method. A thorough analysis for this method is provided in Section 3. Section 4 presents numerical studies and discussions. Section 5 discusses related work and provides additional insights. Finally, Section 6 presents the conclusion and discusses future work. The theoretical proofs and experimental details are provided in the supplementary materials.

## 2. CHARACTERISTIC GENERATIVE LEARNING

Dating back to Moser (1965), Dacorogna and Moser (1990), researchers proposed a continuous dynamic-induced approach for solving the normalizing equation (1.1). In the field of deep generative learning, flow-based models utilize ODE-dynamics to construct probability flows, effectively pushing the prior distribution towards the target distribution. This family of generative models is represented by continuous normalizing flows (CNF) (Chen et al., 2018, Grathwohl et al., 2019) and their variants (Gao et al., 2019, Rozen et al., 2021, Gao et al., 2022, Lipman et al., 2023, Neklyudov et al., 2023, Albergo and Vanden-Eijnden, 2023). The major challenges faced by flow-based models revolve around two key questions:

Q1. During the training phase, how can we estimate the velocity field of the probability flow ODE?

Q2. During the sampling phase, how can we solve the probability flow ODE efficiently?

The goal of this section is to propose the characteristic learning that has potential to address the aforementioned questions. In Section 2.1, we derive a probability flow ODE based on the concept of stochastic interpolants and the method of characteristics. Subsequently, in Section 2.2, we propose a velocity matching approach using least-squares regression, which provides an efficient solution to **Q1**. To tackle **Q2**, we first solve the probability flow ODE numerically in Section 2.3. Then Section 2.4 introduces a regression problem to fit characteristics using the obtained numerical solutions. This leads to an efficient simulation-free sampling method for flow-based generative models. Additionally, the characteristic fitting is improved by incorporating a semi-group penalty strategy. Finally, we summarizes the training and sampling algorithms in Section 2.4.

2.1. **Characteristics and Probability Flow ODE.** In this work, we follow the framework of stochastic interpolant (Albergo and Vanden-Eijnden, 2023, Albergo et al., 2023b,a). Let $X_0$ and $X_1$ be two independent random variables drawn from $\mu_0$ and $\mu_1$, respectively. The spatially linear stochastic interpolant $X_t$ is the stochastic process defined as

$$(2.1) \qquad X_t = \alpha(t)X_0 + \beta(t)X_1, \quad t \in (0,1),$$

where $\alpha(t)$ and $\beta(t)$ are two scalar-valued functions satisfying the following condition.

**Condition 1.** The coefficient functions $\alpha(t), \beta(t) \in C([0,1])$ satisfy

   (i) $\alpha(0) = \beta(1) = 1$ and $\alpha(1) = \beta(0) = 0$,
   (ii) $\alpha^2(t) + \beta^2(t) > 0$ for each $t \in [0,1]$,
   (iii) $\alpha(t)$ and $-\beta(t)$ are monotonically decreasing, and
   (iv) $\dot{\alpha}(t), \ddot{\alpha}(t) \in C([0,1))$, $\dot{\alpha}(t)\alpha(t) \in C^1([0,1])$ and $\dot{\beta}(t), \ddot{\beta}(t) \in C([0,1])$.

In this paper, we focus on two examples shown in Table 3: linear interpolants and Föllmer flow. Both of them are widely used in generative learning, such as Nichol and Dhariwal (2021), Liu et al. (2022), Albergo and Vanden-Eijnden (2023), Albergo et al. (2023a), Lipman et al. (2023), Chang et al. (2024).

TABLE 3. Two examples of spatially linear interpolant.

| | $\alpha(t)$ | $\beta(t)$ |
|---|---|---|
| Linear interpolants (Liu et al., 2022, Lipman et al., 2023) | $1-t$ | $t$ |
| Föllmer flow (Chang et al., 2024, Jiao et al., 2024) | $\sqrt{1-t^2}$ | $t$ |

Denote by $\mu_t$ the distribution of the process $X_t$ for each $t \in (0,1)$. The following proposition demonstrates that $\mu_t$ has a density $\rho_t$ that interpolates between $\rho_0$ and $\rho_1$. Further, the density $\rho_t$ satisfies the continuity equation.

**Proposition 2.1** (Transport equation). *The distribution of the stochastic interpolant $X_t$ has a density function $\rho(t,x)$ satisfying $\rho(0,x) = \rho_0(x)$, $\rho(1,x) = \rho_1(x)$, and*

$$\rho(t,x) = \frac{1}{\beta_t}\int_{\mathbb{R}^d}\rho_0(x_0)\rho_1\Big(\frac{x-\alpha_t x_0}{\beta_t}\Big)\,\mathrm{d}x_0 = \frac{1}{\alpha_t}\int_{\mathbb{R}^d}\rho_0\Big(\frac{x-\beta_t x_1}{\alpha_t}\Big)\rho_1(x_1)\,\mathrm{d}x_1,$$

*for each time $t \in (0,1)$. Further, the density $\rho(t,x)$ solves the linear transport equation*

$$(2.2) \qquad \partial_t\rho(t,x) + \nabla\cdot(b^*(t,x)\rho(t,x)) = 0,$$

*where the velocity field is defined as*

$$(2.3) \qquad b^*(t,x) = \mathbb{E}\big[\dot{\alpha}_t X_0 + \dot{\beta}_t X_1 | X_t = x\big].$$

As our primary objective is to generate samples that obey the target distribution, we now consider the transport equation (2.2) from the lens of particles. It suffices to consider characteristics (Courant and Hilbert., 1989, Section II.2), along which the transport equation becomes an ODE:

$$(2.4) \qquad \mathrm{d}x(t) = b^*(t,x(t))\,\mathrm{d}t,$$

where $x(t)$ is knwon as the characteristic curve of the transport equation (2.2), representing the position of the particle at time $t \in (0,1)$, and $b^*$ is the associated velocity field that moves particles around. This characteristic ODE (Evans, 2010, Section 3.2) is known as the probability flow ODE (Song et al., 2021c).

Further, we define the associated two-parameter continuous normalizing flow $g^*_{t,s}$, which transports particles along the characteristic curves:

$$g^*_{t,s} : \mathbb{R}^d \to \mathbb{R}^d, \quad x_t \mapsto x_s, \quad 0 \le t \le s \le 1.$$

Here $x_s = x(s)$ represents the solution of (2.4) at time $s$ given $x(t) = x_t$. Notice that the flow $g^*_{t,s}$ pushes the distribution $\mu_t$ onto $\mu_s$, that is,

$$(g^*_{t,s})_\sharp \mu_t = \mu_s, \quad 0 \le t \le s \le 1.$$

It is evident that the flow $g^*_{t,s}$ satisfies the semigroup property as follows.

**Proposition 2.2** (Semigroup property). *For each $x \in \mathbb{R}^d$, it holds that*

  (i) $g^*_{t,t}(x) - x = 0$ *for each $0 \le t \le 1$, and*
  (ii) $g^*_{t,s}(x) = g^*_{r,s} \circ g^*_{t,r}(x)$ *for each $0 \le t \le r \le s \le 1$.*

Observe that the flow map $g^*_{0,1}$ satisfies the normalizing equation (1.1), and for each fixed $x_t$, $\{g_{t,s}(x_t)\}_{s \ge t}$ is a part of the characteristic curve. Consequently, the generative learning can be reduced to the problem of fitting the characteristic curves by minimizing the following quadratic risk:

$$(2.5) \qquad \mathcal{R}(g) = \frac{2}{T^2} \int_0^T \int_t^T \mathbb{E}_{Z_0 \sim \mu_0} \left[ \|Z_s - g(t, s, Z_t)\|_2^2 \right] \mathrm{d}s \mathrm{d}t,$$

where $Z_t = g^*_{0,t}(Z_0)$, $Z_s = g^*_{0,s}(Z_0)$, and $T \in (1/2, 1)$ is a pre-specified early-stopping time. The primary goal of our method, "characteristic learning", is dedicated to estimating the flow map $g^*_{t,s}$ specified by the characteristic curves.

Given that the distributions of $Z_t$ and $Z_s$ in (2.5) are unknown, it is necessary to estimate them prior to minimizing (2.5). To achieve this, the velocity field is initially estimated in Section 2.2, followed by the utilization of the Euler method to numerically solve the probability flow ODE in Section 2.3. The resulting numerical solutions provide approximations of $(Z_t, Z_s)$, which are then utilized to approximate the population risk (2.5) in Section 2.4.

2.2. **Velocity Matching.** According to (2.3), for each fixed stopping time $T \in (1/2, 1)$, the velocity field $b^*$ is the minimizer of following functional

$$(2.6) \qquad \mathcal{L}(b) = \frac{1}{T} \int_0^T \mathbb{E}_{(X_0, X_1)} \left[ \|\dot{\alpha}(t) X_0 + \dot{\beta}(t) X_1 - b(t, X_t)\|_2^2 \right] \mathrm{d}t,$$

where $X_t$ is the stochastic interpolant defined as (2.1).

Let $\{X_0^{(i)}\}_{i=1}^n$ and $\{X_1^{(i)}\}_{i=1}^n$ be two sets of independent copies of $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$, respectively. Additionally, let $\{t^{(i)}\}_{i=1}^n$ be a set of $n$ i.i.d. random variables drawn from the

uniform distribution on $[0, T]$. Denote by $\mathcal{S} = \{(t^{(i)}, X_0^{(i)}, X_1^{(i)})\}_{i=1}^n$. Then the empirical risk associated with (2.6) is defined as

$$(2.7) \qquad \widehat{\mathcal{L}}_n(b) = \frac{1}{n} \sum_{i=1}^n \|\dot{\alpha}(t^{(i)}) X_0^{(i)} + \dot{\beta}(t^{(i)}) X_1^{(i)} - b(t^{(i)}, X_t^{(i)})\|_2^2,$$

where $X_t^{(i)} = \alpha(t^{(i)}) X_0^{(i)} + \beta(t^{(i)}) X_1^{(i)}$. This induces the empirical risk minimizer

$$(2.8) \qquad \widehat{b} \in \arg\min_{b \in \mathscr{B}} \widehat{\mathcal{L}}_n(b),$$

where $\mathscr{B}$ is a vector-valued deep neural network class. The detailed velocity matching algorithm is shown in Algorithm 1. This approach for velocity matching is also used by rectified flow (Liu et al., 2022) and flow matching (Lipman et al., 2023).

---

**Algorithm 1** Velocity matching.

---

**Input:** Observations $\{X_1^{(i)}\}_{i=1}^n \sim^{\text{i.i.d.}} \mu_1$.
 1: Drawn $\{X_0^{(i)}\}_{i=1}^n \sim^{\text{i.i.d.}} \mu_0 = N(0, I_d)$.
 2: Drawn $\{t^{(i)}\}_{i=1}^n \sim^{\text{i.i.d.}} \text{Unif}[0, T]$.
 3: Construct empirical interpolants via (2.1).
 4: Initialize the deep neural network $b_\theta : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.
 5: **repeat**
 6:     Compute the empirical risk $\widehat{\mathcal{L}}_n(b_\theta)$ in (2.7).
 7:     Compute the gradient $\nabla_\theta \widehat{\mathcal{L}}_n(b_\theta)$.
 8:     Gradient descent update $\theta \leftarrow \theta - \alpha \nabla_\theta \widehat{\mathcal{L}}_n(b_\theta)$.
 9: **until** converged
**Output:** An estimator $\widehat{b} = b_\theta$ of the velocity field.

---

2.3. **Euler Sampling.** We turn to focus on sampling from the estimated probability flow equation in this section. We replace the velocity $b^*$ in probability flow ODE (2.4) by its estimated counterpart $\widehat{b}$ defined as (2.8), and employ the forward Euler method (Iserles, 2008, Butcher, 2016) to discretize this ODE. Denote by $K \in \mathbb{N}_+$ the total number of steps, then the step size is given as $\tau = T/K$. Define $\{t_k = k\tau\}_{k=0}^K$ as the set of time points. Applying forward Euler method deduces the following scheme:

$$(2.9) \qquad x_k = x_{k-1} + \tau \widehat{b}(t_{k-1}, x_{k-1}), \quad 1 \le k \le K.$$

Similar to the flow $g_{t,s}^*$ associated with ODE (2.4), the flow induced by Euler method (2.9) is defined as $\widehat{E}_{k,\ell}^\tau : x_k \mapsto x_\ell$ for each $0 \le k \le \ell \le K$. The following proposition demonstrates that the Euler flow $\widehat{E}_{k,\ell}^\tau$ inherits the semi-group property of continuous flow $g_{t,s}^*$ in Proposition 2.2.

**Proposition 2.3** (Semi-group property). *For each $x \in \mathbb{R}^d$, it holds that*

    *(i) $\widehat{E}_{k,k}^\tau(x) - x = 0$ for each $0 \le k \le \ell \le K$, and*
    *(ii) $\widehat{E}_{k,\ell}^\tau(x) = \widehat{E}_{j,\ell}^\tau \circ \widehat{E}_{k,j}^\tau(x)$ for each $0 \le k \le j \le \ell \le K$.*

It is true that the Euler flow $\widehat{E}_{k,\ell}^{\tau}$ pushes the distribution $\mu_{k\tau}$ approximately onto the distribution $\mu_{\ell\tau}$ for each $1 \leq k \leq \ell \leq K$. Hence, the Euler flow $\widehat{E}_{k,\ell}^{\tau}$ is an alternative approach for evaluating the flow $g_{t,s}^{*}$. However, it is important to note that the Euler sampling incurs a substantial computational cost, as it necessitates a large number of velocity network evaluations. This makes the Euler sampling encounter challenges in time-sensitive application scenarios. Therefore, there is a pressing need to develop an efficient simulation-free approach for evaluating the flow $g_{t,s}^{*}$.

2.4. **Characteristic Fitting and Semi-Group Penalty.** In order to estimate the flow $g_{t,s}^{*}$ via a simulation-free approach, we leverage a deep neural network to fit characteristics using data samples obtained by Euler flow. Let $\mathcal{Z} = \{\widehat{Z}_{0}^{(i)}\}_{i=1}^{m}$ be a set of $m$ random copies of $\widehat{Z}_{0} \sim \mu_{0}$. Further, one obtains $m$ discrete characteristics $\{(\widehat{Z}_{k}^{(i)})_{k=0}^{K}\}_{i=1}^{m}$ by the Euler method (2.9) with $\widehat{Z}_{k}^{(i)} = \widehat{E}_{0,k}^{\tau}(\widehat{Z}_{0}^{(i)})$. Then the empirical counterpart of (2.5) is defined as

$$(2.10) \quad \widehat{\mathcal{R}}_{T,m,K}^{\text{Euler}}(g) = \frac{2}{mK^2} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \left\{ \frac{1}{2} \|\widehat{Z}_{k}^{(i)} - g(t_k, t_k, \widehat{Z}_{k}^{(i)})\|_2^2 + \sum_{\ell=k+1}^{K-1} \|\widehat{Z}_{\ell}^{(i)} - g(t_k, t_\ell, \widehat{Z}_{k}^{(i)})\|_2^2 \right\}.$$

The characteristic generator can be obtained by the empirical risk minimization

$$(2.11) \qquad \widehat{g} \in \underset{g \in \mathscr{G}}{\arg\min} \, \widehat{\mathcal{R}}_{T,m,K}^{\text{Euler}}(g),$$

where $\mathscr{G}$ is a set of vector-valued deep neural networks. It is important to note that the estimator $\widehat{g}_{t,s}$ serves as a neural network approximation for the flow $g_{t,s}^{*}$, eliminating the need to simulate the ODE when evaluating $\widehat{g}_{t,s}$. Therefore, the characteristic generator $\widehat{g}_{t,s}$ is an efficient alternative to the Euler flow $\widehat{E}_{k,\ell}^{\tau}$ in (2.9). Additionally, the idea of fitting characteristics using deep neural network is also used by previous literature, such as DSNO (Zheng et al., 2023a) and CTM (Kim et al., 2024).

Despite that both the continuous flow $g_{t,s}^{*}$ and Euler flow $\widehat{E}_{k,\ell}^{\tau}$ satisfy semigroup property (Propositions 2.2 and 2.3), the characteristic generator $\widehat{g}$ defined as (2.11) does not satisfy the semi-group property in general. In order to ensure the long-term stability of the estimator, we introduce the semi-group constraint, which requires

$$\Delta_{kj\ell}(g) = \|g_{k\tau,\ell\tau}(\widehat{Z}_{k}^{(i)}) - g_{j\tau,\ell\tau} \circ \widehat{E}_{k,j}^{\tau}(\widehat{Z}_{k}^{(i)})\|_2$$

to be as small as possible for each $0 \leq k \leq j \leq \ell \leq K$. Consequently, we propose the semi-group-penalized risk

$$(2.12) \qquad \widehat{\mathcal{R}}_{m,K}^{\text{Euler},\lambda}(g) = \widehat{\mathcal{R}}_{m,K}^{\text{Euler}}(g) + \lambda \widehat{\mathscr{P}}(g),$$

where $\lambda > 0$ is the penalty parameter, and the semigroup-penalty is defined as

$$\widehat{\mathscr{P}}(g) = \frac{2}{mK^2} \sum_{i=1}^{m} \left\{ \sum_{k=0}^{K-1} \sum_{j=k+1}^{K-1} \sum_{\ell=j+1}^{K-1} \Delta_{kj\ell}^{2}(g) \right\}.$$

The complete training and sampling procedures of the characteristic generator are summarized in Algorithms 2, 3 and 4.

As shown in Algorithm 3, in the sampling phase, one only needs to evaluate the characteristic generator $\widehat{g}_{0,T}$ once. As a consequence, our generator diminishes the sampling time

---

**Algorithm 2** Training procedure of characteristic generator.

---

**Input:** Velocity estimator $\widehat{b}$.

1: # Euler sampling
2: **for** $i = 1, \ldots, m$ **do**
3:     Sample initial value $\widehat{Z}_0^{(i)} \sim \mu_0 = N(0, I_d)$.
4:     **for** $k = 1, \ldots, K$ **do**
5:         $\widehat{Z}_k^{(i)} \leftarrow \widehat{E}_{k-1,k}^{\tau}(\widehat{Z}_{k-1}^{(i)})$ by (2.9).
6:     **end for**
7: **end for**
8: # Characteristic fitting
9: Initialize the neural network $g_\phi : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.
10: **repeat**
11:     Compute the empirical risk $\widehat{\mathcal{R}}_{m,K,n}(g_\phi)$ in (2.10) or (2.12).
12:     Compute the gradient $\nabla_\phi \widehat{\mathcal{R}}_{m,K,n}(g_\phi)$.
13:     Gradient descent update $\phi \leftarrow \phi - \alpha \nabla_\phi \widehat{\mathcal{R}}_{m,K,n}(g_\phi)$.
14: **until** converged

**Output:** Characteristic generator $\widehat{g}_{t,s} : \mathbb{R}^d \to \mathbb{R}^d$ for each $0 \leq t \leq s < T$.

---

---

**Algorithm 3** One-step sampling of characteristic generator.

---

**Input:** Characteristic generator $\widehat{g}_{t,s} : \mathbb{R}^d \to \mathbb{R}^d$ for each $0 \leq t \leq s < T$.

1: Sample initial value $\widehat{Z}_0 \sim \mu_0 = N(0, I_d)$.
2: $\widehat{Z}_T \leftarrow \widehat{g}_{0,T}(\widehat{Z}_0)$.

**Output:** Generated samples $\widehat{Z}_T$.

---

---

**Algorithm 4** Fine-grained sampling of characteristic generator.

---

**Input:** Characteristic generator $\widehat{g}_{t,s} : \mathbb{R}^d \to \mathbb{R}^d$ for each $0 \leq t \leq s < T$.

1: Sample initial value $\widehat{Z}_0 \sim \mu_0 = N(0, I_d)$.
2: Choose a sequence of time points $0 = t_0 < \ldots < t_K = T$.
3: **for** $k = 1, \ldots, K$ **do**
4:     $\widehat{Z}_T \leftarrow \widehat{g}_{t_{k-1},t_k}(\widehat{Z}_{k-1})$.
5: **end for**

**Output:** Generated samples $\widehat{Z}_T$.

---

in comparison to Euler sampling. However, it necessitates a significant number of ODE simulations during the training phase as Algorithm 2. Nevertheless, the benefits outweigh the costs. In practical application scenarios, one can simulate ODE (2.9) and fit the probability flow (2.11) (Algorithm 2) on extensive and high-performance computing platforms. Consequently, the derived estimators $\widehat{g}_{0,T}$ can be deployed in computationally constrained and time-sensitive application scenarios.

The characteristic generator is not restricted to one-step generation as Algorithm 3. On the contrary, it can be employed iteratively to produce refined sampling algorithms, thereby enhancing the quality of generation, albeit with a slight increase in computational cost. Algorithm 4 presents the comprehensive procedure for achieving this fine-grained sampling.

## 3. Convergence Rates Analysis

In this section, we present a comprehensive convergence rate analysis for the characteristic generator. We begin by illustrating Assumptions 1 and 2 in Section 3.1, and then establish regularity properties of the probability flow ODE (2.4) in Section 3.2. In Section 3.3 and Section 3.4, we propose a non-asymptotic error analysis for velocity matching and Euler sampling, respectively. In Section 3.5, a convergence rate analysis for the characteristic generator is established. Finally, in Section 3.6, we apply the aforementioned analysis to two widely-used types of probability flow ODE: linear interpolant and Föllmer flow.

3.1. **Discussions of Assumptions.** Assumption 1 is standard and commonly-used in flow-based generative models (Liu et al., 2022, Lipman et al., 2023, Albergo and Vanden-Eijnden, 2023). The following proposition demonstrates that the velocity field $b^*(t, x)$ is a spatial linear combination of $x$ and the score function $s^*(t, x) = \nabla \log \rho_t(x)$. This connection suggests that stochastic interpolants that satisfy Assumption 1 are closely related to score-based diffusion models (Ho et al., 2020, Song et al., 2021c).

**Proposition 3.1** (Velocity and score)**.** *Suppose Assumption 1 holds. Then the following equality holds:*

$$b^*(t, x) = \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t^2\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)s^*(t, x), \quad (t, x) \in (0, 1) \times \mathbb{R}^d.$$

Assumption 2 necessitates that the target distribution be a Gaussian convolution of a distribution with compact support. Previous research (Lee et al., 2023, Oko et al., 2023, Chen et al., 2023a,d, Beyler and Bach, 2025) has investigated the assumption of compact support in the target distribution. However, the probability flow that pushes a Gaussian prior distribution towards a target distribution with compact support may lack regularity because the velocity field is not uniformly Lipschitz in general. Therefore, it becomes crucial to impose additional intractable regularity properties on the probability flow ODE. In contrast, the probability flow towards a distribution of compact support with Gaussian smoothing exhibits high regularity even without any additional assumptions. The regularity properties of the probability flow ODE are established in Section 3.2. Furthermore, Assumption 2 covers a wide range of target distributions, including Gaussian mixtures (Grenioux et al., 2024, Appendix C). In essence, our generative model is capable of learning arbitrarily complex distributions with compact support, provided they are smoothed via convolution with a Gaussian distribution of sufficiently small variance. This flexibility ensures that the model can be effectively applied to a wide range of generative tasks.

3.2. **Properties of Probability Flow ODE.** In this section, we present elementary properties of the probability flow ODE (2.4). To begin with, the following proposition shows that the velocity fields is local bounded.

**Proposition 3.2** (Local bounded velocity). *Suppose Assumptions 1 and 2 hold. Let $R \in (1, +\infty)$. Then it follows that*

$$\max_{1 \leq k \leq d} |b_k^*(t, x)| \leq B_{\text{vel}} R, \quad 0 \leq t \leq 1, \ x \in \mathbb{B}_R^\infty,$$

*where the constant $B_{\text{vel}}$ only depends on $d$ and $\sigma$.*

With the aid of Proposition 3.2, we show that the probability flow and its time derivatives are also local bounded.

**Corollary 3.3** (Local bounded flow). *Suppose Assumptions 1 and 2 hold. Let $R \in (1, +\infty)$. Then it follows that*

$$\max_{1 \leq k \leq d} |g_k^*(t, s, x)| \leq B_{\text{flow}} R, \quad 0 \leq t \leq s \leq 1, \ x \in \mathbb{B}_R^\infty,$$

*where the constant $B_{\text{flow}}$ only depends on $d$ and $\sigma$.*

**Corollary 3.4** (Local bounded time derivatives of flow). *Suppose Assumptions 1 and 2 hold. Let $R \in (1, +\infty)$. Then it follows that*

$$\max \left\{ \|\partial_t g^*(t, s, x)\|_2, \|\partial_s g^*(t, s, x)\|_2 \right\} \leq B_{\text{flow}}' R, \quad 0 \leq t \leq s \leq 1, \ x \in \mathbb{B}_R^\infty.$$

*where the constant $B_{\text{flow}}'$ only depends on $d$ and $\sigma$.*

Further, the spatial gradient of the velocity field is uniformly bounded in spectral norm, as demonstrated by the next proposition.

**Proposition 3.5** (Bounded spatial gradient of velocity). *Suppose Assumptions 1 and 2 hold. Then there exists a constant $G < \infty$ such that*

$$\|\nabla b^*(t, x)\|_{\text{op}} \leq G, \quad 0 \leq t \leq 1, \ x \in \mathbb{B}_R^\infty,$$

*where the constant $G$ only depends on $d$ and $\sigma$.*

A direct consequence of this proposition is that, under Assumptions 1 and 2, the velocity field is uniformly Lipschitz continuous with respect to the spatial variable, as shown by the following corollary.

**Corollary 3.6** (Lipschitz continuity). *Suppose Assumptions 1 and 2 hold. Then*

$$\|b^*(t, x) - b^*(t, x')\|_2 \leq G \|x - x'\|_2, \quad 0 \leq t \leq 1, \ (x, x') \in \mathbb{R}^d \times \mathbb{R}^d.$$

The uniform Lipschitz continuity of the velocity field or score plays a crucial role in controlling the error accumulation along the ODE or Euler method, as detailed in Corollary 3.11 and Theorem 3.12. Previous studies have often made the direct assumption of Lipschitz continuity for the velocity field or score at each time $t \in (0, 1)$ (Chen et al., 2023d,a,c,b, Benton et al., 2024b, Gao and Zhu, 2024, Fukumizu et al., 2025). However, this assumption

appears to be restrictive, as it is difficult to verify. On the contrary, Assumptions 1 and 2 are easily satisfied and cover a large amount of generative tasks.

Another important consequence of Proposition 3.5 is shown as follows.

**Corollary 3.7** (Bounded spatial gradient of flow)**.** *Suppose Assumptions 1 and 2 hold. Then*

$$\|\nabla g^*(t,s,x)\|_{\mathrm{op}} \leq \exp(G(s-t)), \quad 0 \leq t \leq s \leq 1, \ x \in \mathbb{R}^d.$$

Finally, we state the bound of time derivative of the velocity in the following proposition.

**Proposition 3.8** (Bounded time derivative of velocity)**.** *Suppose Assumptions 1 and 2 hold. Let* $T \in (1/2, 1)$ *and* $R \in (1, +\infty)$*. Then it follows that*

$$\|\partial_t b^*(t,x)\|_2 \leq D\kappa(T)R, \quad \kappa(T) := \sup_{t \in [0,T]} \left( \frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \right), \quad (t,x) \in [0,T] \times \mathbb{B}_R^\infty,$$

*where the constant* $D$ *only depends on* $d$ *and* $\sigma$*.*

Proposition 3.8 establishes the Lipschitz continuity of the velocity in time, a crucial requirement for controlling the discretization error of the Euler method. For detailed illustrations, refer to Theorem 3.12. In contrast, previous work (Gao and Zhu, 2024, Assumption 2) simply assumes the Lipschitz continuity of the score in time.

*Remark* 3.9. Proposition 3.8 reveals a critical property of the velocity field: its time derivative, $\|\partial_t b(t,x)\|_2$, becomes unbounded as $t \to 1$. This behavior necessitates stopping the ODE integration at an endpoint $T < 1$ due to two major consequences:

  (i) From the perspective of velocity approximation: A function with an exploding derivative is practically impossible for a neural network to approximate accurately near the singularity. This is a well-known challenge in flow-based and score-based modeling, as discussed in the literature (Kim et al., 2022, Pidstrigach, 2022, Duong et al., 2025).

 (ii) From the perspective of discretization: The error guarantees of numerical ODE solvers are contingent on the smoothness of the vector field. As $t \to 1$, the unbounded derivative violates these smoothness conditions, causing the solver's discretization error to become uncontrolled.

3.3. **Analysis for Velocity Matching.** In this section, we focus on the time-averaged $L^2$-error of the velocity estimator $\widehat{b}$ in (2.8), that is,

$$(3.1) \qquad \mathcal{E}_T(\widehat{b}) = \frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t} \left[ \|b^*(t, X_t) - \widehat{b}(t, X_t)\|_2^2 \right] dt.$$

The majority of existing literature on theoretical analysis of diffusion and flow-based generative models commonly assumes that the $L^2$-risk of score or velocity matching is sufficiently small (Lee et al., 2022, 2023, Chen et al., 2023d,c, Benton et al., 2024a,b, Gao and Zhu, 2024, Beyler and Bach, 2025). However, this line of research lacks the ability to quantitatively characterize the convergence rate of velocity matching with respect to the number of samples. Additionally, no prior theoretical guidance for the selection of neural networks is provided

in this literature. To the best of our knowledge, only a limited number of studies have specifically focused on investigating the convergence rates of score matching (Oko et al., 2023, Chen et al., 2023b, Han et al., 2024) and velocity matching (Chang et al., 2024, Gao et al., 2024, Jiao et al., 2024).

In this work, we aim to establish a non-asymptotic error bound for the $L^2$-risk of the estimated velocity. The main result is stated as follows.

**Theorem 3.10** (Convergence rate for velocity matching). *Suppose Assumptions 1 and 2 hold. Let $T \in (1/2, 1)$. Set the hypothesis class $\mathscr{B}$ as a deep neural network class, which is defined as*

$$
\mathscr{B} = \left\{ b \in N(L, S) : \begin{array}{l} \|b(t,x)\|_\infty \leq B_{\mathrm{vel}} \log^{1/2} n, \\ \|\partial_t b(t,x)\|_2 \leq 3D\kappa(T) \log^{1/2} n, \\ \|\nabla b(t,x)\|_{\mathrm{op}} \leq 3G, \ (t,x) \in [0,T] \times \mathbb{R}^d \end{array} \right\},
$$

*where the depth and the number of non-zero weights of the neural network are given, respectively, as $L = C$ and $S = Cn^{\frac{d+1}{d+3}}$. Then the following inequality holds*

$$
\mathbb{E}_{\mathcal{S}}\big[\mathcal{E}_T(\widehat{b})\big] \leq C\kappa^2(T) n^{-\frac{2}{d+3}} \log^2 n,
$$

*where $C$ is a constant only depending on $d$ and $\sigma$.*

The rate of velocity matching in Theorem 3.10 is consistent with the minimax optimal rate $\mathcal{O}(n^{-\frac{2}{d+3}})$ in nonparametric regression (Stone, 1982, Yang and Barron, 1999, Gyorfi et al., 2002, Tsybakov, 2009) given that the target function is Lipschitz continuous. Moreover, our theoretical findings align with convergence rates of nonparametric regression using deep neural networks (Bauer and Kohler, 2019, Nakada and Imaizumi, 2020, Schmidt-Hieber, 2020, Kohler and Langer, 2021, Farrell et al., 2021, Kohler et al., 2022, Jiao et al., 2023a). It is noteworthy that our results improve upon the rate $\mathcal{O}(n^{-\frac{2}{d+5}})$ derived by Chen et al. (2023b).

In Theorem 3.10, the hypothesis class $\mathscr{B}$ is set as a deep neural network class with Lipschitz constraints. The assumption of uniformly Lipschitz continuity of velocity estimator plays a crucial role in controlling the discretization error induced by Euler method. For further details, refer to Theorem 3.12. This assumption is standard in the theoretical analysis for flow-based or diffusion models, as considered by Kwon et al. (2022, Assumption A2) and Chen et al. (2023c, Assumption 3). In practical applications, various techniques for restricting the Lipschitz constant of deep neural networks have been proposed, such as weight clipping (Arjovsky et al., 2017), gradient penalty (Gulrajani et al., 2017), spectral normalization (Miyato et al., 2018), and Lipschitz network (Zhang et al., 2022). In the theoretical perspective, the approximation properties of deep neural network with Lipschitz constraint has been studied by Chen et al. (2022), Huang et al. (2022), Jiao et al. (2023b), Ding et al. (2024). In this work, an approximation error bound for deep neural networks with Lipschitz constraint is established in Section J.

Besides the error of velocity matching itself, one is actually interested in the error of profitability flow with estimated velocity, for which

$$
(3.2) \qquad\qquad\qquad \mathrm{d}\widehat{x}(t) = \widehat{b}(t, \widehat{x}(t)) \, \mathrm{d}t.
$$

Denote by $\widehat{\mu}_T$ the push-forward distribution of $\mu_0$ by ODE (3.2) at time $T$. The following corollary characteristic the 2-Wasserstein distance between $\mu_T$ and $\widehat{\mu}_T$.

**Corollary 3.11.** *Under the same conditions as Theorem 3.10. The 2-Wasserstein between the probability flow (2.8) and the estimated flow (3.2) at the stopping time $T$ is bounded as follows:*

$$\mathbb{E}_{\mathcal{S}}\left[W_2^2(\widehat{\mu}_T, \mu_T)\right] \leq C\kappa^2(T)n^{-\frac{2}{d+3}}\log^2(n),$$

*where the constant $C$ only depends on $d$ and $\sigma$.*

Corollary 3.11 highlights that the 2-Wasserstein error of the estimated flow converges to zero at a rate of $\mathcal{O}(n^{-\frac{2}{d+3}})$, omitting some logarithmic factors. In contrast, the 2-Wasserstein error bounds derived by Benton et al. (2024b, Theorem 1) and Albergo and Vanden-Eijnden (2023, Proposition 3) are under a "black-box" assumption that the velocity matching error is sufficiently small. Therefore, these results can not capture how the error converges to zero as the number of samples increases.

3.4. **Analysis for Euler Sampling.** The main objective of this section is to estimate the 2-Wasserstein error of the Euler flow (2.9). Despite that there has been numerous studies on the sampling error of SDE-based diffusion models (Lee et al., 2022, 2023, Chen et al., 2023d, Benton et al., 2024a), as well as flow-based methods (Chen et al., 2023c, Gao and Zhu, 2024, Li et al., 2024c,a), most of these works assume that the velocity matching error is sufficiently small. Furthermore, there is limited work that integrates the sampling error with the velocity matching error (Chang et al., 2024, Gao et al., 2024, Jiao et al., 2024). To address this gap, we propose the following theorem.

**Theorem 3.12** (Error analysis for Euler flow). *Under the same conditions as Theorem 3.10. Let the number of time stepss for Euler method $K$ be a positive integer. Then the following inequality holds*

$$\mathbb{E}_{\mathcal{S}}\left[W_2^2\big((\widehat{E}_{0,K}^\tau)_\sharp \mu_0, \mu_T\big)\right] \leq C\kappa^2(T)\left\{n^{-\frac{2}{d+3}}\log^2 n + \frac{1}{K^2}\log n\right\},$$

*where the constant $C$ only depends on $d$ and $\sigma$. Further, if $K \geq Cn^{\frac{1}{d+3}}$, then it follows that*

$$\mathbb{E}_{\mathcal{S}}\left[W_2^2\big((\widehat{E}_{0,K}^\tau)_\sharp \mu_0, \mu_T\big)\right] \leq C\kappa^2(T)n^{-\frac{2}{d+3}}\log^2 n.$$

The error bound of the Euler flow (2.9), as derived in Theorem 3.12, can be roughly divided into two main components. The first term arises from velocity matching, aligning with the error bound presented in Corollary 3.11. The second term corresponds to the discretization error introduced by Euler method. Moreover, as the number of time stepss $K$ increases, the error bound in Theorem 3.12 converges to that in Corollary 3.11.

Based on Theorem 3.12, we can derive a 2-Wasserstein error bound between the target distribution and the push-forward distribution of $\mu_0$ by Euler flow $\widehat{E}_{0,K}^\tau$ as follows.

**Corollary 3.13.** *Under the same conditions as Theorem 3.12. Suppose the number of time stepss for Euler method $K$ satisfies $K \geq Cn^{\frac{1}{d+3}}$. Then the following inequality holds*

$$\mathbb{E}_{\mathcal{S}}\left[W_2^2\big((\widehat{E}_{0,K}^\tau)_\sharp \mu_0, \mu_1\big)\right] \leq C\kappa^2(T)n^{-\frac{2}{d+3}}\log^2(n) + 2\max\{\alpha_T^2, (1-\beta_T)^2\}W_2^2(\mu_0, \mu_1),$$

*where the constant $C$ only depends on $d$ and $\sigma$.*

Corollary 3.13 presents a bound for the total error of flow-based generative models. The first term in the error bound corresponds to the 2-Wasserstein error of the Euler flow, as derived in Theorem 3.12. The second term captures the convergence of the interpolant distribution $\mu_T$ to the target distribution $\mu_1$. It is worth noting that as the stopping time $T$ approaches one, the first term tends to infinity, while the second term simultaneously decreases. This trade-off within the error bound highlights the importance of carefully selecting the stopping time $T$ and provides practical guidance for determining it in real-world applications. Since the optimal early-stopping time depends on the specific choice of the interpolant coefficients, $\alpha(t)$ and $\beta(t)$, we derive the optimal early-stopping time $T$ for two important cases: linear interpolant and Föllmer flow in Corollaries 3.16 and 3.17.

3.5. **Analysis for Characteristic Generator.** Despite the empirical success of simulation-free one-step approaches for the efficient sampling of flow-based generative models (Salimans and Ho, 2022, Song et al., 2023, Zheng et al., 2023a, Kim et al., 2024, Ren et al., 2024), the theoretical analysis for these line of methods still remains unclear. In this section, we establish a thorough analysis for the characteristic generator. To the best of our knowledge, this is the first analysis for one-step sampling method.

To measure the error of the characteristic generator, we focus on the time-average squared 2-Wasserstein distance between the distribution associated to the characteristic generator $\widehat{g}$ and the target distribution

$$(3.3) \qquad \mathcal{D}(\widehat{g}) = \frac{2}{T^2} \int_0^T \int_t^T W_2^2\big((\widehat{g}_{t,s})_\sharp \mu_t, \mu_s\big)\,\mathrm{d}s\mathrm{d}t.$$

The main result is stated as follows.

**Theorem 3.14** (Error analysis for characteristic generator)**.** *Under the same conditions as Theorem 3.12. Further, set the hypothesis class $\mathscr{G}$ as a deep neural network class, which is defined as*

$$\mathscr{G} = \left\{ g \in N(L,S) : \begin{array}{l} \|g(t,s,x)\|_\infty \leq B_{\mathrm{flow}} \log^{1/2} m, \\ \|\partial_t g(t,s,x)\|_2, \|\partial_s g(t,s,x)\|_2 \leq 3B'_{\mathrm{flow}} \log^{1/2} m, \\ \|\nabla g(t,s,x)\|_{\mathrm{op}} \leq 3\exp(GT),\ 0 \leq t \leq s \leq T,\ x \in \mathbb{R}^d \end{array} \right\},$$

*where the depth and the number of non-zero weights of the neural network are given, respectively, as $L = C$ and $S = Cm^{\frac{d+2}{d+4}}$. Then it follows that*

$$\mathbb{E}_{\mathbb{S}}\mathbb{E}_{\mathbb{Z}}[\mathcal{D}(\widehat{g})] \leq C\kappa^2(T)\Big\{n^{-\frac{2}{d+3}}\log^2 n + \frac{\log n}{K^2}\Big\} + C\Big\{m^{-\frac{2}{d+4}}\log^2 m + \frac{\log m}{K}\Big\},$$

*where the constant $C$ only depends on $d$ and $\sigma$. Furthermore, if the number of time steps $K$ for Euler method and the number of samples $m$ for characteristic fitting satisfy*

$$(3.4) \qquad K \geq \max\big\{Cn^{\frac{1}{d+3}}, C\kappa^{-2}(T)n^{\frac{2}{d+3}}\big\} \quad and \quad m \geq C\kappa^{-(d+4)}(T)n^{\frac{d+4}{d+3}},$$

*respectively, then the following inequality holds*

$$\mathbb{E}_{\mathbb{S}}\mathbb{E}_{\mathbb{Z}}[\mathcal{D}(\widehat{g})] \leq C\kappa^2(T)n^{-\frac{2}{d+3}}\log^2 n.$$

In contrast to the error bound of Euler sampling in Theorem 3.12, the error bound of the characteristic generator in Theorem 3.14 incorporates an additional error term $\mathcal{O}(m^{-\frac{2}{d+4}})$. This error term corresponds to the error of the standard nonparametric regression for characteristic fitting, attaining the minimax optimality (Stone, 1982, Yang and Barron, 1999, Gyorfi et al., 2002, Tsybakov, 2009) given that $g^*$ is Lipschitz continuous.

It is noteworthy that the number of samples $m$ for characteristic fitting can be arbitrarily large because training samples can be generated from copies of $Z_0 \sim \mu_0$ using Euler sampling (2.9). Without loss of generality, we consider the case where $m \gg n$. In this scenario, the error bound in Theorem 3.14 aligns with the convergence rate in Theorem 3.12.

In the context of distillation, Euler flow (2.9) is commonly referred as the "teacher" model. Theorem 3.14 guarantees that, when the number of teacher samples is sufficiently large, the characteristic generator can generate new samples that are as good as those generated by the teacher model. However, the theorem also highlights that the teacher model serves as a bottleneck for the characteristic generator, as it cannot surpass the generative quality of the teacher model. These theoretical findings align with empirical observations reported in prior studies (Salimans and Ho, 2022, Song et al., 2023, Kim et al., 2024). One potential approach to overcome this bottleneck is by combining these one-step generative models with GANs, as demonstrated by Lu et al. (2023b), Kim et al. (2024).

*Remark* 3.15. The error bound in Theorem 3.14 is an average-in-time guarantee, not a uniform one. This is a direct consequence of our least-squares objective (2.5), which is designed to minimize the error averaged over all time pairs $(t, s)$. While this formulation is effective for learning the overall map, it naturally leads to a convergence guarantee in an averaged sense. However, it does not imply a stronger, uniform guarantee for a specific time pair. Some potential approaches to obtain a uniform convergence guarantee will be discussed in Section 6.

3.6. **Applications.** In this section, the theoretical analysis is applied to two types of flow-based method in Table 3: linear interpolant and Föllmer flow. The convergence rates of them are shown in Corollaries 3.16 and 3.17, respectively.

**Corollary 3.16** (Convergence rate of linear interpolant). *Suppose Assumptions 1 and 2 hold. Set the velocity matching network class $\mathscr{B}$ to be Lipschitz-controlled, with depth $L_b = C$ and number of non-zero weights $S_b = C n^{\frac{d+1}{d+3}}$. Set the stopping time $T$ as*

$$T = 1 - C n^{-\frac{1}{3(d+3)}} \log^{\frac{1}{2}} n.$$

*Suppose the number of time step $K$ satisfies $K \geq C n^{\frac{1}{d+3}}$. Then it follows that*

$$\mathbb{E}_{\mathcal{S}}\left[ W_2^2\left( (\widehat{E}_{0,K}^\tau)_\sharp \mu_0, \mu_1 \right) \right] \leq C n^{-\frac{2}{3(d+3)}} \log n.$$

*In addition, set the characteristic network class to be Lipschitz-controlled, with depth $L_g = C$ and number of non-zero weights $S_g = C m^{\frac{d+2}{d+4}}$. Further, let*

$$m \geq C n^{\frac{d+4}{3(d+3)}} \log^{d+4} n.$$

*Then it follows that*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{Z}}\big[\mathcal{D}(\widehat{g})\big] \leq C\,n^{-\frac{2}{3(d+3)}}\,\log n.$$

*Here, the constants $C$ may depend on $d$ and $\sigma$, but is independent of $n$.*

**Corollary 3.17** (Convergence rate of Föllmer flow)**.** *Suppose Assumptions 1 and 2 hold. Set the velocity matching network class $\mathscr{B}$ to be Lipschitz-controlled, with depth $L_b = C$ and number of non-zero weights $S_b = Cn^{\frac{d+1}{d+3}}$. Set the stopping time $T$ as*

$$T = 1 - C\,n^{-\frac{2}{5(d+3)}}\,\log^{\frac{3}{5}}n.$$

*Suppose the number of time step $K$ satisfies $K \geq C\,n^{\frac{1}{d+3}}$. Then it follows that*

$$\mathbb{E}_{\mathcal{S}}\Big[W_2^2\big((\widehat{E}_{0,K}^{\tau})_{\sharp}\mu_0, \mu_1\big)\Big] \leq C\,n^{-\frac{2}{5(d+3)}}\,\log^{\frac{3}{5}}n.$$

*In addition, set the characteristic network class to be Lipschitz-controlled, with depth $L_g = C$ and number of non-zero weights $S_g = C\,m^{\frac{d+2}{d+4}}$. Further, let*

$$m \;\geq\; C\,n^{\frac{d+4}{5(d+3)}}\,\log^{\frac{6}{5}(d+4)}n.$$

*Then it follows that*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{Z}}\big[\mathcal{D}(\widehat{g})\big] \leq C\,n^{-\frac{2}{5(d+3)}}\,\log^{\frac{3}{5}}n.$$

*Here, the constants $C$ may depend on $d$ and $\sigma$, but is independent of $n$.*

3.7. **Mitigate the Curse of Dimensionality.** As shown in Sections 3.10 and 3.14, the convergence of both velocity matching and characteristic fitting suffer from the curse of dimensionality: the sample complexity increases exponentially with the ambient dimension $d$ of the data. However, we will show that under Assumption 3, the convergence rate of the characteristic generator overcomes this limitation. The key to this result is the following decomposition lemma.

**Lemma 3.18** (Low-dimensional decomposition)**.** *Suppose Assumptions 1 and 3 hold. For each $(t, s, x)$ with $0 \leq t \leq s \leq 1$ and $x \in \mathbb{R}^d$,*

$$b^*(t, x) \equiv P\,\widetilde{b}^*(t, P^{\top}x) + \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2}\,(I_d - PP^{\top})x,$$

$$g^*(t, s, x) \equiv P\,\widetilde{g}^*(t, s, P^{\top}x) + \sqrt{\frac{\alpha_s^2 + \sigma^2 \beta_s^2}{\alpha_t^2 + \sigma^2 \beta_t^2}}\,(I_d - PP^{\top})x.$$

*The vector field $\widetilde{b}^* : \mathbb{R} \times \mathbb{R}^{d^*} \to \mathbb{R}^{d^*}$ is defined as:*

$$\widetilde{b}^*(t, \widetilde{x}) := \mathbb{E}[\dot{\alpha}_t \widetilde{X}_0 + \dot{\beta}_t \widetilde{X}_1 | \widetilde{X}_t = \widetilde{x}], \quad \widetilde{x} \in \mathbb{R}^{d^*},$$

*where $\widetilde{X}_0 \sim N(0, I_{d^*})$ and $\widetilde{X}_1 \sim N(0, \sigma^2 I_{d^*}) * \widetilde{\nu}$ are independent, and $\widetilde{X}_t := \alpha_t \widetilde{X}_0 + \beta_t \widetilde{X}_1$. Further, the vector field $\widetilde{g}^* : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d^*} \to \mathbb{R}^{d^*}$ is defined as the flow map of $\mathrm{d}\widetilde{x}(t) = \widetilde{b}^*(t, \widetilde{x}(t))\,\mathrm{d}t$.*

This lemma is the cornerstone of our result. It reveals that both $b^*$ and $g^*$ have a crucial low-dimensional structure. They decompose into two orthogonal parts:

  (i) A complex, non-linear component governed by the low-dimensional functions $\widetilde{b}^*$ and $\widetilde{g}^*$, which only act on the $d^*$-dimensional data manifold.

(ii) A simple, linear component that acts on the space orthogonal to the manifold.

This decomposition implies that the learning problem is greatly simplified. Instead of learning a function in the ambient $d$ dimensions, we only need to learn the low-dimensional components $\widetilde{b}^*$ and $\widetilde{g}^*$ in the intrinsic $d$ dimensions. This is precisely why our method's convergence rate depends on $d^*$ instead of $d$, effectively mitigating the curse of dimensionality. The necessary regularity conditions for these low-dimensional functions are established in the appendix.

Now, we formally state the convergence results of linear interpolant and Föllmer flow under manifold hypothesis, which explicitly ease the curse of dimensionality.

**Corollary 3.19** (Convergence of linear interpolant under manifold hypothesis). *Suppose Assumptions 1 and 3 hold. Set the velocity matching network class $\mathscr{B}$ to be Lipschitz-controlled, with depth $L_b = C$ and number of non-zero weights $S_b = Cn^{\frac{d^*+1}{d^*+3}}$. Set the stopping time $T$ as*

$$T = 1 - C\,n^{-\frac{1}{3(d^*+3)}}\,\log^{\frac{1}{2}} n.$$

*Suppose the number of time step $K$ satisfies $K \geq C\,n^{\frac{1}{d^*+3}}$. Then it follows that*

$$\mathbb{E}_\mathcal{S}\Big[W_2^2\big((\widehat{E}_{0,K}^\tau)_\sharp \mu_0, \mu_1\big)\Big] \leq C\,n^{-\frac{2}{3(d^*+3)}}\,\log n.$$

*In addition, set the characteristic network class to be Lipschitz-controlled, with depth $L_g = C$ and number of non-zero weights $S_g = C\,m^{\frac{d^*+2}{d^*+4}}$. Further, let*

$$m \;\geq\; C\,n^{\frac{d^*+4}{3(d^*+3)}}\,\log^{d^*+4} n.$$

*Then it follows that*

$$\mathbb{E}_\mathcal{S}\mathbb{E}_\mathcal{Z}\big[\mathcal{D}(\widehat{g})\big] \leq C\,n^{-\frac{2}{3(d^*+3)}}\,\log n.$$

*Here, the constants $C$ may depend on $d$, $d^*$, and $\sigma$, but is independent of $n$.*

**Corollary 3.20** (Convergence of Föllmer flow under manifold hypothesis). *Suppose Assumptions 1 and 3 hold. Set the velocity matching network class $\mathscr{B}$ to be Lipschitz-controlled, with depth $L_b = C$ and number of non-zero weights $S_b = Cn^{\frac{d^*+1}{d^*+3}}$. Set the stopping time $T$ as*

$$T = 1 - C\,n^{-\frac{2}{5(d^*+3)}}\,\log^{\frac{3}{5}} n.$$

*Suppose the number of time step $K$ satisfies $K \geq C\,n^{\frac{1}{d^*+3}}$. Then it follows that*

$$\mathbb{E}_\mathcal{S}\Big[W_2^2\big((\widehat{E}_{0,K}^\tau)_\sharp \mu_0, \mu_1\big)\Big] \leq C\,n^{-\frac{2}{5(d^*+3)}}\,\log^{\frac{3}{5}} n.$$

*In addition, set the characteristic network class to be Lipschitz-controlled, with depth $L_g = C$ and number of non-zero weights $S_g = C\,m^{\frac{d^*+2}{d^*+4}}$. Further, let*

$$m \;\geq\; C\,n^{\frac{d^*+4}{5(d^*+3)}}\,\log^{\frac{6}{5}(d^*+4)} n.$$

*Then it follows that*

$$\mathbb{E}_\mathcal{S}\mathbb{E}_\mathcal{Z}\big[\mathcal{D}(\widehat{g})\big] \leq C\,n^{-\frac{2}{5(d^*+3)}}\,\log^{\frac{3}{5}} n.$$

*Here, the constants $C$ may depend on $d$, $d^*$, and $\sigma$, but is independent of $n$.*

In contrast to the results in Corollaries 3.16 and 3.17, the convergence rates under the manifold assumption depend on the intrinsic dimension $d^*$ rather than the ambient dimension $d$. This demonstrates that our characteristic generator can mitigate the curse of dimensionality when the data possesses an intrinsic low-dimensional structure.

## 4. NUMERICAL STUDIES

In this section, we delve into the numerical performance of characteristic learning. To begin with, we introduce several technical improvements in Section 4.1. Subsequently, the experimental results and discussions are presented in Section 4.2.

4.1. **Technical improvements.** In this section several technical enhancements to the algorithms in Section 2 are introduced. Empirical evidence indicates that these methods exhibit superior numerical performance. It is noteworthy that improved algorithms in this section are mathematically equivalent to the previous ones in Section 2. Consequently, these improvements remain within the established mathematical framework and theoretical analysis. Specifically, in Section 4.1.1, we adopt a denoiser matching algorithm as a replacement for the velocity matching in Section 2.2. Additionally, in Section 4.1.2, we replace Euler method with the technique of exponential integrator. Finally, the modified characteristic learning algorithm is summarized in Section 4.1.3.

4.1.1. *Denoiser matching.* To begin with, we define the denoiser denoiser $D^*$ as

$$D^*(t, x) = \mathbb{E}\big[X_1 \big| X_t = x\big], \quad (t, x) \in (0, 1) \times \mathbb{R}^d,$$

which recovers $X_1$ from noised observation $X_t = \alpha_t X_0 + \beta_t X_1$. It is apparent that $D^*$ minimizes the following functional for each $T \in (0, 1)$,

$$(4.1) \qquad \mathcal{L}(D) = \frac{1}{T} \int_0^T \mathbb{E}\Big[\|X_1 - D(t, X_t)\|_2^2\Big] \, \mathrm{d}t.$$

An estimator $\widehat{D}$ of the denoiser $D^*$ can be obtained by the empirical risk minimization similar to (2.8) using data set $\mathcal{S} = \{(t^{(i)}, X_0^{(i)}, X_1^{(i)})\}_{i=1}^n$. In the context of distillation for diffusion models, the denoiser network $\widehat{D}$ is referred as the "teacher" network.

By an argument similar to Proposition 3.1, the velocity field is a spatial linear combination of $x$ and denoiser $D^*(t, x)$, that is,

$$(4.2) \qquad b^*(t, x) = \frac{\dot{\alpha}_t}{\alpha_t} x + \beta_t \Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big) D^*(t, x).$$

Thus the denoiser matching is equivalent to estimating the velocity field, but the former has better numerical stability (Karras et al., 2022, Kim et al., 2024). Furthermore, the semi-linear form (4.2) enable us to use the exponential integrator, which is more stable than Euler method. See Section 4.1.2 for detailed discussions.

4.1.2. *Exponential integrator.* With the aid of the denoiser estimator $\widehat{D}$, the estimated probability flow (3.2) is replaced by

$$(4.3) \qquad \frac{\mathrm{d}\widehat{x}(t)}{\mathrm{d}t} = \frac{\dot{\alpha}_t}{\alpha_t}\widehat{x}(t) + \beta_t\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\widehat{D}(t,\widehat{x}(t)), \quad t \in (0,1).$$

Observe that the solution of the semi-linear ODE (4.3) can be exactly formulated by the "variation of constants" formula as

$$(4.4) \qquad \widehat{x}(s) = \Phi(t,s)\widehat{x}(t) + \int_t^s \psi(\tau,s)\widehat{D}(\tau,\widehat{x}(\tau))\,\mathrm{d}\tau,$$

where $\Phi(t,s)$ and $\psi(t,s)$ are defined as

$$\Phi(t,s) = \exp\Big(\int_t^s \frac{\dot{\alpha}_\tau}{\alpha_\tau}\,\mathrm{d}\tau\Big), \quad \psi(t,s) = \Phi(t,s)\beta_t\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big), \quad 0 \le t \le s \le T.$$

By a similar argument to Euler method (2.9), we replace $\widehat{D}(\tau,\widehat{x}(\tau))$ in (4.4) by $\widehat{D}(t,\widehat{x}(t))$ and implies an explicit scheme

$$(4.5) \qquad \widehat{x}(s) \approx \Phi(t,s)\widehat{x}(t) + \Psi(t,s)\widehat{D}(t,\widehat{x}(t)), \quad 0 \le t \le s \le T.$$

where $\Psi(t,s)$ is a integral defined as

$$\Psi(t,s) = \int_t^s \Phi(\tau,s)\beta_\tau\Big(\frac{\dot{\beta}_\tau}{\beta_\tau} - \frac{\dot{\alpha}_\tau}{\alpha_\tau}\Big)\,\mathrm{d}\tau.$$

Notice that the integrals $\Phi$ and $\Psi$ can be computed analytically given the interpolant coefficients $\alpha_t$ and $\beta_t$. The integral scheme (4.5) is commonly referred to as the first-order exponential integrator (Hochbruck and Ostermann, 2010), which has been utilized in sampling of diffusion models by Zhang and Chen (2023), Lu et al. (2022, 2023a), Zheng et al. (2023b). We display the generated images and corresponding FID using exponential integrator (4.5) in Figure 1.
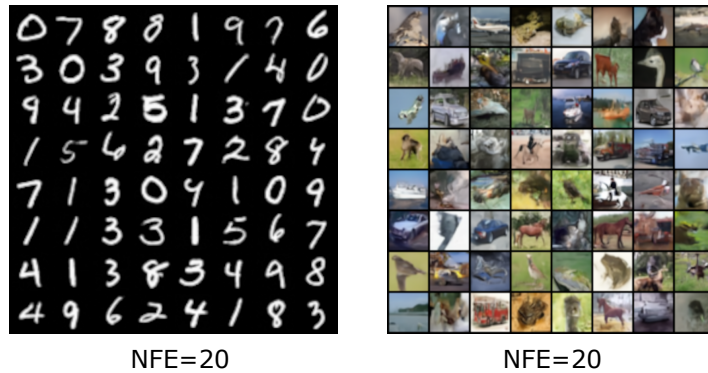


NFE=20      NFE=20

FIGURE 1. Samples generated by probability flow ODE with exponential integrator.

In practice, we find the first-order exponential integrator outperforms other methods, such as Euler and Heun methods. Nevertheless, it is important to note that this integral scheme remains a first-order method, akin to the Euler method, and does not improve the

convergence rate of the discretization error. Consequently, our analysis encompasses this integral scheme as well.

4.1.3. *Characteristic fitting.* In Section 2.4, we directly parameterize the probability flow by a deep neural network. However, in practice training such a neural network is unstable. In this section we present some technical tricks to get a more stable training algorithm by exploiting as much of the structure of the problem as possible without changing its mathematical nature.

Recall the explicit formulation (4.4) of the solution

$$x(s) = \Phi(t,s)x(t) + \Psi(t,s)\frac{\int_t^s \psi(\tau,s)D^*(\tau,x(\tau))\,\mathrm{d}\tau}{\int_t^s \psi(\tau,s)\,\mathrm{d}\tau},$$

where the fraction in the second term can be viewed as a weighted average of $D^*(\tau,x(\tau))$ on $[s,t]$. Our main idea is to approximate this term using a deep neural network $D_{\mathcal{S}}$, which is referred to the student model. Then the probability flow is parameterized by

$$(4.6) \qquad g(t,s,x) = \Phi(t,s)x + \Psi(t,s)D_{\mathcal{S}}(t,s,x), \quad 0 \le t \le s \le T.$$

Since the exact denoiser $D^*$ is unknown, the student network $D_{\mathcal{S}}$ can only be trained from the denoiser estimator $\widehat{D}$. Therefore, $\widehat{D}$ is referred to the teacher model and denoted by $D_{\mathcal{T}} = \widehat{D}$ thereafter.

We next design the objective functional for the student model $D_{\mathcal{S}}$ to utilize as much of the structure of the problem as possible. First, it is apparent that

$$\lim_{s \to t^+} \frac{\int_t^s \psi(\tau,s)D^*(\tau,x(\tau))\,\mathrm{d}\tau}{\int_t^s \psi(\tau,s)\,\mathrm{d}\tau} = D^*(t,x(t)).$$

This allows us to reuse the denoiser matching objective functional (4.1) as the local risk to ensure the local consistency of the student model

$$(4.7) \qquad \mathcal{R}_{\mathrm{loc}}(D_{\mathcal{S}}) = \int_0^T \mathbb{E}\Big[\|X_1 - D_{\mathcal{S}}(t,t,X_t)\|_2^2\Big]\,\mathrm{d}t.$$

On the other hand, the outputs of generator (4.6) are required to align with the numerical solutions (4.5) and satisfy the semi-group properties, as discussed in Section 2.4. This implies the following global risk

$$(4.8) \qquad \mathcal{R}_{\mathrm{glo}}(D_{\mathcal{S}}) = \int_0^T \int_t^T \int_s^T \mathbb{E}\Big[\|g_{s,T}^{\mathrm{off}} \circ g_{u,s} \circ g_{t,u}^{\mathrm{int}}(X_t) - g_{s,T}^{\mathrm{off}} \circ g_{t,s}^{\mathrm{off}}(X_t)\|_2^2\Big]\,\mathrm{d}u\mathrm{d}s\mathrm{d}t,$$

where $g$ is induced by the student model $D_{\mathcal{S}}$ defined as (4.6), $g^{\mathrm{int}}$ denotes the exponential integrator given by teacher model $D_{\mathcal{T}}$, and $g^{\mathrm{off}}$ denotes an offline copy of $g$ for training stability. The population risk (4.8) can be considered as a variant of the original objective functional (2.12), ensuring the long-range consistency of the generator. Combining the short-range denoiser matching risk (4.7) and the long-range characteristic fitting risk (4.8) yields the final training procedure for the characteristic generator. We conclude the practical characteristic learning algorithm in Algorithm 5. The one-step sampling procedure is the same as that in Algorithm 2. For better sampling quality, one can divide the time interval into pieces as Algorithm 4.

---

**Algorithm 5** Practical training procedure of characteristic generator.

---

**Input:** Observations $X_1 \sim \mu_1$, and pre-trained denoiser $D_{\mathcal{T}}$.
1:  Initialize the student neural network $D_{\mathcal{S},\phi} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.
2:  Choose the loss parameter $\lambda$.
3:  **repeat**
4:      # Short-range denoiser matching
5:      Drawn $X_0 \sim \mu_0 = N(0, I_d)$ and $t \sim \mathrm{Unif}[0, T]$.
6:      Construct stochastic interpolant $X_t = \alpha_t X_0 + \beta_t X_1$.
7:      $\widehat{\mathcal{R}}_{\mathrm{loc}}(D_{\mathcal{S},\phi}) \leftarrow \|D_{\mathcal{S},\phi}(t, t, X_t) - X_0\|_2^2$.
8:      # Long-range characteristic matching
9:      Drawn $s \sim \mathrm{Unif}[t, T]$ and $u \sim \mathrm{Unif}[s, T]$.
10:     $\widehat{\mathcal{R}}_{\mathrm{glo}}(D_{\mathcal{S},\phi}) \leftarrow \|g_{s,T}^{\mathrm{off}} \circ g_{u,s} \circ g_{t,u}^{\mathrm{int}}(X_t) - g_{s,T}^{\mathrm{off}} \circ g_{t,s}^{\mathrm{off}}(X_t)\|_2^2$.
11:     # Combined objective functional
12:     Compute the gradient $\nabla_\phi\{\lambda\widehat{\mathcal{R}}_{\mathrm{loc}}(D_{\mathcal{S},\phi}) + \widehat{\mathcal{R}}_{\mathrm{glo}}(D_{\mathcal{S},\phi})\}$.
13:     Gradient descent update $\phi \leftarrow \phi - \alpha\nabla_\phi\{\lambda\widehat{\mathcal{R}}_{\mathrm{loc}}(D_{\mathcal{S},\phi}) + \widehat{\mathcal{R}}_{\mathrm{glo}}(D_{\mathcal{S},\phi})\}$.
14: **until** converged

**Output:** Characteristic generator $\widehat{g}(t, s, x) = \Phi(t, s)x + \Psi(t, s)D_{\mathcal{S},\phi}(t, s, x)$.

---

*Remark* 4.1 (Comparison with CTM (Kim et al., 2024)). Kim et al. (2024) proposed a similar method, but our approach differs from the CTM in two significant aspects. First, the integral scheme $g^{\mathrm{int}}$ employed by CTM is Euler-based. While Euler method coincides with the first-order exponential integrator for VE-ODE (Song et al., 2021c), its numerical stability cannot be guaranteed for general probability flow ODEs. In contrast, the exponential integrator used in our method may ease potential training instability as it fully exploits the semi-linearity of the ODE system. Secondly, CTM relies on GAN training in their approach, borrowing a pre-trained discriminator and treating $g$ as the generator. This reliance on GAN training may cause potential training instability and require extra training of the discriminator. Our method, however, does not require this additional GAN training burden. Detailed comparisons between characteristic generator and CTM can be found in Appendix N.

4.2. **Experiment results and discussions.** In this section, we validate the generation quality and sampling efficiency of the characteristic generator using both synthetic and real data through numerical experiments. Föllmer flow is taken as the underlying ODE, and all results can be generalized to arbitrary probability flow ODE without loss of generality. We use the Fréchet inception distance (FID) to measure sampling quality on image data. Lower FID means better performance. See Appendix L for the list of full hyper-parameters employed in our numerical experiments. In addition, we provide some remarks on the training time consumption of our numerical experiments in Appendix M.

4.2.1. *Synthetic 2-dimensional dataset.* On a 2-dimensional dataset where the target distribution shapes like a Swiss roll, the characteristic generator trained by Algorithm 2 works well.

We display the original dataset and samples generated by Euler method (NFE=100) and the characteristic generator (NFE=1) in Figure 2.
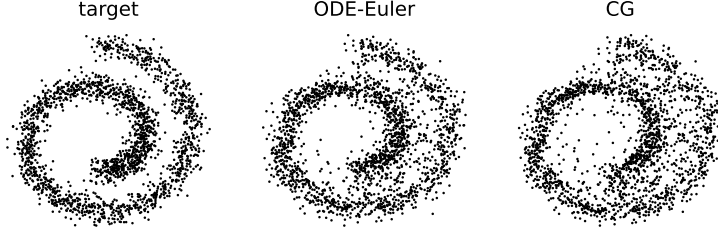


FIGURE 2. Original Swiss roll samples and samples generated by ODE model with Euler method (ODE-Euler) and the characteristic generator (CG).

From Figure 2, it can be observed that the generative quality of the one-step characteristic generator closely resembles that of Euler method with 100 function evaluations (NFE). This indicates that the original training procedure (Algorithm 2) can yield commendable generation outcomes when dealing with relatively simple target distribution.

4.2.2. *Real dataset.* In this section, we apply the characteristic generator to two real dataset: MNIST and CIFAR-10. The characteristic generators are trained by Algorithm 5.

Generated images by the characteristic generator are displayed in Figure 3 (MNIST on top and CIFAR-10 on bottom). The experimental results illustrate that one-step generation has high generation quality, which can be significantly improved by iterating the characteristic generator by a few steps.
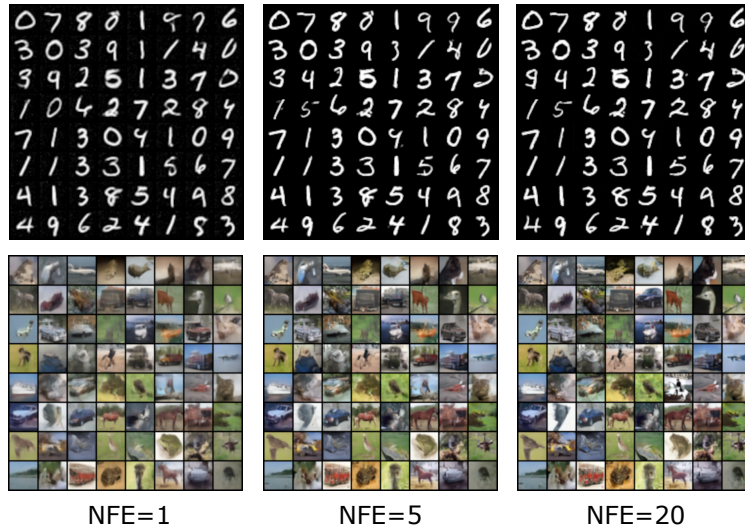


FIGURE 3. Samples generated by the characteristic generator on CIFAR-10.

We compare the images generated by the numerical sampler and characteristic generator with different numbers of function evaluations (NFE) in Figure 4. The numerical sampler

fails to accurately generate images at small NFE values such as 1 and 2. In fact, with 1 NFE, the solution is actually close to the mean of the target distribution. It is necessary to have five or more NFE for the numerical ODE solvers to work properly. In contrast, the characteristic generator is capable of generating high-quality images even with only 1 NFE. In essence, the characteristic generator distills the knowledge of a precise, multi-step solver into a fast, single step network. Our theory correctly predicts that its error should be much lower than that of a naive one-step Euler sampler, which aligns perfectly with our empirical findings in Figure 4. See Appendix O for detailed discusssion.

Furthermore, we compare the convergence of FID by NFE for the characteristic generator in Table 4. It is evident that the characteristic generator converges faster than the numerical sampler and achieves better scores at smaller NFE values.
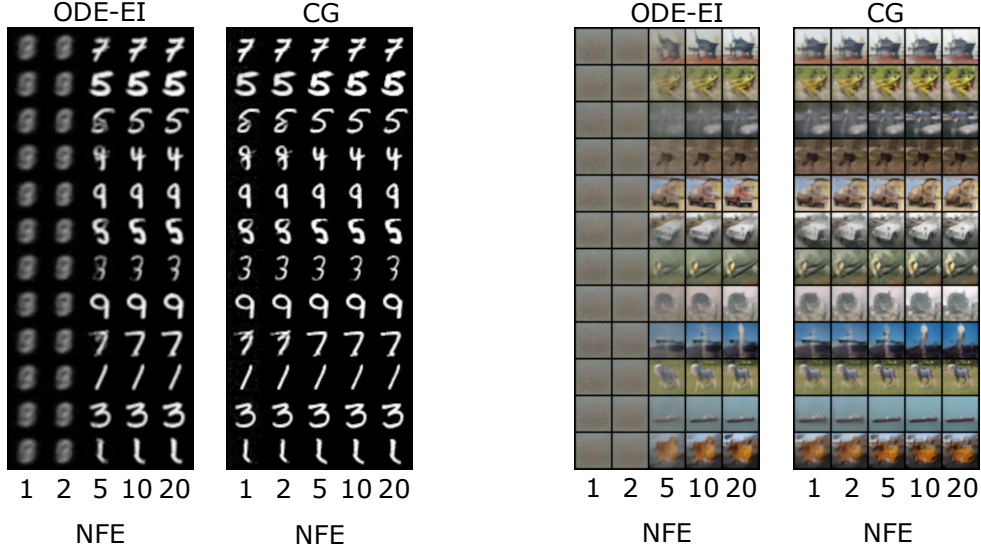


FIGURE 4. Comparison of samples generated by ODE with exponential integrator (ODE-EI) and characteristic generators (CG) under different NFE on MNIST and CIFAR-10.

TABLE 4. Comparison of FID by NFE between the exponential integrator (EI) and characteristic generator (CG) in MNIST and CIFAR-10.

| Dataset | Method | NFE=1 | NFE=2 | NFE=5 | NFE=10 | NFE=20 |
|---------|--------|-------|-------|-------|--------|--------|
| MNIST | ODE-EI | 46.55 | 46.71 | 2.69 | 0.66 | 0.28 |
| MNIST | CG | 1.97 | 1.15 | 0.28 | 0.20 | 0.13 |
| CIFAR-10 | ODE-EI | 14.06 | 15.42 | 5.38 | 3.16 | 2.50 |
| CIFAR-10 | CG | 4.59 | 3.50 | 2.90 | 2.76 | 2.63 |

4.2.3. *Comparison with other generative models.* Table 5 presents the FID on CIFAR-10 achieved by various generative models. Our characteristic generator demonstrates superior generation quality compared to models without GAN, regardless of whether it is one-step or few-step generation. Notably, our proposed method achieves a comparable FID to the state-of-the-art method CTM (Kim et al., 2024), without the requirement of additional GAN training as employed by CTM. Moreover, our proposed characteristic generator with NFE=4 achieves similar or even superior generation performance compared to GAN models.

TABLE 5. Performance comparisons on CIFAR-10.

| Model | NFE ↓ | FID ↓ |
|---|---|---|
| **GAN Models** | | |
| BigGAN (Brock et al., 2019) | 1 | 8.51 |
| StyleGAN-Ada (Karras et al., 2019) | 1 | 2.92 |
| **Diffusion + Sampler** | | |
| DDPM (Ho et al., 2020) | 1000 | 3.17 |
| DDIM (Song et al., 2021a) | 100 | 4.16 |
| Score SDE (Song et al., 2021c) | 2000 | 2.20 |
| EDM (Karras et al., 2022) | 35 | 2.01 |
| **Diffusion + Distillation** | | |
| KD (Luhman and Luhman, 2021) | 1 | 9.36 |
| DFNO (Zheng et al., 2023a) | 1 | 5.92 |
| Rectified Flow (Liu, 2022) | 1 | 4.85 |
| PD (Salimans and Ho, 2022) | 1 | 9.12 |
| CD (Song et al. (2023), retrained by Kim et al. (2024)) | 1 | 10.53 |
| CTM (without GAN) (Kim et al., 2024, Table 3) | 1 | 5.19 |
| CG (ours) | 1 | **4.59** |
| PD (Salimans and Ho, 2022) | 2 | 4.51 |
| CTM (without GAN) (Kim et al., 2024, Table 3) | 18 | 3.00 |
| CG (ours) | 2 | **3.50** |
| CG (ours) | 4 | **2.83** |
| **Diffusion + Distillation + GAN** | | |
| CD (with GAN) (Lu et al., 2023b) | 1 | 2.65 |
| CTM (with GAN) (Kim et al., 2024) | 1 | 1.98 |
| CTM (with GAN) (Kim et al., 2024) | 2 | 1.87 |

4.2.4. *Self-Distillation on higher resolution.* We also explore to apply the proposed method on higher resolution image datasets such as CelebA HQ (down-sampled to $256 \times 256$ and $512 \times 512$), and manage to eliminate the dependence on the teacher model, resulting a self-distillation scheme.

The main idea is to define a teacher-free reference solution $\overline{g}_{t,s}(X_t)$ from a characteristic estimator $\widehat{g}$, by

$$(4.9) \qquad \overline{g}_{t,s}(X_t) = \widehat{g}_{u,s} \circ \widehat{g}_{t,u}(X_t), \ u \in (t,s).$$

With the reference solution $\overline{g}$, we can get rid of the extra teacher network when training the characteristic generator, significantly reducing the required memory. Also, the formation of $\overline{g}$ naturally incorporates the semigroup constraint, as the optimization target is to minimize the distance between $\overline{g}$ and $\widehat{g}$. Notice that we do not specify the choice of $u$ in (4.9) as it should be arbitrary. Recent progress (Frans et al., 2025) finds $u = (t+s)/2$ works well enough and shows some symmetry, and we follow this setup. During the early stage of training, the local consistency loss plays the dominant role as the global counterpart relied on the reference solution, which requires the estimator $\widehat{g}$ fitted on a smaller time scale. During the late stage of training, the local consistency is already ensured, and the global consistency loss gradually takes effect.

We carry out the self-distillation variant on the CelebA HQ dataset (down-sampled to $256 \times 256$ and $512 \times 512$). We carry out the training and generation on the $8\times$ down-sampled latent space using the sd-vae-ft-mse autoencoder (Rombach et al., 2022). We also move from the U-Net architecture to the emerging DiT architectures (Peebles and Xie, 2023) (B/2 on the 256 resolution and XL/2 on the 512 resolution). We report the FID in Table 6 and the generated images in Figure 5.

TABLE 6. Comparison of FID by NFE by the characteristic generator on CelebAHQ-256 and CelebAHQ-512.

| Resolution | NFE=1 | NFE=2 | NFE=4 | NFE=8 |
|---|---|---|---|---|
| 256 | 30.25 | 19.02 | 15.82 | 14.29 |
| 512 | 31.18 | 22.45 | 18.61 | 15.95 |

## 5. RELATED WORKS

5.1. **Fast sampling method for diffusion and flow-based models.** Diffusion and flow-based models have demonstrated impressive generative performance across various applications. However, their iterative sampling process requires a substantial number of evaluations of the score or velocity neural network, which currently limits their real-time application. In recent years, there has been a surge of fast sampling methods aimed at accelerating the sampling process of diffusion or flow-based models.

The sampling process of the diffusion or flow-based model can be considered as numerically solving SDE or ODE. Therefore, one approach to address this issue is to develop acceleration algorithms for these equations (Zhang and Chen, 2023, Lu et al., 2022, 2023a, Zheng et al., 2023b, Gao and Zhu, 2024, Li et al., 2024a). For instance, Lu et al. (2022) effectively utilizes the semi-linear structure of the probability flow ODE by employing an exponential
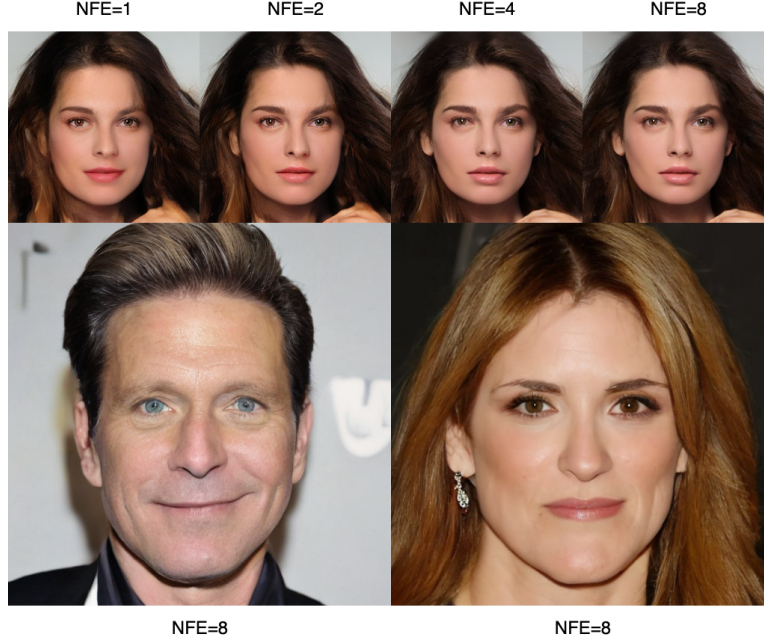
FIGURE 5. Generated images on CelebAHQ-256 (top) and CelebAHQ-512(bottom) (scaled down by a factor of 0.28 to fit the page).

integrator. Furthermore, this algorithm incorporates adaptive step size schedules and high-order approximations. While these strategies can achieve high-quality generation requiring 10-15 neural network evaluations, generating samples in a single step still poses a significant challenge.

There is another line of recent works that aim to propose a simulation-free one-step sampling method. It is important to note that SDE has probabilistic trajectories, while the trajectory of ODE is deterministic, which is known as "self-consistency" (Song et al., 2023). This line of work is referred to the distillation, which can be divided into two distinct categories.

In the first category, researchers aim to train a deep neural network that maps noise to the endpoint of the probability flow ODE, while disregarding the information at intermediate time points. This class of methods includes knowledge distillation (KD) (Luhman and Luhman, 2021), Euler particle transport (EPT) (Gao et al., 2022), rectified flow (Liu et al., 2022), and diffusion model sampling with neural operator (DSNO) (Zheng et al., 2023a). Unfortunately, these methods are hindered by training instability and low generation quality, as they solely focus on long-range information and are unable to capture the short-range structure at intermediate time points.

The second category of distillation, which is highly relevant to our work, aims to fit the characteristics at each time point using deep neural networks, as demonstrated by Salimans and Ho (2022), Song et al. (2023), Kim et al. (2024), Zhou et al. (2024). These methods take into account both the long-range and short-range structures of the original probability flow, resulting in a high quality of one-step generation. Furthermore, these models exhibit

the potential for further enhancement through a few-step iteration process. However, despite their impressive generation quality and training stability, these methods have not yet undergone rigorous theoretical analysis. In contrast, our paper presents a comprehensive framework for generative models utilizing characteristic matching and establishes a rigorous convergence analysis, providing theoretical guarantees for these methods. Additionally, we incorporate the exponential integrator into the characteristic matching procedure. Notably, our characteristic generator surpasses the generation quality achieved by Salimans and Ho (2022), Song et al. (2023), Kim et al. (2024) without the assistance of GANs.

5.2. **Error analysis for diffusion and flow-based models.** Although a large body of literature has been devoted to the theoretical analysis for diffusion and flow-based generative models, a majority of these works rely on intractable assumptions, such as the regularity of the probability flow SDEs or ODEs. In contrast, our theoretical findings are established under fewer and milder assumptions on the prior and target distribution (Assumptions 1 and 2).

The errors of diffusion-based and flow-based one-step generative models primarily focus on three aspects: velocity matching error, discretization error, and characteristic fitting error. Existing literature on theoretical analysis of these generative models commonly assumes that the $L^2$-risk of score or velocity matching is sufficiently small (Lee et al., 2022, 2023, Chen et al., 2023d,c, Benton et al., 2024a,b). However, only a limited number of studies have specifically investigated the convergence rates of score matching (Oko et al., 2023, Chen et al., 2023b, Han et al., 2024) and velocity matching (Chang et al., 2024, Gao et al., 2024, Jiao et al., 2024). The convergence rate of the velocity, as derived in Theorem 3.10, achieves minimax optimality under the assumption of Lipschitz continuity of the target function, which improves upon the rates proposed by Chen et al. (2023b), Chang et al. (2024). The discretization error of the numerical sampler has been explored for both diffusion models (Lee et al., 2022, 2023, Chen et al., 2023d, Benton et al., 2024a), and flow-based methods (Chen et al., 2023c, Gao and Zhu, 2024, Li et al., 2024c,a). To the best of our knowledge, Theorem 3.14 is the first to systematically analyze these three errors, providing theoretical guidance for selecting suitable neural networks and determining the number of numerical discretization steps.

## 6. Conclusions and Future Work

In this paper, we propose the characteristic generator, a novel one-step generative model that combines sampling efficiency with high generation quality. In terms of theoretical analysis, we have examined the errors in velocity matching, Euler discretization, and characteristic fitting, enabling us to establish a non-asymptotic convergence rate for the characteristic generator in 2-Wasserstein distance. This analysis represents the first comprehensive investigation into simulation-free one-step generative models, refining the error analysis of flow-based generative models in prior research. We have validated the effectiveness of our method through experiments on synthetic and real datasets. The results demonstrate that the characteristic generator achieves high generation quality with just a single evaluation of

the neural network. This highlights the efficiency and stability of our model in generating high-quality samples.

Finally, we would like to emphasize that our framework of one-step generation is highly versatile and can be extended to conditional generative learning directly. By incorporating the encoder-decoder technique, our approach can be generalized to the latent space, enabling its application in a wide range of practical scenarios, including video generation. The characteristic generator improved by these techniques lays a technical foundation for deploying large-scale generative models on end devices.

On the theoretical front, we intend to exploit the regularity of the velocity field in our error analysis for velocity matching. This will allow us to achieve a faster convergence rate. Additionally, we plan to analyze higher order and more stable numerical schemes, such as the high-order exponential integrator, in order to provide a comprehensive understanding of their effectiveness. Furthermore, we aim to establish a theoretical foundation for the role of semi-group penalties in the characteristic fitting. This will contribute to a deeper understanding of their impact and significance in our framework. A significant extension would be to move from our current average-in-time error guarantee (Theorem 3.14) to a stronger, uniform-in-time bound. Achieving this would likely require designing novel objective functionals for the characteristic fitting, potentially based on $L^\infty$-risk minimization or adversarial (minimax) formulations. Investigating the theoretical properties and empirical performance of these approaches is an important open question.

This work provides both theoretical analysis and experimental validation. However, it is important to clarify the role of our theoretical bounds, as their direct numerical verification is challenging due to the well-known gap between deep learning theory and practice. This gap arises from several fundamental factors: (i) Our bounds, like many in the field, serve to characterize scaling laws and qualitative relationships, e.g., how error scales with dimension, sample size, or the number of Euler steps. They are not intended to be numerically predictive of the exact empirical error. The bounds often contain large, abstract constants that are intractable to compute in practice, making a direct numerical comparison infeasible. (ii) Since the theoretical bounds often represent a worst-case scenario, there exists a gap between the empirical error and the theoretical bound. (iii) Our theory analyzes the properties of the empirical risk minimizer, whereas in practice, non-convex optimization methods find local minima. This creates a gap between the object of our theoretical analysis and the model obtained empirically. Bridging this theory-practice gap and developing bounds that more closely reflect the performance of practically trained models remains a crucial challenge for the field.

## References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023a. arXiv:2303.08797.

Michael S. Albergo, Mark Goldstein, Nicholas M. Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings, 2023b. arXiv:2310.03725.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. Cambridge university press Cambridge, 1999.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Benedikt Bauer and Michael Kohler. On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly $d$-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in wasserstein distance, 2025. arXiv:2508.03210.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. Expert Certification.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024.

J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Ltd., Third edition, 2016.

Jinyuan Chang, Zhao Ding, Yuling Jiao, Ruoxuan Li, and Jerry Zhijian Yang. Deep conditional generative learning: Model and error analysis, 2024. arXiv:2402.01460.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4735–4763. PMLR, 23–29 Jul 2023a.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks, 2022. arXiv:2002.03938.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4672–4712. PMLR, 23–29 Jul 2023b.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68552–68575. Curran Associates, Inc., 2023c.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023d.

Richard Courant and David Hilbert. *Methods of Mathematical Physics*, volume 2. John Wiley & Sons, Ltd, 1989.

Bernard Dacorogna and Jürgen Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 7(1):1–26, 1990. ISSN 0294-1449.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

Zhao Ding, Chenguang Duan, Yuling Jiao, and Jerry Zhijian Yang. Semi-supervised deep Sobolev regression: Estimation, variable selection and beyond, 2024. arXiv:2401.04535.

Chenguang Duan, Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, and Jerry Zhijian Yang. Convergence rate analysis for deep Ritz method. *Communications in Computational Physics*, 31(4):1020–1048, 2022. ISSN 1991-7120.

Chenguang Duan, Yuling Jiao, Lican Kang, Xiliang Lu, and Jerry Zhijian Yang. On convergence rates of deep nonparametric regression under covariate shift, 2023.

Richard Duong, Jannis Chemseddine, Peter K. Friz, and Gabriele Steidl. Telegrapher's generative model via kac flows, 2025. arXiv:2506.20641.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. arXiv:2403.03206.

Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society (AMS), Second edition, 2010.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances, 2024. arXiv:2401.17958.

Yuan Gao, Yuling Jiao, Yang Wang, Yao Wang, Can Yang, and Shunkang Zhang. Deep generative learning via variational gradient flow. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2093–2101. PMLR, 09–15 Jun 2019.

Yuan Gao, Jian Huang, Yuling Jiao, Jin Liu, Xiliang Lu, and Zhijian Yang. Deep generative learning via Euler particle transport. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 336–368. PMLR, 16–19 Aug 2022.

Yuan Gao, Jian Huang, and Yuling Jiao. Gaussian interpolation flows, 2023. arXiv:2311.11475.

Yuan Gao, Jian Huang, Yuling Jiao, and Shurong Zheng. Convergence of continuous normalizing flows for learning probability distributions, 2024. arXiv:2404.00551.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 27. MIT Press, 2014.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019.

Louis Grenioux, Maxence Noble, Marylou Gabrié, and Alain Oliviero Durmus. Stochastic localization via iterative posterior sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16337–16376. PMLR, 21–27 Jul 2024.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Laszlo Gyorfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*, volume 1. Springer, 2002.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *The Twelfth International Conference on Learning Representations*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc., 2022.

Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19: 209–286, 2010. doi: 10.1017/S0962492910000048.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. *Journal of Machine Learning Research*, 23(116):1–43, 2022.

Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Second edition, 2008. doi: 10.1017/CBO9780511995569.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep Nonparametric Regression on Approximate Manifolds: Nonasymptotic Error Bounds with Polynomial Prefactors. *The Annals of Statistics*, 51(2):691 – 716, 2023a.

Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and GANs. *Applied and Computational Harmonic Analysis*, 65:249–278, 2023b. ISSN 1063-5203.

Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Convergence analysis of flow matching in latent space with transformers, 2024. arXiv:2404.02538.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10124–10134, June 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577. Curran Associates, Inc., 2022.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11201–11228. PMLR, 17–23 Jul 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael Kohler and Sophie Langer. On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a Function of Low Local Dimensionality by Deep Neural Networks. *IEEE transactions on information theory*, 68(6): 4032–4042, 2022.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the Wasserstein distance. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20205–20217. Curran Associates, Inc., 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22870–22882. Curran Associates, Inc., 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona, editors, *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 946–985. PMLR, 20 Feb–23 Feb 2023.

Bo Li, Shanshan Tang, and Haijun Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Communications in Computational Physics*, 27(2):379–411, 2019. ISSN 1991-7120.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, Provably. In *Forty-first International*

*Conference on Machine Learning*, 2024a.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models, 2024b. arXiv:2402.07802.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024c.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset Rademacher complexity. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1260–1285, Paris, France, 03–06 Jul 2015. PMLR.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport, 2022. arXiv:2209.14577.

Shiao Liu, Yunfei Yang, Jian Huang, Yuling Jiao, and Yang Wang. Non-asymptotic error bounds for bidirectional GANs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12328–12339. Curran Associates, Inc., 2021.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. arXiv:2209.03003.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787. Curran Associates, Inc., 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models, 2023a. arXiv:2211.01095.

Haoye Lu, Yiwei Lu, Dihong Jiang, Spencer Ryan Szabados, Sun Sun, and Yaoliang Yu. CM-GAN: Stabilizing GAN training with consistency models. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023b.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed, 2021. arXiv:2101.02388.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, Second edition, 2018.

Jürgen Moser. On the volume elements on a manifold. *Transactions of the American Mathematical Society*, 377(3):286–294, 1965. ISSN 1088-6850.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25858–25889. PMLR, 23–29 Jul 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26517–26582. PMLR, 23–29 Jul 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, volume 35, pages 35852–35865. Curran Associates, Inc., 2022.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. arXiv:1511.06434.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 2022. arXiv:2204.06125.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q.

Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-SD: Trajectory segmented consistency model for efficient image synthesis, 2024. arXiv:2404.13686.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17669–17680. Curran Associates, Inc., 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Saeed Saremi, Ji Won Park, and Francis Bach. Chain of log-concave markov chains. In *The Twelfth International Conference on Learning Representations*, 2024.

Johannes Schmidt-Hieber. Nonparametric Regression using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Zuowei Shen. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1415–1428. Curran Associates, Inc., 2021b.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021c.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 23–29 Jul 2023.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics (SSS). Springer New York, NY, 2009.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics (SSS). Springer Cham, second edition, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften (GL)*. Springer Berlin, Heidelberg, First edition, 2009.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for Gaussian mixture models. In *Forty-first International Conference on Machine Learning*, 2024.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564 – 1599, 1999. doi: 10.1214/aos/1017939142.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking Lipschitz neural networks and certified robustness: A boolean function perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 19398–19413. Curran Associates, Inc., 2022.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a prceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42390–42402. PMLR, 23–29 Jul 2023a.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. DPM-Solver-v3: Improved diffusion ODE solver with empirical model statistics. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55502–55542. Curran Associates, Inc., 2023b.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity Distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation, 2024. arXiv:2404.04057.

Xingyu Zhou, Yuling Jiao, Jin Liu, and Jian Huang. A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543):1837–1848, 2023.

OUTLINE OF THE SUPPLEMENTARY MATERIAL

The supplementary material provides detailed derivations, proofs, and auxiliary results, organized into the following appendices:

(I) Appendix A summarizes the notation used throughout the paper.

(II) Appendix B provides auxiliary definitions and lemmas.

(III) Appendix C presents the proofs for the theoretical results in Section 2.

(IV) Appendix D establishes regularity properties of the probability flow ODE and provides the proofs for the propositions in Section 3.2.

(V) Appendix E provides the proofs for the results in Section 3.3.

(VI) Appendix F provides the proofs for the results in Section 3.4.

(VII) Appendix G provides the proofs for the results in Section 3.5.

(VIII) Appendix H provides the proofs for the results in Section 3.7.

(IX) Appendix I details a generalization error analysis for nonparametric regression, which is used to establish the oracle inequalities for velocity matching (Appendix E) and characteristic fitting (Appendix G).

(X) Appendix J establishes approximation results for deep neural networks with a Lipschitz constraint, which are also used for the convergence rates of velocity matching (Appendix E) and characteristic fitting (Appendix G).

(XI) Appendix K specifies the parameterization of the denoiser employed in Section 4.

(XII) Appendix L reports the full hyperparameters used in our numerical experiments, including model size, training iterations, batch size, learning rate, optimizer, and augmentation choices.

(XIII) Appendix M provides some remarks on the training time consumption of our numerical experiments.

(XIV) Appendix N conducts a thorough comparison of the implementation details between our method and CTM.

(XV) Appendix O conducts a thorough theoretical comparison between the one-step Euler sampler and characteristic generator.

## Appendix A. Summary of Notations

Table 7 summarizes the notations used in throughout the paper for easy reference and cross checking.

Table 7: The list of notations used in Sections 2, and 3.

| Symbols | Description |
|---|---|
| $\mu_0$ | Prior distribution, $\mu_0 = N(0, I_d)$ (Assumption 1). |
| $\mu_1$ | Target distribution, $\mu_1 = N(0, \sigma^2 I_d) * \nu$ (Assumption 2). |
| $\rho_0, \rho_1$ | Densities of $\mu_0$ and $\mu_1$. |
| $\mu_t, \rho(t, x)$ | Interpolant distribution and its density, see Prop. 2.1. |
| $(g_{t,s}^*)_\sharp \mu_t$ | Push-forward of $\mu_t$ by the flow $g_{t,s}^*$, exactly $\mu_s$ (§2.1). |
| $\varphi_d$ | Standard Gaussian density in $\mathbb{R}^d$ (Assumption 2). |
| $\nu$ | Base distribution with $\operatorname{supp}(\nu) \subseteq [0, 1]^d$ (Assumption 2). |
| $X_0, X_1$ | Independent samples from $\mu_0$ and $\mu_1$. |
| $X_t$ | Stochastic interpolant $X_t = \alpha(t)X_0 + \beta(t)X_1$ (Eq. (2.1)). |
| $\alpha(t), \beta(t)$ | Interpolant coefficients (Cond. 1); write $\alpha_t, \beta_t$. |
| $s^*(t, x)$ | Score $\nabla \log \rho_t(x)$ (§3.1). |
| $b^*(t, x)$ | Velocity field (Eq. (2.3)); transport Eq. (2.2). |
| $g_{t,s}^*$ | Continuous probability flow map: $x_t \mapsto x_s$ (§2.1). |
| $Z_t, Z_s$ | Flowed variables $Z_t = g_{0,t}^*(Z_0)$, $Z_s = g_{0,s}^*(Z_0)$ (§2.1). |
| $T$ | Stopping time in $(1/2, 1)$ used for training/evaluation. |
| $\mathcal{S}, n$ | Training dataset for velocity matching $\{(t^{(i)}, X_0^{(i)}, X_1^{(i)})\}_{i=1}^n$ and its size. |
| $\mathcal{L}(b)$ | Population risk for velocity (Eq. (2.6)). |
| $\widehat{\mathcal{L}}_n(b)$ | Empirical risk for velocity (Eq. (2.7)). |
| $\mathscr{B}$ | Hypothesis class for velocity networks (Thm. 3.10). |
| $\widehat{b}$ | Velocity estimator via ERM (Eq. (2.8)). |
| $\mathcal{E}_T(\widehat{b})$ | Time-averaged velocity error (Eq. (3.1)). |
| $\widehat{\mu}_T$ | Push-forward at time $T$ under ODE with $\widehat{b}$ (§3.3). |
| $K, \tau, t_k$ | Number of steps, step size $\tau = T/K$, and grid $t_k = k\tau$ (§2.3). |
| $\widehat{E}_{k,\ell}^\tau$ | Euler flow map (Eq. (2.9)). |
| $\mathcal{Z}, m$ | Teacher trajectories and their number for fitting (§2.4). |
| $\mathcal{R}(g)$ | Population characteristic fitting risk (Eq. (2.5)). |
| $\widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(g)$ | Empirical characteristic risk using Euler data (Eq. (2.10)). |
| $\lambda, \widehat{\mathscr{P}}(g)$ | Semi-group penalty weight and penalty (Eq. (2.12)). |
| $\widehat{\mathcal{R}}_{m,K}^{\mathrm{Euler},\lambda}(g)$ | Penalized empirical risk (Eq. (2.12)). |
| $\Delta_{kj\ell}(g)$ | Semi-group discrepancy term (§2.4). |
| $\mathscr{G}$ | Hypothesis class for characteristic networks (Thm. 3.14). |

Continued from previous page

| Symbols | Description |
|:---:|:---|
| $\widehat{g}_{t,s}$ | Learned characteristic generator (ERM; Eq. (2.11)). |
| $\mathcal{D}(\widehat{g})$ | Time-averaged squared Wasserstein metric for $\widehat{g}$ (Eq. (3.3)). |
| $B_{\mathrm{vel}}, B_{\mathrm{flow}}, B'_{\mathrm{flow}}$ | Local bounds for velocity/flow and time-derivatives (§3.2). |
| $G$ | Uniform bound on spatial Jacobians (Prop. 3.5). |
| $D$ | Time-regularity constant for $\partial_t b^*$ (Prop. 3.8). |
| $\kappa(T)$ | Function in time-regularity bounds (Prop. 3.8). |
| $C$ | Universal constants depending on $d, \sigma$ (various theorems). |
| $d^*$ | Intrinsic dimension with $d^* \ll d$ (§3.7). |
| $P \in \mathbb{R}^{d \times d^*}$ | Column-orthonormal embedding matrix (§3.7). |
| $\widetilde{\nu}, P_\sharp \widetilde{\nu}$ | Base distribution on $[0,1]^{d^*}$ and its push-forward to $\mathbb{R}^d$. |
| $\widetilde{b}^*, \widetilde{g}^*$ | Intrinsic-coordinate velocity/flow (Lemma 3.18). |

## Appendix B. Supplemental Definitions and Lemmas

In this section, we introduce some supplemental definitions and lemmas that are used in the proofs.

We first introduce the sub-Gaussian random variable Vershynin (2018), Wainwright (2019).

**Definition B.1** (Sub-Gaussian). A random variable $X$ with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number $\sigma$ such that

$$\log \mathbb{E}[\exp(\lambda(X - \mu))] \leq \frac{\sigma^2 \lambda^2}{2}, \quad \lambda \in \mathbb{R}.$$

Here the constant $\sigma$ is referred to as the variance proxy.

The following results (Lemmas B.2 to B.4) shows some important properties of sub-Gaussian variables, whose proofs can be found in (Wainwright, 2019, Section 2.1).

**Lemma B.2** (Chernoff bound). *Let $X$ be a $\sigma^2$-sub-Gaussian random variable with zero mean. Then it holds that*

$$\Pr\{|X| \geq t\} \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

**Lemma B.3.** *Let $X$ be a $\sigma^2$-sub-Gaussian random variable with zero mean. Then it holds that*

$$\mathbb{E}\left[\exp\left(\frac{\lambda X^2}{2\sigma^2}\right)\right] \leq \frac{1}{\sqrt{1-\lambda}}, \quad \lambda \in [0, 1).$$

**Lemma B.4.** *Let $X_0$ be a $\sigma_0^2$-sub-Gaussian, and let $X_1$ be a $\sigma_1^2$-sub-Gaussian independent of $X_0$. Then the randon variable $\alpha X_0 + \beta X_1$ is sub-Gaussian with variance proxy $\alpha^2 \sigma_0^2 + \beta^2 \sigma_1^2$.*

Lemmas B.5 and B.6 show bounds of the tail probability and the expectation of the maximum of $N$ sub-Gaussian variables.

**Lemma B.5.** *Let $\{X_n\}_{n=1}^N$ be a set of $\sigma^2$-sub-Gaussian random variables with zero mean, then it follows that*

$$\Pr\Big\{ \max_{1\leq n\leq N} |X_n| \geq t \Big\} \leq 2N \exp\Big( -\frac{t^2}{2\sigma^2} \Big).$$

The following lemma states that the expectation of the maximum of the squares of $N$ sub-Gaussian variables is bounded by $\log N$.

**Lemma B.6.** *Let $\{X_n\}_{n=1}^N$ be a set of $\sigma^2$-sub-Gaussian random variables with zero mean, then it follows that*

$$\mathbb{E}\Big[ \max_{1\leq n\leq N} X_n^2 \Big] \leq 4\sigma^2 (\log N + 1).$$

*Proof of Lemma B.5.* It is straightforward that

$$\Pr\Big\{ \max_{1\leq n\leq N} |X_n| \geq t \Big\} \leq \sum_{n=1}^N \Pr\{ |X_n| \geq t \} \leq 2N \exp\Big( -\frac{t^2}{2\sigma^2} \Big),$$

where the last inequality holds from Lemma B.2. This completes the proof. $\square$

*Proof of Lemma B.6.* By Jensen's inequality, it is straightforward that

$$\exp\Big( \frac{\lambda}{2\sigma^2} \mathbb{E}\Big[ \max_{1\leq n\leq N} \xi_n^2 \Big] \Big) \leq \mathbb{E}\Big[ \max_{1\leq n\leq N} \exp\Big( \frac{\lambda \xi_n^2}{2\sigma^2} \Big) \Big] \leq N\mathbb{E}\Big[ \exp\Big( \frac{\lambda \xi_1^2}{2\sigma^2} \Big) \Big] \leq \frac{N}{\sqrt{1-\lambda}},$$

where the last inequality holds from Lemma B.3 for each $\lambda \in [0,1)$. Letting $\lambda = 1/2$ yields the desired inequality. $\square$

**Lemma B.7** (Fourth moment of standard Gaussian). *Let $\epsilon \sim N(0, I_d)$. Then $\mathbb{E}[\|\epsilon\|_2^4] = d^2 + 2d$.*

*Proof of Lemma B.7.* It is straightforward that

$$\mathbb{E}[\|\epsilon\|_2^4] = \mathbb{E}\Big[ \sum_{k=1}^d \epsilon_k^4 + \sum_{k\neq\ell} \epsilon_k^2 \epsilon_\ell^2 \Big] = \sum_{k=1}^d \mathbb{E}[\epsilon_k^4] + \sum_{k\neq\ell} \mathbb{E}[\epsilon_k^2]\mathbb{E}[\epsilon_\ell^2] = d^2 + 2d,$$

where we used the fact that $\mathbb{E}[X^4] = 3$ for $X \sim N(0,1)$. $\square$

We next introduce the notation of covering number and Vapnik-Chervonenkis dimension (VC-dimension), both of which measure the complexity of a function class Mohri et al. (2018), Vaart and Wellner (2023). They are used in the error analysis for velocity matching (Section E) and characteristic fitting (Section G).

**Definition B.8** (Covering number). Let $\mathscr{F}$ be a class of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$. Suppose $\mathcal{D} = \{X^{(i)}\}_{i=1}^n$ is a set of samples in $\mathbb{R}^d$. Define the $L^\infty(\mathcal{D})$-norm of the function $f \in \mathscr{F}$ as $\|f\|_{L^\infty(\mathcal{D})} = \max_{1\leq i\leq n} |f(X^{(i)})|$. A function set $\mathscr{F}_\delta$ is called an $L^\infty(\mathcal{D})$ $\delta$-cover of $\mathscr{F}$ if for each $f \in \mathscr{F}$, there exits $f_\delta \in \mathscr{F}_\delta$ such that $\|f - f_\delta\|_{L^\infty(\mathcal{D})} \leq \delta$. Furthermore,

$$N(\delta, \mathscr{F}, L^\infty(\mathcal{D})) = \inf\Big\{ |\mathscr{F}_\delta| : \mathscr{F}_\delta \text{ is a } L^\infty(\mathcal{D}) \text{ } \delta\text{-cover of } \mathscr{F} \Big\}$$

is called the $L^\infty(\mathcal{D})$ $\delta$-covering number of $\mathscr{F}$.

**Definition B.9** (VC-dimension)**.** Let $\mathscr{F}$ be a class of functions from $\mathbb{R}^d$ to $\{\pm 1\}$. For any non-negative integer $m$, we define the growth function of $\mathscr{F}$ as

$$\Pi_{\mathscr{F}}(m) = \max_{\{X^{(i)}\}_{i=1}^m \subseteq \mathbb{R}^d} |\{(f(X^{(1)}), \dots, f(X^{(m)})) : f \in \mathscr{F}\}|.$$

A set $\{X^{(i)}\}_{i=1}^m$ is said to be shattered by $\mathscr{F}$ when $|\{(f(X^{(1)}), \dots, f(X^{(m)})) : f \in \mathscr{F}\}| = 2^m$. The Vapnik-Chervonenkis dimension of $\mathscr{F}$, denoted $\mathrm{VCdim}(\mathscr{F})$, is the size of the largest set that can be shattered by $\mathscr{F}$, that is, $\mathrm{VCdim}(\mathscr{F}) = \max\{m : \Pi_{\mathscr{F}}(m) = 2^m\}$. For a class $\mathscr{F}$ of real-valued functions, we define $\mathrm{VCdim}(\mathscr{F}) = \mathrm{VCdim}(\mathrm{sign}(\mathscr{F}))$.

**Lemma B.10.** *Let $\mathscr{F}$ be a class of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$, and let $\mathscr{H}$ be a function class defined as $\mathscr{H} = \{(x, f) \mapsto h(f, x) \in \mathbb{R} : f \in \mathscr{F}\}$. Suppose $\mathcal{D} = \{X^{(i)}\}_{i=1}^n$ is a set of samples in $\mathbb{R}^d$. If there exists a constant $L > 0$ such that for each $f, f' \in \mathscr{F}$,*

$$\max_{1 \le i \le n} |h(f, X^{(i)}) - h(f', X^{(i)})| \le L \max_{1 \le i \le n} |f(X^{(i)}) - f'(X^{(i)})|,$$

*then the following inequality holds for each $\delta > 0$,*

$$N(L\delta, \mathscr{H}, L^\infty(\mathcal{D})) \le N(\delta, \mathscr{F}, L^\infty(\mathcal{D})).$$

*Proof of Lemma B.10.* Let $\mathscr{F}_\delta$ be an $L^\infty(\mathcal{D})$ $\delta$-cover of $\mathscr{F}$ with $|\mathscr{F}_\delta| = N(L\delta, \mathscr{H}, L^\infty(\mathcal{D}))$. Define $\mathscr{H}_\delta = \{(x, f) \mapsto h(f, x) \in \mathbb{R} : f \in \mathscr{F}_\delta\}$. Then for each $h(f, \cdot) \in \mathscr{H}$, there exists $h(f_\delta, \cdot) \in \mathscr{H}_\delta$, such that

$$\max_{1 \le i \le n} |h(f, X^{(i)}) - h(f_\delta, X^{(i)})| \le L \max_{1 \le i \le n} |f(X^{(i)}) - f_\delta(X^{(i)})| \le L\delta.$$

Thus $\mathscr{H}_\delta$ is an $L^\infty(\mathcal{D})$ $(L\delta)$-cover of $\mathscr{H}$. This completes the proof. $\qquad \square$

We then bound the covering number by VC-dimension as following lemma.

**Lemma B.11** ((Anthony et al., 1999, Theorem 12.2))**.** *Let $\mathscr{F}$ be a class of functions mapping from $\mathbb{R}^d$ to $[0, B]$. Then it follows that for each $n \ge \mathrm{VCdim}(\mathscr{F})$,*

$$\sup_{\mathcal{D} \in (\mathbb{R}^d)^n} \log N(\delta, \mathscr{F}, L^\infty(\mathcal{D})) \le \mathrm{VCdim}(\mathscr{F}) \log\left(\frac{enB}{\delta\,\mathrm{VCdim}(\mathscr{F})}\right).$$

The following lemma provides a VC-dimension bound for neural network classes with piecewise-polynomial activation functions.

**Lemma B.12** ((Bartlett et al., 2019, Theorem 7))**.** *The VC-dimension of a neural network class with piecewise-polynomial activation functions is bounded as follows*

$$\mathrm{VCdim}(N(L, S)) \le CLS \log(S),$$

*where $C$ is an absolute constant.*

## Appendix C. Proof of Results in Section 2

The proof of Proposition 2.1 follows from the proof of (Albergo et al., 2023b, Theorem 2) and (Albergo et al., 2023a, Theorem 2.6).

*Proof of Proposition 2.1.* The characteristic function of $X_t$ (2.1) is given as

$$
\begin{aligned}
\varphi_{X_t}(\xi) &= \mathbb{E}_{X_t}[\exp(i\xi \cdot X_t)] \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp(i\xi \cdot (\alpha_t x_0 + \beta_t x_1))\rho_0(x_0)\rho_1(x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1 \\
&= \int_{\mathbb{R}^d} \exp(i\xi \cdot (\alpha_t x_0))\rho_0(x_0)\,\mathrm{d}x_0 \int_{\mathbb{R}^d} \exp(i\xi \cdot (\beta_t x_1))\rho_1(x_1)\,\mathrm{d}x_1 \\
&= \mathbb{E}_{X_0}[\exp(i\alpha_t\xi \cdot X_0)]\mathbb{E}_{X_1}[\exp(i\beta_t\xi \cdot X_1)] = \varphi_{X_0}(\alpha_t\xi)\varphi_{X_1}(\beta_t\xi),
\end{aligned}
$$

where $\xi \in \mathbb{R}^d$, and we used the fact that $X_0$ is independent of $X_1$. On the other hand,

$$
\begin{aligned}
&\int_{\mathbb{R}^d} \exp(i\xi \cdot x)\Big(\frac{1}{\beta_t}\int_{\mathbb{R}^d} \rho_0(x_0)\rho_1\Big(\frac{x - \alpha_t x_0}{\beta_t}\Big)\,\mathrm{d}x_0\Big)\,\mathrm{d}x \\
&= \frac{1}{\beta_t}\int_{\mathbb{R}^d \times \mathbb{R}^d} \exp(i\alpha_t\xi \cdot x_0)\exp(i\xi \cdot (x - \alpha_t x_0))\rho_0(x_0)\rho_1\Big(\frac{x - \alpha_t x_0}{\beta_t}\Big)\,\mathrm{d}x_0\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \exp(i\alpha_t\xi \cdot x_0)\rho_0(x_0)\Big\{\int_{\mathbb{R}^d} \exp\Big(i\xi \cdot (x - \alpha_t x_0)\Big)\rho_1\Big(\frac{x - \alpha_t x_0}{\beta_t}\Big)\,\mathrm{d}\Big(\frac{x - \alpha_t x_0}{\beta_t}\Big)\Big\}\,\mathrm{d}x_0 \\
&= \int_{\mathbb{R}^d} \exp(i\alpha_t\xi \cdot x_0)\rho_0(x_0)\Big\{\int_{\mathbb{R}^d} \exp(i\beta_t\xi \cdot x_1)\rho_1(x_1)\,\mathrm{d}x_1\Big\}\,\mathrm{d}x_0 = \varphi_{X_0}(\alpha_t\xi)\varphi_{X_1}(\beta_t\xi).
\end{aligned}
$$

Combining the above two equality deduces that for each $\xi \in \mathbb{R}^d$,

$$
\varphi_{X_t}(\xi) = \int_{\mathbb{R}^d} \exp(i\xi \cdot x)\Big(\frac{1}{\beta_t}\int_{\mathbb{R}^d} \rho_0(x_0)\rho_1\Big(\frac{x - \alpha_t x_0}{\beta_t}\Big)\,\mathrm{d}x_0\Big)\,\mathrm{d}x.
$$

By using Fourier inversion theorem, we obtain the density function of $X_t$ as

$$
\rho_t(x) = \frac{1}{\beta_t}\int_{\mathbb{R}^d} \rho_0(x_0)\rho_1\Big(\frac{x - \alpha_t x_0}{\beta_t}\Big)\,\mathrm{d}x_0.
$$

By a same argument, we have

$$
\rho_t(x) = \frac{1}{\alpha_t}\int_{\mathbb{R}^d} \rho_0\Big(\frac{x - \beta_t x_1}{\alpha_t}\Big)\rho_1(x_1)\,\mathrm{d}x_1.
$$

We next turn to verify that the density function solves the transport equation. Let $\phi$ be an arbitrary smooth testing function. By the definition of the interpolant (2.1), it holds that

$$
\mathbb{E}_{X_t}[\phi(X_t)] = \mathbb{E}_{(X_0, X_1)}[\phi(\alpha_t X_0 + \beta_t X_1)].
$$

Taking derivative with respect to $t$ on the left-hand side of the equality yields

$$
(C.1) \qquad \frac{\partial}{\partial t}\Big(\int_{\mathbb{R}^d} \phi(x)\rho_t(x)\,\mathrm{d}x\Big) = \int_{\mathbb{R}^d} \phi(x)\partial_t\rho_t(x)\,\mathrm{d}x.
$$

Similarly, for the right-hand side of the equality, we have

$$\frac{\partial}{\partial t}\Big(\int_{\mathbb{R}^d\times\mathbb{R}^d}\phi(\alpha_t x_0+\beta_t x_2)\rho_0(x_0)\rho_1(x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1\Big)$$

$$=\int_{\mathbb{R}^d\times\mathbb{R}^d}(\dot\alpha_t x_0+\dot\beta_t x_1)\cdot\nabla\phi(\alpha_t x_0+\beta_t x_1)\rho_0(x_0)\rho_1(x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1$$

$$=\int_{\mathbb{R}^d}\mathbb{E}[\dot\alpha_t X_0+\dot\beta_t X_1|X_t=x]\cdot\nabla\phi(x)\rho_t(x)\,\mathrm{d}x$$

$$\text{(C.2)}\qquad =\int_{\mathbb{R}^d}b_t^*(x)\cdot\nabla\phi(x)\rho_t(x)\,\mathrm{d}x=-\int_{\mathbb{R}^d}\phi(x)\nabla\cdot(b_t^*(x)\rho_t(x))\,\mathrm{d}x,$$

where the first equality follows from the chain rule, the forth equality holds from the definition of $b_t^*$ (2.3), and the last equality is due to integration by parts and the divergence theorem (Evans, 2010, Theorem 1 in Section C.2). Combining (C.1) and (C.2) gives the desired transport equation. $\qquad\square$

### APPENDIX D. PROPERTIES OF THE PROBABILITY FLOW ODE

In this section, we present some auxiliary properties of the probability flow ODE in Section D.1. Proofs of the propositions in Section 3.2 are given in Section D.2.

D.1. **Auxiliary Properties.** Lemmas D.1 to D.4 are established under Assumption 1.

**Lemma D.1.** *Suppose Assumption 1 holds. Then the conditional score function is given as*

$$\nabla_x\log\rho_{t|1}(x|x_1)=-\frac{1}{\alpha_t^2}x+\frac{\beta_t}{\alpha_t^2}x_1,\quad t\in(0,1).$$

*Proof of Lemma D.1.* Given $X_1=x_1$, using the definition of stochastic interpolant (2.1) implies

$$(X_t|X_1=x_1)\sim N(\beta_t x_1,\alpha_t^2 I_d),\quad t\in(0,1),$$

which implies the desired result immediately. $\qquad\square$

**Lemma D.2.** *Suppose Assumption 1 holds. Then the following holds*

$$\mathbb{E}[X_1|X_t=x]=\frac{1}{\beta_t}x+\frac{\alpha_t^2}{\beta_t}\nabla_x\log\rho_t(x).$$

*Proof of Lemma D.2.* It is straightforward that

$$\nabla_x\log\rho_t(x)=\frac{\nabla_x\rho_t(x)}{\rho_t(x)}=\frac{1}{\rho_t(x)}\nabla_x\Big(\int_{\mathbb{R}^d}\rho_{t|1}(x|x_1)\rho_1(x_1)\,\mathrm{d}x_1\Big)$$

$$=\int_{\mathbb{R}^d}\frac{\rho_{t|1}(x|x_1)\rho_1(x_1)}{\rho_t(x)}\nabla_x\log\rho_{t|1}(x|x_1)\,\mathrm{d}x_1$$

$$=\int_{\mathbb{R}^d}\rho_{1|t}(x_1|x)\Big(-\frac{1}{\alpha_t^2}x+\frac{\beta_t}{\alpha_t^2}x_1\Big)\,\mathrm{d}x_1$$

$$=-\frac{1}{\alpha_t^2}x+\frac{\beta_t}{\alpha_t^2}\mathbb{E}[X_1|X_t=x],$$

where the we used the definition of conditional density and Lemma D.1. This completes the proof. $\qquad\square$

**Lemma D.3.** *Suppose Assumptions 1 holds. Then it follows that*

$$\nabla b^*(t,x) = \frac{\dot\alpha_t}{\alpha_t} I_d + \Big(\frac{\dot\beta_t}{\beta_t} - \frac{\dot\alpha_t}{\alpha_t}\Big)\frac{\beta_t^2}{\alpha_t^2}\,\mathrm{Cov}(X_1|X_t = x).$$

*Proof of Lemma D.3.* According to the proof of Proposition 3.1, we have

(D.1) $$b^*(t,x) = \frac{\dot\alpha_t}{\alpha_t} x + \beta_t\Big(\frac{\dot\beta_t}{\beta_t} - \frac{\dot\alpha_t}{\alpha_t}\Big)\mathbb{E}[X_1|X_t = x],$$

which deduces

(D.2) $$\nabla b^*(t,x) = \frac{\dot\alpha_t}{\alpha_t} I_d + \beta_t\Big(\frac{\dot\beta_t}{\beta_t} - \frac{\dot\alpha_t}{\alpha_t}\Big)\nabla\mathbb{E}[X_1|X_t = x].$$

Hence it suffices to estimate the gradient of the conditional expectation. In fact,

$$\nabla_x\mathbb{E}[X_1|X_t = x] = \nabla_x\Big(\frac{1}{\rho_t(x)}\int_{\mathbb{R}^d} x_1 \rho_{t|1}(x|x_1)\rho_1(x_1)\,\mathrm{d}x_1\Big)$$

$$= -\frac{\nabla_x\rho_t(x)}{\rho_t^2(x)}\int_{\mathbb{R}^d} x_1^T \rho_{t|1}(x|x_1)\rho_1(x_1)\,\mathrm{d}x_1 + \frac{1}{\rho_t(x)}\int_{\mathbb{R}^d} \nabla_x\rho_{t|1}(x|x_1)x_1^T\rho_1(x_1)\,\mathrm{d}x_1$$

(D.3) $$= -\nabla_x\log\rho_t(x)\int_{\mathbb{R}^d} x_1^T \rho_{1|t}(x_1|x)\,\mathrm{d}x_1 + \int_{\mathbb{R}^d} \nabla_x\log\rho_{t|1}(x|x_1)x_1^T\rho_{1|t}(x_1|x)\,\mathrm{d}x_1,$$

where we used the definition of the conditional density that

$$\rho_{1|t}(x_1|x) = \frac{\rho_{t|1}(x|x_1)\rho_1(x_1)}{\rho_t(x)}.$$

For the first term in the right-hand side of (D.3), applying Lemma D.2 implies

(D.4) $$-\nabla_x\log\rho_t(x)\int_{\mathbb{R}^d} x_1^T \rho_{1|t}(x_1|x)\,\mathrm{d}x_1$$
$$= \Big(\frac{1}{\alpha_t^2} x - \frac{\beta_t}{\alpha_t^2}\mathbb{E}[X_1|X_t = x]\Big)\mathbb{E}[X_1|X_t = x]^T.$$

For the second term, it follows from Lemma D.1 that

(D.5) $$\int_{\mathbb{R}^d} \log\nabla_x\rho_{t|1}(x|x_1)x_1^T\rho_{1|t}(x_1|x)\,\mathrm{d}x_1$$
$$= -\frac{1}{\alpha_t^2} x\mathbb{E}[X_1|X_t = x]^T + \frac{\beta_t}{\alpha_t^2}\mathbb{E}[X_1 X_1^T|X_t = x].$$

Plugging (D.4) and (D.5) into (D.3) yields

(D.6) $$\nabla\mathbb{E}[X_1|X_t = x] = \frac{\beta_t}{\alpha_t^2}\,\mathrm{Cov}(X_1|X_t = x).$$

Substituting (D.6) into (D.2) completes the proof. $\qquad\square$

The following lemma gives an explicit expression of time derivative of the velocity.

**Lemma D.4** ((Gao et al., 2023, Proposition 63)). *Suppose Assumption 1 holds. Then it follows that for each $(t, x) \in (0, 1) \times \mathbb{R}^d$,*

$$\partial_t b^*(t, x) = \Big(\frac{\ddot{\alpha}_t}{\alpha_t} - \frac{\dot{\alpha}_t^2}{\alpha_t^2}\Big)x + \Big(\alpha_t^2 \frac{\ddot{\beta}_t}{\beta_t} - \dot{\alpha}_t \alpha_t \frac{\dot{\beta}_t}{\beta_t} - \ddot{\alpha}_t \alpha_t + \dot{\alpha}_t^2\Big)\frac{\beta_t}{\alpha_t^2}\mathbb{E}[X_1 | X_t = x]$$

$$+ \frac{\beta_t^2}{\alpha_t^2}\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\Big(\frac{\dot{\beta}_t}{\beta_t} - 2\frac{\dot{\alpha}_t}{\alpha_t}\Big)\mathrm{Cov}(X_1 | X_t = x)x$$

$$- \frac{\beta_t^3}{\alpha_t^2}\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)^2\Big(\mathbb{E}[X_1 X_1^T X_1 | X_t = x] - \mathbb{E}[X_1 X_1^T | X_t = x]\mathbb{E}[X_1 | X_t = x]\Big).$$

Lemmas D.5 to D.8 are established under Assumptions 1 and 2.

**Lemma D.5.** *Suppose Assumptions 1 and 2 hold. Then it follows that*

$$(X_1 | X_t = x) \stackrel{\mathrm{d}}{=} \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2}U_{t,x} + \sqrt{\frac{\sigma^2 \alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2}}\epsilon + \frac{\sigma^2 \beta_t}{\alpha_t^2 + \sigma^2 \beta_t^2}x,$$

*where $U_{t,x}$ is a random variable satisfying $\mathrm{supp}(U_{t,x}) = \mathrm{supp}(\nu)$, and $\epsilon \sim N(0, I_d)$.*

*Proof of Lemma D.5.* According to the definition of stochastic interpolant (2.1), we have

(D.7) $$\rho_{t|1}(x|x_1) = \frac{1}{(2\pi)^{d/2}\alpha_t^d}\exp\Big(-\frac{1}{\alpha_t^2}\|\beta_t x_1 - x\|_2^2\Big).$$

Using Assumption 2, the target distribution is given as

(D.8) $$\rho_1(x_1) = \frac{1}{(2\pi)^{d/2}\sigma^d}\int_{\mathbb{R}^d}\exp\Big(-\frac{1}{\sigma^2}\|x_1 - u\|_2^2\Big)\mathrm{d}\nu(u).$$

Combining (D.7) and (D.8) implies

$$\rho_{1|t}(x_1|x) = \frac{\rho_{t|1}(x|x_1)\rho_1(x_1)}{\rho_t(x)}$$

$$= \frac{1}{(2\pi\sigma\alpha_t)^d}\frac{1}{\rho_t(x)}\int_{\mathbb{R}^d}\exp\Big(-\frac{1}{\alpha_t^2}\|\beta_t x_1 - x\|_2^2\Big)\exp\Big(-\frac{1}{\sigma^2}\|x_1 - u\|_2^2\Big)\mathrm{d}\nu(u)$$

$$= \frac{1}{(2\pi)^{d/2}}\Big(\frac{\sigma^2 \alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2}\Big)^{d/2}\int_{\mathbb{R}^d}\exp\Big(-\frac{\alpha_t^2 + \sigma^2 \beta_t^2}{\sigma^2 \alpha_t^2}\Big\|x_1 - \frac{\sigma^2 \beta_t x + \alpha_t^2 u}{\alpha_t^2 + \sigma^2 \beta_t^2}\Big\|_2^2\Big)g(t, x, u)\,\mathrm{d}\nu(u),$$

$$= N\Big(\frac{\sigma^2 \beta_t x + \alpha_t^2 u}{\alpha_t^2 + \sigma^2 \beta_t^2}, \frac{\sigma^2 \alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2}\Big) * \nu_{t,x}(u),$$

where $g(t, x, u)$ is a function such that $\int \rho_{1|t}(x_1|x)\,\mathrm{d}x_1 = 1$ for each $(t, x) \in (0, 1) \times \mathbb{R}^d$, and the measure $\nu_{t,x}$ is defined as $\mathrm{d}\nu_{t,x}(u) = g(t, x, u)\,\mathrm{d}\nu(u)$. It is apparent that $\mathrm{supp}(\nu_{t,x}) = \mathrm{supp}(\nu)$. Therefore,

$$(X_1 | X_t = x) \stackrel{\mathrm{d}}{=} \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2}U_{t,x} + \sqrt{\frac{\sigma^2 \alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2}}\epsilon + \frac{\sigma^2 \beta_t}{\alpha_t^2 + \sigma^2 \beta_t^2}x,$$

where $U_{t,x} \sim \nu_{t,x}$ and $\epsilon \sim N(0, I_d)$ are two independent random variables. This completes the proof. $\square$

**Lemma D.6** (Conditional expectation). *Suppose Assumptions 1 and 2 hold. Then the following inequalities hold for each $t \in (0, 1)$ and $x \in \mathbb{R}^d$,*

$$|\mathbb{E}[X_{1,k} | X_t = x]| \leq M(1 + |x_k|), \quad 1 \leq k \leq d,$$

*where $M$ is a constant only depending on $d$ and $\sigma$.*

*Proof of Lemma D.6.* According to Lemma D.5, it holds that

$$\mathbb{E}[X_1|X_t = x] = \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}] + \frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}x,$$

which implies the desired inequalities directly. $\square$

**Lemma D.7** (Conditional covariance). *Suppose Assumptions 1 and 2 hold. Then the following inequalities hold for each $t \in (0,1)$ and $x \in \mathbb{R}^d$,*

$$\frac{\sigma^2\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}I_d \preceq \mathrm{Cov}(X_1|X_t = x) \preceq \frac{\sigma^2\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}I_d + d\Big(\frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2 I_d.$$

*Proof of Lemma D.7.* It is straightforward from Lemma D.5 that

$$\mathbb{E}[X_1|X_t = x] = \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}] + \frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}x,$$

which implies

$$\mathbb{E}[X_1|X_t = x]\mathbb{E}[X_1|X_t = x]^T$$
$$= \Big(\frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2\mathbb{E}[U_{t,x}]\mathbb{E}[U_{t,x}]^T + \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}]x^T + \Big(\frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2 xx^T.$$

On the other hand, using Lemma D.5 deduces

$$\mathbb{E}[X_1X_1^T|X_t = x] = \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}]x^T + \Big(\frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2 xx^T$$
$$+ \Big(\frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2\mathbb{E}[U_{t,x}U_{t,x}^T] + \frac{\sigma^2\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}I_d.$$

Combining the above two equalities yields

$$\mathrm{Cov}(X_1|X_t = x) = \mathbb{E}[X_1X_1^T|X_t = x] - \mathbb{E}[X_1|X_t = x]\mathbb{E}[X_1|X_t = x]^T$$
$$= \Big(\frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2\mathrm{Cov}(U_{t,x}) + \frac{\sigma^2\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}I_d.$$

According to Lemma D.5, the random variable $\|U_{t,x}\|_\infty \le 1$ and thus $\mathrm{Cov}(U_{t,x}) \preceq dI_d$. Consequently,

$$\frac{\sigma^2\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}I_d \preceq \mathrm{Cov}(X_1|X_t = x) \preceq \frac{\sigma^2\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}I_d + d\Big(\frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\Big)^2 I_d,$$

for each $(t,x) \in (0,1) \times \mathbb{R}^d$. This completes the proof. $\square$

By a same argument as Lemmas D.6 and D.7, we also have the following inequality. See (Gao et al., 2024, Lemma A.8) for a detailed proof.

**Lemma D.8.** *Suppose Assumptions 1 and 2 hold. Then the following inequalities hold for each $t \in (0,1)$ and $x \in \mathbb{R}^d$,*

$$\|\mathbb{E}[X_1X_1^TX_1|X_t = x] - \mathbb{E}[X_1^TX_1|X_t = x]\mathbb{E}[X_1|X_t]\|_2 \le M\alpha_t^2(1 + \|x\|_2),$$

*where $M$ is a constant only depending on $d$ and $\sigma$.*

D.2. **Proof of Results in Section 3.2.** Then we show proofs of propositions and corollaries in Section 3.2.

*Proof of Proposition 3.1.* Using the definition of velocity (2.3), we have

$$b^*(t,x) = \mathbb{E}[\dot{\alpha}_1 X_0 + \dot{\beta}_t X_1 | X_t = x] = \mathbb{E}\Big[\dot{\beta}_t X_1 + \frac{\dot{\alpha}_t}{\alpha_t}(X_t - \beta_t X_1)\Big| X_t = x\Big]$$

$$= \beta_t\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\mathbb{E}[X_1 | X_t = x] + \frac{\dot{\alpha}_t}{\alpha_t}x = \alpha_t^2\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\nabla_x \log \rho_t(x) + \frac{\dot{\beta}_t}{\beta_t}x,$$

where the second equality holds from the definition of stochastic interpolant (2.1), and the last equality is due to Lemma D.2. This completes the proof. □

*Proof of Proposition 3.2.* Using the definition of velocity (2.3), we have

$$b^*(t,x) = \mathbb{E}[\dot{\alpha}_1 X_0 + \dot{\beta}_t X_1 | X_t = x] = \mathbb{E}\Big[\dot{\beta}_t X_1 + \frac{\dot{\alpha}_t}{\alpha_t}(X_t - \beta_t X_1)\Big| X_t = x\Big]$$

$$= \beta_t\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\mathbb{E}[X_1 | X_t = x] + \frac{\dot{\alpha}_t}{\alpha_t}x$$

$$= \beta_t\Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\Big(\frac{\alpha_t^2}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}] + \frac{\sigma^2\beta_t}{\alpha_t^2 + \sigma^2\beta_t^2}x\Big) + \frac{\dot{\alpha}_t}{\alpha_t}x$$

$$= \frac{\alpha_t(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}] + \frac{\alpha_t\dot{\alpha}_t + \sigma^2\beta_t\dot{\beta}_t}{\alpha_t^2 + \sigma^2\beta_t^2}x$$

where the second equality holds from the definition of stochastic interpolant (2.1), and the fourth equality follows from Lemma D.6. This completes the proof. □

*Proof of Corollary 3.3.* For each $1 \le k \le d$, it follows from (2.4) that

$$x_k(s) = x_k(t) + \int_t^s b_k^*(\tau, x(\tau)) \, d\tau.$$

Using the triangular inequality and Jensen's inequality, we have

$$|x_k(s)| \le |x_k(t)| + \int_t^s |b_k^*(\tau, x(\tau))| \, d\tau$$

$$\le |x_k(t)| + \int_t^s \Big|\frac{\alpha_\tau(\alpha_\tau\dot{\beta}_\tau - \dot{\alpha}_\tau\beta_\tau)}{\alpha_\tau^2 + \sigma^2\beta_\tau^2}\Big| \, d\tau + \int_t^s \Big|\frac{\alpha_t\dot{\alpha}_t + \sigma^2\beta_t\dot{\beta}_t}{\alpha_t^2 + \sigma^2\beta_t^2}\Big| |x_k(\tau)| \, d\tau,$$

where the second inequality follows from Lemma D.6, and $M$ is a constant only depending on $d$ and $\sigma$. Applying Gronwall's inequality (Evans, 2010, Section B.2) yields

$$|x_k(s)| \le C \exp\Big(\int_t^s \Big|\frac{\alpha_t\dot{\alpha}_t + \sigma^2\beta_t\dot{\beta}_t}{\alpha_t^2 + \sigma^2\beta_t^2}\Big| \, d\tau\Big)\Big(|x_k(t)| + \int_t^s \Big|\frac{\alpha_\tau(\alpha_\tau\dot{\beta}_\tau - \dot{\alpha}_\tau\beta_\tau)}{\alpha_\tau^2 + \sigma^2\beta_\tau^2}\Big| \, d\tau\Big),$$

where the constant $C$ depends only on $d$ and $\sigma$. This completes the proof. □

*Proof of Corollary 3.4.* First, we observe that

$$g^*(t, s, x) = x + \int_t^s b^*(\tau, g^*(t, \tau, x)) \, d\tau.$$

Taking the partial derivative with respect to $s$ yields $\partial_s g^*(t, s, x) = b^*(s, g^*(t, s, x))$. Since $x \in \mathbb{B}_R^\infty$, by Corollary 3.3 and Proposition 3.2, we get $\|\partial_s g^*(t, s, x)\|_2 \le \sqrt{d}B_{\text{vel}}B_{\text{flow}}R$.

Now, we consider the partial derivative of $g^*(t, s, x)$ with respect to $t$, which leads to

$$(D.9) \qquad \partial_t g^*(t, s, x) = -b^*(t, x) + \int_t^s \nabla b^*(\tau, g^*(t, \tau, x)) \partial_t g^*(t, \tau, x) \, \mathrm{d}\tau \,.$$

Note that $\partial_t g^*(t, t, x) := \partial_t g^*(t, s, x)|_{s=t} = -b^*(t, x)$. Denote by $\zeta(t, s, x) := \partial_t g^*(t, s, x)$, and further take the partial derivative with respect to $s$ on both sides of (D.9):

$$\partial_s \zeta(t, s, x) = \nabla b^*(s, g^*(t, s, x)) \zeta(t, s, x) \,.$$

Then, it holds that

$$\partial_s \|\zeta(t, s, x)\|_2^2 = 2 \langle \zeta(t, s, x), \partial_s \zeta(t, s, x) \rangle = 2 \langle \zeta(t, s, x), \nabla b^*(s, g^*(t, s, x)) \zeta(t, s, x) \rangle \,.$$

Using Cauchy-Schwarz inequality, we have that

$$\partial_s \|\zeta(t, s, x)\|_2^2 \le 2 \|\nabla b^*(t, x)\|_{\mathrm{op}} \|\zeta(t, s, x)\|_2^2 \le 2G \|\zeta(t, s, x)\|_2^2 \,.$$

For any $t \le s \le T$, with Gronwall's inequality (Evans, 2010, Section B.2), it holds that

$$\|\zeta(t, s, x)\|_2^2 \le \|\zeta(t, t, x)\|_2^2 \exp \Big( \int_t^s 2G \, \mathrm{d}\tau \Big) \le \|b^*(x, t)\|_2^2 \exp(2G) \,.$$

Thus, we conclude that $\|\partial_t g^*(t, s, x)\|_2 \le \|b^*(x, t)\|_2 \exp(G) \le \sqrt{d} B_{\mathrm{vel}} R \exp(G)$. $\qquad \square$

*Proof of Proposition 3.5.* It sufficient to show that

$$(D.10) \qquad \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} I_d \preceq \nabla b^*(t, x) \preceq \Big( \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} + d \frac{\alpha_t \beta_t (\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t)}{(\alpha_t^2 + \sigma^2 \beta_t^2)^2} \Big) I_d.$$

For the upper bound, it follows from Lemma D.3 and Lemma D.7 that

$$\nabla b^*(t, x) \preceq \frac{\dot{\alpha}_t}{\alpha_t} I_d + \Big( \frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t} \Big) \frac{\beta_t^2}{\alpha_t^2} \Big( \frac{\sigma^2 \alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2} I_d + d \Big( \frac{\alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2} \Big)^2 I_d \Big)$$

$$(D.11) \qquad = \Big( \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} + d \frac{\alpha_t \beta_t (\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t)}{(\alpha_t^2 + \sigma^2 \beta_t^2)^2} \Big) I_d.$$

By a similar argument, we have

$$(D.12) \qquad \nabla b^*(t, x) \succeq \frac{\dot{\alpha}_t}{\alpha_t} I_d + \Big( \frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t} \Big) \frac{\beta_t^2}{\alpha_t^2} \frac{\sigma^2 \alpha_t^2}{\alpha_t^2 + \sigma^2 \beta_t^2} I_d = \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2} I_d.$$

Combining (D.11) and (D.12) yields (D.10). This competes the proof. $\qquad \square$

With the aid of above auxiliary lemmas, we provide the following proof of Corollary 3.6.

*Proof of Corollary 3.6.* We first show the Lipschitz continuity of the velocity. It is straightforward that for each $t \in (0, 1)$ and $x, x' \in \mathbb{R}^d$,

$$\|b^*(t, x) - b^*(t, x')\|_2 = \Big\| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\tau} b^*(t, x' + \tau(x - x')) \, \mathrm{d}\tau \Big\|_2$$

$$= \Big\| \int_0^1 \nabla b^*(t, x' + \tau(x - x')) \, \mathrm{d}\tau (x - x') \Big\|_2$$

$$\le \int_0^1 \|\nabla b^*(t, x' + \tau(x - x'))\|_{\mathrm{op}} \, \mathrm{d}\tau \|x - x'\|_2$$

$$(D.13) \qquad \le \Big( \int_0^1 G \, \mathrm{d}\tau \Big) \|x - x'\|_2 = G \|x - x'\|_2,$$

where the first inequality follows from the definition of the operator norm $\|\cdot\|_{\mathrm{op}}$ and Jensen's inequality, and the second inequality is due to Proposition 3.5. This shows the Lipschitz continuity of the velocity.

We now turn to focus on the Lipschitz continuity of the flow. It suffices to show that the solution at time $s$ depends Lipschitz continuously on the solution $x_t$ at time $t$. Let $x(\cdot)$ and $x'(\cdot)$ be two continuous vector-valued functions satisfy the ODE (2.4) with different values at $t$. Then it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}(x_k(t) - x'_k(t)) = b^*_k(t, x(t)) - b^*_k(t, x'(t)), \quad 1 \le k \le d.$$

By the definition of $\ell_2$-norm, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - x'(t)\|_2 = \frac{1}{2\|x(t) - x'(t)\|_2} \sum_{k=1}^{d} \frac{\mathrm{d}}{\mathrm{d}t}(x_k(t) - x'_k(t))^2$$

$$= \frac{1}{\|x(t) - x'(t)\|_2} \sum_{k=1}^{d} (x_k(t) - x'_k(t))(b^*_k(t, x(t)) - b^*_k(t, x'(t)))$$

$$\le \|b^*(t, x(t)) - b^*(t, x'(t))\|_2 \le G\|x(t) - x'(t)\|_2,$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality is due to (D.13). Then applying Gronwall's inequality (Evans, 2010, Section B.2) completes the proof. $\square$

*Proof of Proposition 3.8.* According to Lemma D.6, we find

$$\|\mathbb{E}[X_1|X_t = x]\|_2 \le M_1 R, \quad (t, x) \in (0, 1) \times \mathbb{B}_R^{\infty}, \tag{D.14}$$

where $M_2$ is a constant only depending on $d$ and $\sigma$. For the conditional covariance, using Lemma D.7 implies

$$\|\operatorname{Cov}(X_1|X_t = x)\|_{\mathrm{op}} \le M_2 \alpha_t^2, \quad (t, x) \in (0, 1) \times \mathbb{B}_R^{\infty}, \tag{D.15}$$

where $M_2$ is a constant only depending on $d$ and $\sigma$. In addition, applying Lemma D.8 yields

$$\|\mathbb{E}[X_1 X_1^T X_1|X_t = x] - \mathbb{E}[X_1^T X_1|X_t = x]\mathbb{E}[X_1|X_t]\|_2 \le M_3 \alpha_t^2 R, \tag{D.16}$$

for each $(t, x) \in (0, 1) \times \mathbb{B}_R^{\infty}$. Substituting (D.14), (D.15) and (D.16) to Lemma D.4 achieves the desired result. $\square$

## APPENDIX E. PROOF OF RESULTS IN SECTION 3.3

In the section, we present the proofs of Theorem 3.10 and Corollary 3.11.

E.1. **Proof of Theorem 3.10.** In this section, we prove Theorem 3.10. Specifically, we propose the oracle inequality in Lemma E.1, which decomposes the $L^2$-risk into approximation error, the generalization error, and the truncation error. Then we provide an approximation error bound in Lemma E.2. By making a trade-off between three errors, we finally obtain the convergence rate for the velocity estimator, which completes the proof Theorem 3.10.

Recall the weighted $L^2$-risk (3.1) of a measurable function $b : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ as

$$\mathcal{E}_T(b) = \frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t} \left[ \|b(t, X_t) - b^*(t, X_t)\|_2^2 \right] \mathrm{d}t,$$

and for the sake of notation simplicity, define the truncated $L^2$-risk with truncation parameter $R > 1$ as

$$\mathcal{E}_{T,R}(b) = \frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t} \left[ \|b(t, X_t) - b^*(t, X_t)\|_2^2 \mathbb{1}\{\|X_t\|_\infty \leq R\} \right] \mathrm{d}t.$$

**Lemma E.1** (Oracle inequality for velocity estimation)**.** *Suppose that Assumptions 1 and 2 hold. Let $T \in (1/2, 1)$ and $R \in (1, +\infty)$. Further, assume that for each $b \in \mathscr{B}$,*

(E.1) $$\max_{1 \leq k \leq d} |b_k(t, x)| \leq B_{\mathrm{vel}} R, \quad (t, x) \in [0, T] \times \mathbb{R}^d,$$

*Then the following inequality holds for each $n \geq \max_{1 \leq k \leq d} \mathrm{VCdim}(\Pi_k \mathscr{B})$,*

$$\mathbb{E}_{\mathcal{S}}[\mathcal{E}_T(\hat{b})] \leq 2 \inf_{b \in \mathscr{B}} \mathcal{E}_{T,R}(b) + C\lambda(T)R^2 \left( \max_{1 \leq k \leq d} \frac{\mathrm{VCdim}(\Pi_k \mathscr{B})}{n \log^{-1}(n)} + \frac{1}{\exp(\theta R^2)} \right),$$

*where the constant $\theta$ only depends on $\sigma$, the constant $C$ only depends on $d$ and $\sigma$, and the constant $\lambda(T)$ is defined as $\lambda(T) = \max\{1, \sup_{t \in [0,T]} \dot{\alpha}_t^2\}$.*

*Proof of Lemma E.1.* Before proceeding, we introduce some notations, aiming to reformulate the original velocity matching problem to a standard regression model.

Given a pair of random variables $(X_0, X_1)$ sampled from $\mu_0 \times \mu_1$, define a stochastic process $Y_t = \dot{\alpha}_t X_0 + \dot{\beta}_t X_1$ for each $t \in [0, T]$. Recall the stochastic interpolant $X_t = \alpha_t X_0 + \beta_t X_1$. We define the noise term as $\varepsilon_t = Y_t - b^*(t, X_t)$. Since $b^*(t, x) = \mathbb{E}[Y_t | X_t = x]$, we have $\mathbb{E}[\varepsilon_t | X_t = x] = 0$ for each $(t, x) \in [0, T] \times \mathbb{R}^d$. Therefore, in the rest of this proof, it suffices to consider the following regression model:

(E.2) $$Y_t = b^*(t, X_t) + \varepsilon_t, \quad X_t \sim \mu_t, \ t \sim \mathrm{Unif}[0, T].$$

Recall the data set $\mathcal{S} = \{(t^{(i)}, X_0^{(i)}, X_1^{(i)})\}_{i=1}^n$. Then we define the data set corresponding to the regression model (E.2) as $\{(t^{(i)}, X_t^{(i)}, Y_t^{(i)})\}$, for which

$$X_t^{(i)} = \alpha(t^{(i)}) X_0^{(i)} + \beta(t^{(i)}) X_1^{(i)} \quad \text{and} \quad Y_t^{(i)} = \dot{\alpha}(t^{(i)}) X_0^{(i)} + \dot{\beta}(t^{(i)}) X_1^{(i)}.$$

The noise terms $\{\varepsilon_t^{(i)}\}_{i=1}^n$ can be defined by $\varepsilon_t^{(i)} = Y_t^{(i)} - b^*(t^{(i)}, X_t^{(i)})$ for each $1 \leq i \leq n$.

We divide the proof into four steps.

*Step 1. Sub-Gaussian noise.*

In this step, we show that the noise term $(\varepsilon_t | X_t = x)$ in (E.2) is sub-Gaussian for each $t \in [0, T]$ and $x \in \mathbb{R}^d$, and aim to estimate its variance proxy. According to the definition of stochastic interpolant $X_t$ (2.1), we have

$$Y_t = \dot{\alpha}_t X_0 + \dot{\beta}_t X_1 = \frac{\dot{\alpha}_t}{\alpha_t} X_t + \frac{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t}{\alpha_t} X_1, \quad t \in [0, T],$$

which implies from Lemma D.5 that

$$(Y_t|X_t = x) = (\dot{\alpha}_t X_0 + \dot{\beta}_t X_1 | X_t = x) = \frac{\dot{\alpha}_t}{\alpha_t} x + \frac{\alpha_t \dot{\beta}_t - \dot{\alpha}_t \beta_t}{\alpha_t}(X_1|X_t = x)$$

$$\overset{\mathrm{d}}{=} \frac{\dot{\alpha}_t}{\alpha_t} x + \frac{\alpha_t(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)}{\alpha_t^2 + \sigma^2\beta_t^2} U_{t,x} + \sqrt{\frac{\sigma^2(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)^2}{\alpha_t^2 + \sigma^2\beta_t^2}}\epsilon + \frac{\sigma^2\beta_t(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)}{\alpha_t(\alpha_t^2 + \sigma^2\beta_t^2)}x,$$

where $U_{t,x} \in [0,1]^d$ and $\epsilon \sim N(0, I_d)$ are two independent variables. Then taking expectation on both sides of the equality yields

$$\mathbb{E}[Y_t|X_t = x] = \frac{\dot{\alpha}_t}{\alpha_t} x + \frac{\alpha_t(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)}{\alpha_t^2 + \sigma^2\beta_t^2}\mathbb{E}[U_{t,x}] + \frac{\sigma^2\beta_t(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)}{\alpha_t(\alpha_t^2 + \sigma^2\beta_t^2)}x.$$

Therefore, by the definition of the noise term, the following equality holds

$$(\varepsilon_t|X_t = x) = (Y_t|X_t = x) - \mathbb{E}[Y_t|X_t = x]$$

$$\overset{\mathrm{d}}{=} \frac{\alpha_t(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)}{\alpha_t^2 + \sigma^2\beta_t^2}(U_{t,x} - \mathbb{E}[U_{t,x}]) + \sqrt{\frac{\sigma^2(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)^2}{\alpha_t^2 + \sigma^2\beta_t^2}}\epsilon.$$

Since that the random variable $U_{t,x} \in [0,1]^d$, using Hoeffding's lemma (Mohri et al., 2018, Lemma D.1) implies that $U_{t,x}$ is 1-sub-Gaussian. Further, applying Lemma B.4 deduces that each element of $(\varepsilon_t|X_t = x)$ is sub-Gaussian for each $t \in [0, T]$ and $x \in \mathbb{R}^d$ with variance proxy

$$(\text{E.3}) \qquad \sigma_T^2 = \sup_{t \in [0,T]} \left\{ \frac{\alpha_t^2(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)^2}{(\alpha_t^2 + \sigma^2\beta_t^2)^2} + \frac{\sigma^2(\alpha_t\dot{\beta}_t - \dot{\alpha}_t\beta_t)^2}{\alpha_t^2 + \sigma^2\beta_t^2} \right\} \le C\lambda(T),$$

where $C$ is a constant only depending on $\sigma$.

*Step 2. Truncation.*

Notice that the velocity fields $b^*$ is defined on $\mathbb{R}^d$. It is necessary to restrict the original problem onto a compact subset of $\mathbb{R}^d$ by the technique of truncation. To begin with, we define the truncated population and empirical excess risks with radius $R > 1$, respectively, as

$$\mathcal{E}_{T,R}(b) = \frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t}\left[ \|b^*(t, X_t) - b(t, X_t)\|_2^2 \mathbb{1}\{\|X_t\|_\infty \le R\} \right] \mathrm{d}t,$$

$$\widehat{\mathcal{E}}_{T,R,n}(b) = \frac{1}{n} \sum_{i=1}^n \|b^*(t^{(i)}, X_t^{(i)}) - b(t^{(i)}, X_t^{(i)})\|_2^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty \le R\}.$$

The population excess risk of the estimator $\widehat{b}$ can be decomposed by

$$(\text{E.4}) \quad \mathbb{E}_\mathbb{S}[\mathcal{E}_T(\widehat{b})] \le \mathbb{E}_\mathbb{S}\left[ \mathcal{E}_T(\widehat{b}) - \mathcal{E}_{T,R}(\widehat{b}) \right] + \mathbb{E}_\mathbb{S}\left[ \sup_{b \in \mathscr{B}} \mathcal{E}_{T,R}(b) - 2\widehat{\mathcal{E}}_{T,R,n}(b) \right] + 2\mathbb{E}_\mathbb{S}\left[ \widehat{\mathcal{E}}_{T,R,n}(\widehat{b}) \right].$$

The first term in the right-hand side of (E.4) corresponds to the truncation error, which is estimated in the rest of this step. The second term of (E.4) is studied in *Step 3*. Finally, we bound the last term of (E.4) in *Step 4*.

For each hypothesis $b \in \mathscr{B}$, it follows that

$$\mathbb{E}_{X_t \sim \mu_t}\Big[\|b^*(t, X_t) - b(t, X_t)\|_2^2 \mathbb{1}\{\|X_t\|_\infty > R\}\Big]$$

$$\leq \mathbb{E}_{X_t \sim \mu_t}^{1/2}\Big[\|b^*(t, X_t) - b(t, X_t)\|_2^4\Big] \mathbb{E}_{X_t \sim \mu_t}^{1/2}\Big[\mathbb{1}\{\|X_t\|_\infty > R\}\Big]$$

$$(\text{E.5}) \qquad \leq 8\Big(\mathbb{E}_{X_t \sim \mu_t}^{1/2}\Big[\|b(t, X_t)\|_2^4\Big] + \mathbb{E}_{X_t \sim \mu_t}^{1/2}\Big[\|b^*(t, X_t)\|_2^4\Big]\Big) \Pr^{1/2}\{\|X_t\|_\infty > R\},$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality is due to the triangular inequality. The boundedness of the hypothesis (E.1) deduces

$$(\text{E.6}) \qquad \mathbb{E}_{X_t \sim \mu_t}^{1/2}\Big[\|b(t, X_t)\|_2^4\Big] \leq dB_{\text{vel}}^2 R^2.$$

Then we consider the fourth moment of $b^*(t, X_t)$ in (E.5). By using Assumption 2, we have

$$Y_t = \dot{\alpha}_t X_0 + \dot{\beta}_t X_1 \stackrel{\text{d}}{=} \dot{\alpha}_t X_0 + \dot{\beta}_t U + \sigma \dot{\beta}_t \epsilon, \quad U \sim \nu, \ \epsilon \sim N(0, I_d),$$

which implies

$$\mathbb{E}^{1/2}\Big[\|Y_t\|_2^4\Big] \leq \mathbb{E}^{1/2}\Big[(\|\dot{\alpha}_t X_0\|_2 + \|\dot{\beta}_t U\|_2 + \|\sigma \dot{\beta}_t \epsilon\|_2)^4\Big]$$

$$\leq 27\Big(\dot{\alpha}_t^4 \mathbb{E}\Big[\|X_0\|_2^4\Big] + \dot{\beta}_t^4 \mathbb{E}\Big[\|U\|_2^4\Big] + \sigma^4 \dot{\beta}_t^4 \mathbb{E}\Big[\|\epsilon\|_2^4\Big]\Big)^{1/2}$$

$$\leq 81d(\dot{\alpha}_t^2 + \dot{\beta}_t^2 + \sigma^2 \dot{\beta}_t^2) \leq C\lambda(T), \quad t \in [0, T],$$

where the second inequality holds from the triangular inequality, the last inequality follows from Lemma B.7, and the constant $C$ only depends in $d$ and $\sigma$. Consequently,

$$(\text{E.7}) \qquad \mathbb{E}_{X_t}^{1/2}\Big[\|b^*(t, X_t)\|_2^4\Big] = \mathbb{E}_{X_t}^{1/2}\Big[\|\mathbb{E}[Y_t|X_t]\|_2^4\Big] \leq \mathbb{E}^{1/2}\Big[\|Y_t\|_2^4\Big] \leq C\lambda(T),$$

where the first inequality is due to the definition of velocity, and the second inequality follows from Jensen's inequality. We next consider the tail probability of $X_t$ in (E.5). According to Assumption 2, we find

$$X_t \stackrel{\text{d}}{=} \alpha_t X_0 + \beta_t U + \sigma \beta_t \epsilon, \quad U \sim \nu_{t,x}, \ \epsilon \sim N(0, I_d),$$

which implies from Hoeffding's lemma (Mohri et al., 2018, Lemma D.1) and Lemma B.4 that $X_t$ is a $(\alpha_t^2 + \beta_t^2 + \sigma^2 \beta_t^2)$-sub-Gaussian random variable. Then it follows from Lemma B.5 that

$$(\text{E.8}) \qquad \sup_{t \in (0,1)} \Pr\{\|X_t\|_\infty > R\} \leq 2d \sup_{t \in (0,1)} \exp\Big(-\frac{R^2}{2(\alpha_t^2 + \beta_t^2 + \sigma^2 \beta_t^2)}\Big) \leq \frac{2d}{\exp(\theta R^2)},$$

where $\theta$ is a constant only depending on $\sigma$. Substituting (E.6), (E.7), and (E.8) into (E.5) yields

$$\mathbb{E}_{X_t \sim \mu_t}\Big[\|b^*(t, X_t) - b(t, X_t)\|_2^2 \mathbb{1}\{\|X_t\|_\infty > R\}\Big] \leq \frac{C\lambda(T)R^2}{\exp(\theta R^2)}, \quad t \in [0, T],$$

where $C$ is a constant only depending on $d$ and $\sigma$. As a consequence, for each hypothesis $b \in \mathscr{B}$, it follows that

$$(\text{E.9}) \qquad \mathcal{E}_T(b) - \mathcal{E}_{T,R}(b) \leq \frac{C\lambda(T)R^2}{\exp(\theta R^2)}.$$

Hence the first term in the right-hand side of (E.4) can be bounded by

$$(E.10) \qquad \mathbb{E}_{\mathbb{S}}\Big[\mathcal{E}_T(\widehat{b}) - \mathcal{E}_{T,R}(\widehat{b})\Big] \leq \frac{C\lambda(T)R^2}{\exp(\theta R^2)}.$$

We next bound another truncation term by a similar argument, which will be used in *Step 4*. It follows from Cauchy-Schwarz inequality and (E.8) that

$$(E.11) \qquad \mathbb{E}_{\mathbb{S}}\Big[(\varepsilon_{t,k}^{(i)})^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty > R\}\Big] \leq \mathbb{E}_{\mathbb{S}}^{1/2}\Big[(\varepsilon_{t,k}^{(i)})^4\Big] \Pr^{1/2}\{\|X_t^{(i)}\|_\infty > R\} \leq \frac{C\lambda(T)R^2}{\exp(\theta R^2)},$$

where $C$ is a constant only depending on $d$ and $\sigma$, and the last inequality follows from the fact that $\varepsilon_{t,k}^{(i)}$ is sub-Gaussian with variance proxy in (E.3).

*Step 3. Relate the truncated population excess risk of the estimator with its empirical counterpart.*

In this step, we prove the following inequality:

$$(E.12) \qquad \mathbb{E}_{\mathbb{S}}\Big[\sup_{b\in\mathscr{B}} \mathcal{E}_{T,R}(b) - 2\widehat{\mathcal{E}}_{T,R,n}(b)\Big] \leq CR^2 \sum_{k=1}^d \frac{\text{VCdim}(\Pi_k\mathscr{B})}{n\log^{-1}(n)},$$

where $C$ is a constant only depending on $d$ and $\sigma$, and $n \geq \text{VCdim}(\Pi_k\mathscr{B})$ for each $1 \leq k \leq d$. For simplicity of notation, we define the $k$-th term of excess risks as

$$\mathcal{E}_{T,R}^k(b) = \frac{1}{T}\int_0^T \mathbb{E}_{X_t\sim\mu_t}\Big[(b_k^*(t,X_t) - b_k(t,X_t))^2 \mathbb{1}\{\|X_t\|_\infty \leq R\}\Big]\,dt,$$

$$\widehat{\mathcal{E}}_{T,R,n}^k(b) = \frac{1}{n}\sum_{i=1}^n (b_k^*(t^{(i)}, X_t^{(i)}) - b_k(t^{(i)}, X_t^{(i)}))^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}.$$

Applying Proposition 3.2, (E.1), and Lemma I.1 yields the following inequality

$$\mathbb{E}_{\mathbb{S}}\Big[\sup_{b\in\mathscr{B}} \mathcal{E}_{T,R}^k(b) - 2\widehat{\mathcal{E}}_{T,R,n}^k(b)\Big] \leq CR^2 \frac{\text{VCdim}(\Pi_k\mathscr{B})}{n\log^{-1}(n)}, \quad 1 \leq k \leq d,$$

which implies (E.12) by summing with respect to $1 \leq k \leq d$.

*Step 4. Estimate the empirical excess risk.*

In this section, we shown the following bound for the empirical excess risk of the estimator

$$(E.13) \qquad \mathbb{E}_{\mathbb{S}}\Big[\widehat{\mathcal{E}}_{T,R,n}(\widehat{b})\Big] \leq 2\inf_{b\in\mathscr{B}} \mathcal{E}_{T,R}(b) + C\lambda(T)R^2\Big(\sum_{k=1}^d \frac{\text{VCdim}(\Pi_k\mathscr{B})}{n\log^{-1}n} + \frac{1}{\exp(\theta R^2)}\Big),$$

where $C$ is a constant only depending on $d$ and $\sigma$, and $n \geq \text{VCdim}(\Pi_k\mathscr{B})$ for each $1 \leq k \leq d$. It is straightforward that

$$\widehat{\mathcal{L}}_{T,R,n}(\widehat{b}) = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^d (\varepsilon_{t,k}^{(i)} + b_k^*(t^{(i)}, X_t^{(i)}) - \widehat{b}_k(t^{(i)}, X_t^{(i)}))^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}$$

$$= \widehat{\mathcal{E}}_{T,R,n}(\widehat{b}) + \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^d (\varepsilon_{t,k}^{(i)})^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}$$

$$+ \frac{2}{n}\sum_{i=1}^n \sum_{k=1}^d \varepsilon_{t,k}^{(i)}(b_k^*(t^{(i)}, X_t^{(i)}) - \widehat{b}_k(t^{(i)}, X_t^{(i)}))\mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}.$$

Since $\widehat{b}$ is a minimizer of $\widehat{\mathcal{L}}_{T,n}(\cdot)$ over the hypothesis class $\mathscr{B}$, it holds that $\widehat{\mathcal{L}}_{T,R,n}(\widehat{b}) \leq \widehat{\mathcal{L}}_{T,n}(\widehat{b}) \leq \widehat{\mathcal{L}}_{T,n}(b)$ for each $b \in \mathscr{B}$. Consequently,

$$
\begin{aligned}
\mathbb{E}_{\mathbb{S}}\Big[\widehat{\mathcal{E}}_{T,R,n}(\widehat{b})\Big] &\leq \mathcal{L}_T(b) - \mathbb{E}_{\mathbb{S}}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}(\varepsilon_{t,k}^{(i)})^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}\Big] \\
&\quad + 2\mathbb{E}_{\mathbb{S}}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}\varepsilon_{t,k}^{(i)}\widehat{b}_k(t^{(i)}, X_t^{(i)})\mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}\Big],
\end{aligned}
$$
(E.14)

where we used the fact that $\mathbb{E}[\widehat{\mathcal{L}}_{T,n}(b)] = \mathcal{L}_T(b)$ and $\mathbb{E}[\varepsilon_{t,k}^{(i)}b^*(t^{(i)}, X_t^{(i)})] = 0$. For the first term in the right-hand side of (E.14), we have

$$
\begin{aligned}
&\mathcal{L}_T(b) - \mathbb{E}_{\mathbb{S}}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}(\varepsilon_{t,k}^{(i)})^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}\Big] \\
&\quad = \mathcal{E}_T(b) + \mathbb{E}_{\mathbb{S}}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}(\varepsilon_{t,k}^{(i)})^2 \mathbb{1}\{\|X_t^{(i)}\|_\infty > R\}\Big] \leq \mathcal{E}_{T,R}(b) + 2\frac{C\lambda(T)R^2}{\exp(\theta R^2)},
\end{aligned}
$$
(E.15)

where the inequality follows from (E.9) and (E.11). It remains to bound the second term in the right-hand side of (E.14). Plugging Proposition 3.2 and (E.3) into Lemma I.2, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{S}}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}\varepsilon_{t,k}^{(i)}\widehat{b}_k(t^{(i)}, X_t^{(i)})\mathbb{1}\{\|X_t^{(i)}\|_\infty \leq R\}\Big] \\
&\quad \leq \frac{1}{4}\mathbb{E}_{\mathbb{S}}\Big[\widehat{\mathcal{E}}_{T,R,n}(\widehat{b})\Big] + C\lambda(T)R^2 \sum_{k=1}^{d}\frac{\mathrm{VCdim}(\Pi_k\mathscr{B})}{n\log^{-1} n},
\end{aligned}
$$
(E.16)

where $C$ is a constant only depending $d$ and $\sigma$. Substituting (E.15) and (E.16) into (E.14) yields (E.13).

Finally, plugging (E.10), (E.12), and (E.13) into (E.4) completes the proof. $\qquad\square$

**Lemma E.2** (Approximation error). *Let $T \in (1/2, 1)$ and $R \in (1, +\infty)$. Set the hypothesis class $\mathscr{B}$ as a deep neural network class, which is defined as*

$$
\mathscr{B} = \left\{ b \in N(L, S) : \begin{array}{l} \|b(t,x)\|_\infty \leq B_{\mathrm{vel}}R, \ \|\partial_t b(t,x)\|_2 \leq 3D\kappa(T)R, \\ \|\nabla b(t,x)\|_{\mathrm{op}} \leq 3G, \ (t,x) \in [0,T] \times \mathbb{R}^d \end{array} \right\},
$$

*where the depth and the width of the neural network are given, respectively, as $L = C$ and $S = CN^{d+1}$. Then the following inequality holds for each $N \in \mathbb{N}_+$,*

$$
\inf_{b\in\mathscr{B}} \mathcal{E}_{T,R}(b) \leq C\kappa^2(T)R^2 N^{-2},
$$

*where $C$ is a constant only depending on $d$ and $\sigma$.*

*Proof of Lemma E.2.* Denote by $\mathbb{B}_{R,T}^\infty = [0,T] \times \mathbb{B}_R^\infty$. According to Corollary J.6, for each element $1 \leq k \leq d$, there exists a real-valued deep neural network $b_k$ with depth $L_k =$

$\lceil \log_2(d+1) \rceil + 3$ and number of parameters $S_k = (22(d+1)+6)(N+1)^{d+1}$, such that

$$\frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t} \Big[ (b_k(t, X_t) - b_k^*(t, X_t))^2 \mathbb{1}\{ \|X_t\|_\infty \le R \} \Big] \, \mathrm{d}t$$

$$\le \|b_k - b_k^*\|_{L^\infty(\mathbb{B}_{R,T}^\infty)}^2 \le C \Big( T^2 \|\partial_t b_k^*\|_{L^\infty(\mathbb{B}_{R,T}^\infty)}^2 + R^2 \sum_{\ell=1}^d \|\partial_\ell b_k^*\|_{L^\infty(\mathbb{B}_{R,T}^\infty)}^2 \Big) N^{-2},$$

where $C$ is a constant only depending on $d$ and $\sigma$, and the first inequality follows from Hölder's inequality. Then we construct a vector-valued deep neural network $b(t, x) = (b_k(t, x))_{k=1}^d$ with depth $L = \max_{1 \le k \le d} L_k$ and number of parameters $S = \sum_{k=1}^d S_k$, such that

$$\frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t} \Big[ \|b_k(t, X_t) - b_k^*(t, X_t)\|_2^2 \mathbb{1}\{ \|X_t\|_\infty \le R \} \Big] \, \mathrm{d}t$$

$$= \sum_{k=1}^d \frac{1}{T} \int_0^T \mathbb{E}_{X_t \sim \mu_t} \Big[ (b_k(t, X_t) - b_k^*(t, X_t))^2 \mathbb{1}\{ \|X_t\|_\infty \le R \} \Big] \, \mathrm{d}t$$

$$\le C \Big( T^2 \sum_{k=1}^d \|\partial_t b_k^*\|_{L^\infty(\mathbb{B}_{R,T}^\infty)}^2 + R^2 \sum_{\ell=1}^d \sum_{k=1}^d \|\partial_\ell b_k^*\|_{L^\infty(\mathbb{B}_{R,T}^\infty)}^2 \Big) N^{-2} \le C' \kappa^2(T) R^2 N^{-2},$$

where $C$ and $C'$ are two constants only depending on $d$ and $\sigma$, and the last inequality holds from Propositions 3.5 and 3.8. This completes the proof. $\qquad \square$

*Proof of Theorem 3.10.* Set the hypothesis class as that in Lemma E.2. According to Lemma E.2, there exists a vector-valued deep neural network $b \in \mathscr{B}$ such that

$$(\text{E.17}) \qquad\qquad \inf_{b \in \mathscr{B}} \mathcal{E}_{T,R}(b) \le C \kappa^2(T) R^2 N^{-2},$$

where $C$ is a constant only depending on $d$ and $\sigma$. On the other hand, by applying Lemma B.12, the VC-dimension of this deep neural network class $\mathscr{B}$ is given as

$$(\text{E.18}) \qquad\qquad \mathrm{VCdim}(\Pi_k \mathscr{B}) \le C N^{d+1} \log N,$$

where $C$ is an absolute constant. Plugging (E.17) and (E.18) into Lemma E.1 yields

$$\mathbb{E}_\mathbb{S}[\mathcal{E}_T(\widehat{b})] \le C R^2 \Big( \frac{\kappa^2(T)}{N^2} + \frac{\lambda(T) N^{d+1} \log N}{n \log^{-1} n} + \frac{\lambda(T)}{\exp(\theta R^2)} \Big)$$

$$\le C \kappa^2(T) R^2 \Big( \frac{1}{N^2} + \frac{N^{d+1} \log N}{n \log^{-1} n} + \frac{1}{\exp(\theta R^2)} \Big),$$

where $\theta$ is a constant only depending on $\sigma$, and $C$ is a constant only depending on $d$ and $\sigma$. Here the last inequality follows from the fact that

$$\lambda(T) = \sup_{t \in [0,T]} \dot{\alpha}_t^2 \le C' \sup_{t \in [0,T]} \Big( \frac{\dot{\alpha}_t^2}{\alpha_t^2} + \frac{|\ddot{\alpha}_t|}{\alpha_t} \Big) = C' \kappa^2(T),$$

where $C'$ is a constant only depending on $\sigma$. By setting $N = C n^{\frac{1}{d+3}}$, we obtain that

$$(\text{E.19}) \qquad\qquad \mathbb{E}_\mathbb{S}[\mathcal{E}_T(\widehat{b})] \le C \kappa^2(T) R^2 \Big( n^{-\frac{2}{d+3}} \log(n) + \exp(-\theta R^2) \Big)$$

Then by substituting $R^2 = \log(n) \theta^{-1}$, we obtain the desired result. $\qquad\qquad \square$

E.2. **Proof of Corollary 3.11.** Before proceeding, recall the probability flow ODE with exact velocity field (2.4) and estimated velocity field (3.2), respectively, as

$$
\mathrm{d}Z(t) = b^*(t, Z(t))\,\mathrm{d}t, \quad t \in (0, T),
$$
$$
Z(0) = Z_0,
$$
(E.20)

and

$$
\mathrm{d}\widehat{Z}(t) = \widehat{b}(t, \widehat{Z}(t))\,\mathrm{d}t, \quad t \in (0, T),
$$
$$
\widehat{Z}(0) = Z_0.
$$
(E.21)

The following lemma bounds the particle error by the velocity error.

**Lemma E.3.** *Let $\|\mathscr{B}\|_{\mathrm{Lip}}$ be the uniform Lipschitz constant of $b \in \mathscr{B}$. Then it follows that*

$$
\|Z(T) - \widehat{Z}(T)\|_2 \le \exp(\|\mathscr{B}\|_{\mathrm{Lip}}T) \int_0^T \|b^*(t, Z(t)) - \widehat{b}(t, Z(t))\|_2 \, \mathrm{d}t.
$$

*Proof of Lemma E.3.* It is straightforward that

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z(t) - \widehat{Z}(t)\|_2^2 = \sum_{k=1}^d \frac{\mathrm{d}}{\mathrm{d}t}(Z_k(t) - \widehat{Z}_k(t))^2
$$
$$
= 2\sum_{k=1}^d (Z_k(t) - \widehat{Z}_k(t))(b_k^*(t, Z(t)) - \widehat{b}_k(t, \widehat{Z}(t)))
$$
$$
\le 2\|Z(t) - \widehat{Z}(t)\|_2 \|b^*(t, Z(t)) - \widehat{b}(t, \widehat{Z}(t))\|_2,
$$
(E.22)

where the second inequality follows from (E.20) and (E.21), and the inequality holds from Cauchy-Schwarz inequality. On the other hand, we find

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z(t) - \widehat{Z}(t)\|_2^2 = 2\|Z(t) - \widehat{Z}(t)\|_2 \frac{\mathrm{d}}{\mathrm{d}t}\|Z(t) - \widehat{Z}(t)\|_2.
$$
(E.23)

Combining (E.22) and (E.23) implies

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z(t) - \widehat{Z}(t)\|_2 \le \|b^*(t, Z(t)) - \widehat{b}(t, \widehat{Z}(t))\|_2.
$$

Then using the triangular inequality implies

$$
\frac{\mathrm{d}}{\mathrm{d}t}\|Z(t) - \widehat{Z}(t)\|_2 \le \|b^*(t, Z(t)) - \widehat{b}(t, Z(t))\|_2 + \|\widehat{b}(t, Z(t)) - \widehat{b}(t, \widehat{Z}(t))\|_2
$$
$$
\le \|b^*(t, Z(t)) - \widehat{b}(t, Z(t))\|_2 + \|\widehat{b}\|_{\mathrm{Lip}}\|Z(t) - \widehat{Z}(t)\|_2.
$$

By using Gronwall's inequality (Evans, 2010, Section B.2), we have

$$
\|Z(T) - \widehat{Z}(T)\|_2 \le \exp(\|\widehat{b}\|_{\mathrm{Lip}}T) \int_0^T \|b^*(t, Z(t)) - \widehat{b}(t, Z(t))\|_2 \, \mathrm{d}t.
$$

This completes the proof. $\qquad\square$

Then we turn to estimate the distribution error using the particle error bound derived in Lemma E.3.

*Proof of Corollary 3.11.* According to Lemma E.3, we have

$$\|Z(T) - \widehat{Z}(T)\|_2 \le \exp(\|\mathscr{B}\|_{\mathrm{Lip}}T) \int_0^T \|b^*(t, Z(t)) - \widehat{b}(t, Z(t))\|_2 \, \mathrm{d}t.$$

Taking expectation with respect to $Z_0 \sim \mu_0$ implies

$$W_2^2(\mu_T, \widehat{\mu}_T) \le \mathbb{E}_{Z_0 \sim \mu_0}\big[\|Z(T) - \widehat{Z}(T)\|_2^2\big]$$

$$\le \exp(2\|\mathscr{B}\|_{\mathrm{Lip}}T) \int_0^T \mathbb{E}_{X_t \sim \mu_t}\big[\|b^*(t, X_t) - \widehat{b}(t, X_t)\|_2^2\big] \, \mathrm{d}t,$$

where the first inequality follows from the definition of 2-Wasserstein distance and Jensen's inequality. Substituting Theorem 3.10 into the above inequality completes the proof. $\qquad\square$

## APPENDIX F. PROOFS OF RESULTS IN SECTION 3.4

In this section, we present the proof of Theorem 3.12 and Corollary 3.13. Recall Euler scheme (2.9) as

$$
\begin{aligned}
\widehat{Z}_k &= \widehat{Z}_{k-1} + \widehat{b}(t_{k-1}, \widehat{Z}_{k-1})\tau, \quad 1 \le k \le K, \\
\widehat{Z}_0 &= Z_0.
\end{aligned}
\tag{F.1}
$$

The following lemma states the discretization error of Euler method, which uses some standard techniques in the numerical analysis for the forward Euler method (Iserles, 2008, Theorem 1.1).

**Lemma F.1** (Discretization error of Euler method). *Let $\|\mathscr{B}\|_{\mathrm{Lip}}$ be the uniform Lipschitz constant of $b \in \mathscr{B}$. Then it follows that*

$$\|\widehat{Z}(T) - \widehat{Z}_K\|_2 \le \frac{T}{K} \frac{\exp(\|\mathscr{B}\|_{\mathrm{Lip}}T) - 1}{\|\mathscr{B}\|_{\mathrm{Lip}}} D\kappa(T)R.$$

*Proof.* The proof is divided into two steps.

*Step 1. Local truncation error estimate.*

Consider the Taylor expansion of $\widehat{Z}(t_{k+1})$ around $t = t_k$,

$$
\begin{aligned}
\widehat{Z}(t_{k+1}) &= \widehat{Z}(t_k) + \frac{\mathrm{d}}{\mathrm{d}t}\widehat{Z}(t_k)\tau + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\widehat{Z}(\theta)\tau^2 \\
&= \widehat{Z}(t_k) + \widehat{b}(t_k, \widehat{Z}(t_k))\tau + \partial_t\widehat{b}(\theta, \widehat{Z}(\theta))\tau^2,
\end{aligned}
\tag{F.2}
$$

where $\theta \in [t_k, t_{k+1}]$, and the second equality holds from the ODE (2.4). Recalling the forward Euler method (2.9)

$$\widehat{Z}_{k+1} = \widehat{Z}_k + \widehat{b}(t_k, \widehat{Z}_k)\tau. \tag{F.3}$$

Subtracting (F.3) from (F.2) yields

$$
\begin{aligned}
\|\widehat{Z}(t_{k+1}) - \widehat{Z}_{k+1}\|_2 &\le \|\widehat{Z}(t_k) - \widehat{Z}_k\|_2 + \|\widehat{b}(t_k, \widehat{Z}(t_k)) - \widehat{b}(t_k, \widehat{Z}_k)\|_2\tau + D\kappa(T)R\tau^2 \\
&\le (1 + \|\mathscr{B}\|_{\mathrm{Lip}}\tau)\|\widehat{Z}(t_k) - \widehat{Z}_k\|_2 + D\kappa(T)R\tau^2,
\end{aligned}
\tag{F.4}
$$

where the first inequality holds from the triangular inequality and Proposition 3.8, and $\|\mathscr{B}\|_{\mathrm{Lip}}$ denotes the Lipschitz constant of $b$ with respect to the second variable. According to the definition of the hypothesis class $\mathscr{B}$ in Theorem 3.10, we have $\|\mathscr{B}\|_{\mathrm{Lip}} \le 3G$.

*Step 2. Global truncation error estimate.*

We now show that the following inequality holds

$$\text{(F.5)} \qquad \|\widehat{Z}(t_k) - \widehat{Z}_k\|_2 \leq \frac{(1 + \|\mathscr{B}\|_{\text{Lip}}\tau)^k - 1}{\|\mathscr{B}\|_{\text{Lip}}} D\kappa(T)R\tau, \quad 1 \leq k \leq K.$$

We prove (F.5) by induction. When $k = 1$, since that $\widehat{Z}(0) = \widehat{Z}_0$, it follows from (F.4) that

$$\|\widehat{Z}(t_1) - \widehat{Z}_1\|_2 \leq \delta(t_0, \widehat{Z}(t_0))\tau + D\kappa(T)R\tau^2,$$

which satisfies (F.5). For general $k \geq 2$ we assume that (F.5) holds up to $k - 1$. Then applying (F.4) implies that

$$\|\widehat{Z}(t_k) - \widehat{Z}_k\|_2 \leq (1 + \|\mathscr{B}\|_{\text{Lip}}\tau)\|\widehat{Z}(t_{k-1}) - \widehat{Z}_{k-1}\|_2 + D\kappa(T)R\tau^2$$

$$\leq \frac{(1 + \|\mathscr{B}\|_{\text{Lip}}\tau)^k - 1}{\|\mathscr{B}\|_{\text{Lip}}} D\kappa(T)R\tau,$$

which proves that (F.5) is true for $k$. Therefore, we have verified the inequality (F.5). Substituting $K\tau = T$ and $(1 + \|\mathscr{B}\|_{\text{Lip}}\tau)^k \leq \exp(\|\mathscr{B}\|_{\text{Lip}}k\tau)$ into (F.5) completes the proof. $\qquad\square$

*Proof of Theorem 3.12.* Combining Lemmas E.3 and F.1, we have

$$\|Z(T) - \widehat{Z}_K\|_2 \leq \|Z(T) - \widehat{Z}(T)\|_2 + \|\widehat{Z}(T) - \widehat{Z}_K\|_2$$

$$\leq \exp(\|\mathscr{B}\|_{\text{Lip}}T)\Big(\int_0^T \|b^*(t, Z(t)) - \widehat{b}(t, Z(t))\|_2 \, dt + \frac{TD\kappa(T)}{K}R\Big).$$

Taking expectation with respect to $Z_0 \sim \mu_0$ implies

$$W_2^2(\mu_T, \widehat{\mu}_K) \leq \mathbb{E}_{Z_0 \sim \mu_0}\big[\|Z(T) - \widehat{Z}_K\|_2^2\big]$$

$$\leq 2\exp(2\|\mathscr{B}\|_{\text{Lip}}T)\Big(\int_0^T \mathbb{E}_{X_t \sim \mu_t}\big[\|b^*(t, X_t) - \widehat{b}(t, X_t)\|_2^2\big] \, dt + \frac{T^2 D^2 \kappa^2(T)}{K^2}R^2\Big).$$

Combining this inequality with Theorem 3.10 competes the proof. $\qquad\square$

*Proof of Corollary 3.13.* We first show that

$$\text{(F.6)} \qquad W_2(\mu_T, \mu_1) \leq \max\{\alpha_T, 1 - \beta_T\}W_2(\mu_0, \mu_1).$$

Indeed, let $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$ be two independent random variables. Then $X_T = \alpha_T X_0 + \beta_T X_1$ is a random variable obeying $\mu_T$. It follows that

$$\|X_T - X_1\|_2 = \|\alpha_T X_0 - (1 - \beta_T)X_1\|_2 \leq \max\{\alpha_T, 1 - \beta_T\}\|X_0 - X_1\|_2.$$

Taking expectation on both sides of the inequality with respect to $X_0$ and $X_1$ and recalling the definition of 2-Wasserstein distance implies (F.6).

According to the triangular inequality of the Wasserstein distance (Villani, 2009, Chapter 6), we have

$$W_2(\widehat{\mu}_K, \mu_1) \leq W_2(\widehat{\mu}_K, \mu_T) + W_2(\mu_T, \mu_1)$$

$$\leq W_2(\widehat{\mu}_K, \mu_T) + 2\max\{\alpha_t, 1 - \beta_t\}W_2(\mu_0, \mu_1),$$

where we used Lemma F.6. Combining this with Theorem 3.12 implies the desired result. $\qquad\square$

## Appendix G. Proof of Results in Section 3.5

In this section, we aim to prove Theorem 3.14. Towards this end, we first relate the averaged 2-Wasserstein distance of the characteristic generator to its $L^2$-risk by Lemma G.1. Then an oracle inequality of $L^2$-risk are proposed in Lemma G.2. Finally, by substituting approximation and generalization error bounds into the oracle inequality and using Theorem 3.12 completes the proof.

**Lemma G.1.** *Let $\widehat{g}_{t,s}$ be the estimator defined as* (2.11). *Then it follows that*

$$
\frac{2}{T^2} \int_0^T \int_t^T W_2^2\Big((\widehat{g}_{t,s})_\sharp \mu_t, \mu_s\Big) \,\mathrm{d}s\mathrm{d}t
$$
$$
\leq \mathbb{E}_{Z_0 \sim \mu_0}\Big[\frac{2}{T^2} \int_0^T \int_t^T \|g^*(t,s,Z_t) - \widehat{g}(t,s,Z_t)\|_2^2 \,\mathrm{d}s\mathrm{d}t\Big].
$$

*Proof of Lemma G.1.* According to the definition of 2-Wasserstein distance as Definition 1.3, it follows that

$$
W_2^2\Big((\widehat{g}_{t,s})_\sharp \mu_t, \mu_s\Big) \leq \mathbb{E}_{Z_0 \sim \mu_0}\Big[\|\widehat{g}(t,s,Z_t) - Z_s\|_2^2\Big].
$$

Integrating both sides of the inequality with respect to $0 \leq t \leq s \leq T$ deduces

$$
\frac{2}{T^2} \int_0^T \int_t^T W_2^2\Big((\widehat{g}_{t,s})_\sharp \mu_t, \mu_s\Big) \,\mathrm{d}s\mathrm{d}t
$$
$$
\leq \frac{2}{T^2} \int_0^T \int_t^T \mathbb{E}_{Z_0 \sim \mu_0}\Big[\|Z_s - \widehat{g}(t,s,Z_t)\|_2^2\Big] \,\mathrm{d}s\mathrm{d}t
$$
$$
= \frac{2}{T^2} \int_0^T \int_t^T \mathbb{E}_{Z_t \sim \mu_t}\Big[\|g^*(t,s,Z_t) - \widehat{g}(t,s,Z_t)\|_2^2\Big] \,\mathrm{d}s\mathrm{d}t
$$
$$
= \mathbb{E}_{Z_0 \sim \mu_0}\Big[\frac{2}{T^2} \int_0^T \int_t^T \|g^*(t,s,Z_t) - \widehat{g}(t,s,Z_t)\|_2^2 \,\mathrm{d}s\mathrm{d}t\Big],
$$

which completes the proof. $\qquad\square$

**Lemma G.2** (Oracle inequality for characteristic fitting). *Suppose Assumptions 1 and 2 hold. Let $T \in (1/2, 1)$ and $R \in (1, +\infty)$. Further, assume the hypothesis class $\mathscr{G}$ satisfies the following conditions for each $(t,x) \in [0,T] \times \mathbb{R}^d$:*

*(i)* $\sup_{1 \leq k \leq d} |g_k(t,s,x)| \leq B_{\mathrm{flow}} R$,

*(ii)* $\|\partial_t g(t,s,x)\|_2, \|\partial_s g(t,s,x)\|_2 \leq 3B_{\mathrm{vel}} R$, *and*

*(iii)* $\|\nabla g(t,s,x)\|_{\mathrm{op}} \leq 3 \exp(\|\mathscr{B}\|_{\mathrm{Lip}} T)$.

*Then the following inequality holds for each $m \geq \max_{1 \leq k \leq d} \mathrm{VCdim}(\Pi_k \mathscr{G})$,*

$$
\mathbb{E}_{\mathcal{Z}}\big[\mathcal{R}_T(\widehat{g})\big] \leq \inf_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) + CW_2^2(\widehat{\mu}_K, \mu_T)
$$
$$
+ CR^2 \max_{1 \leq k \leq d} \frac{\mathrm{VCdim}(\Pi_k \mathscr{G})}{m \log^{-1}(m)} + \frac{CR^2}{K} + \frac{CR^2}{\exp(\theta R^2)},
$$

*where the constant $\theta$ only depends on $\sigma$, and the constant $C$ only depends on $d$ and $\sigma$.*

*Proof of Lemma G.2.* Recall the set of $m$ random variables $\mathcal{Z} = \{Z_0^{(i)}\}_{i=1}^m$ i.i.d. sampled from $\mu_0$. Further, let $Z_t^{(i)}$ denote the solution of the ODE (2.4) at time $t \in [0,1]$ given initial value

$Z_0^{(i)}$ for each $1 \leq i \leq m$, and let $\widehat{Z}_k^{(i)}$ denote the solution of Euler method (2.9) given the same initial value $Z_0^{(i)}$ at time $k\tau$ for $1 \leq k \leq d$.

We first recall the empirical risk of the characteristic fitting:

$$
\begin{aligned}
\text{(G.1)} \quad \widehat{\mathcal{R}}_{T,m,K}^{\text{Euler}}(g) = \frac{2}{mK^2} \sum_{i=1}^{m} \Bigg\{ & \sum_{k=0}^{K-1} \frac{1}{2} \|\widehat{Z}_k^{(i)} - g(k\tau, k\tau, \widehat{Z}_k^{(i)})\|_2^2 \\
& + \sum_{k=0}^{K-1} \sum_{\ell=k+1}^{K-1} \|\widehat{Z}_\ell^{(i)} - g(k\tau, \ell\tau, \widehat{Z}_k^{(i)})\|_2^2 \Bigg\}.
\end{aligned}
$$

Then replacing Euler solutions $\{\widehat{Z}_k^{(i)} : 1 \leq k \leq d\}_{i=1}^{m}$ in the empirical risk by exact solutions $\{Z_k^{(i)} : 1 \leq k \leq d\}_{i=1}^{m}$ of the ODE (2.4) yields an auxiliary empirical risk

$$
\begin{aligned}
\text{(G.2)} \quad \widehat{\mathcal{R}}_{T,m,K}(g) = \frac{2}{mK^2} \sum_{i=1}^{m} \Bigg\{ & \sum_{k=0}^{K-1} \frac{1}{2} \|g^*(k\tau, k\tau, Z_{k\tau}^{(i)}) - g(k\tau, k\tau, Z_{k\tau}^{(i)})\|_2^2 \\
& + \sum_{k=0}^{K-1} \sum_{\ell=k+1}^{K-1} \|g^*(k\tau, \ell\tau, Z_{k\tau}^{(i)}) - g(k\tau, \ell\tau, Z_{k\tau}^{(i)})\|_2^2 \Bigg\}.
\end{aligned}
$$

Next we introduce a spatial truncation to (G.2), which implies the following risk:

$$
\begin{aligned}
\text{(G.3)} \quad \widehat{\mathcal{R}}_{T,R,m,K}(g) = \frac{2}{mK^2} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \Bigg\{ & \frac{1}{2} \|g^*(k\tau, k\tau, Z_{k\tau}^{(i)}) - g(k\tau, k\tau, Z_{k\tau}^{(i)})\|_2^2 \\
& + \sum_{\ell=k+1}^{K-1} \|g^*(k\tau, \ell\tau, Z_{k\tau}^{(i)}) - g(k\tau, \ell\tau, Z_{k\tau}^{(i)})\|_2^2 \Bigg\} \mathbb{1}\{\|Z_{k\tau}^{(i)}\|_\infty \leq R\}.
\end{aligned}
$$

We then define the following semi-discretized risk, which replaces the empirical average with respect to first two variables in (G.3) by its population :

$$
\text{(G.4)} \quad \widehat{\mathcal{R}}_{T,R,m}(g) = \frac{1}{m} \sum_{i=1}^{m} \Big\{ \frac{2}{T^2} \int_0^T \int_t^T \|g^*(t, s, Z_t^{(i)}) - g(t, s, Z_t^{(i)})\|_2^2 \mathbb{1}\{\|Z_t^{(i)}\|_\infty \leq R\} \, \mathrm{d}s \mathrm{d}t \Big\}.
$$

Finally, recall the population risk of the characteristic fitting

$$
\text{(G.5)} \quad \mathcal{R}_T(g) = \mathbb{E}_{Z_0 \sim \mu_0} \Big[ \frac{2}{T^2} \int_0^T \int_t^T \|g^*(t, s, Z_t) - g(t, s, Z_t)\|_2^2 \, \mathrm{d}s \mathrm{d}t \Big],
$$

of which the spatial truncated counterpart is given as

$$
\text{(G.6)} \quad \mathcal{R}_{T,R}(g) = \mathbb{E}_{Z_0 \sim \mu_0} \Big[ \frac{2}{T^2} \int_0^T \int_t^T \|g^*(t, s, Z_t) - g(t, s, Z_t)\|_2^2 \mathbb{1}\{\|Z_t\|_\infty \leq R\} \, \mathrm{d}s \mathrm{d}t \Big].
$$

According to definitions (G.1) to (G.6), it is straightforward that for each $g \in \mathcal{G}$,

$$
\begin{aligned}
\mathcal{R}_T(\widehat{g}) \leq & \Big( \mathcal{R}_T(\widehat{g}) - \mathcal{R}_{T,R}(\widehat{g}) \Big) + \Big( \mathcal{R}_{T,R}(\widehat{g}) - 2\widehat{\mathcal{R}}_{T,R,m}(\widehat{g}) \Big) + 2\Big( \widehat{\mathcal{R}}_{T,R,m}(\widehat{g}) - \widehat{\mathcal{R}}_{T,R,m,K}(\widehat{g}) \Big) \\
& + 2\Big( \widehat{\mathcal{R}}_{T,R,m,K}(\widehat{g}) - \widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) \Big) + \Big( 2\widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) - \widehat{\mathcal{R}}_{T,m,K}^{\text{Euler}}(\widehat{g}) \Big) + \widehat{\mathcal{R}}_{T,m,K}^{\text{Euler}}(g),
\end{aligned}
$$

where the inequality follows from the fact that $\widehat{g}$ is the minimizer of the empirical risk minimizer $\widehat{\mathcal{R}}_{T,m,K}^{\text{Euler}}$ over the hypothesis class $\mathcal{G}$. Taking expectation on both sides of the

inequality with respect to $\mathcal{Z} \sim \mu_0^m$ yields

$$\mathbb{E}_{\mathcal{Z}}[\mathcal{R}_T(\widehat{g})] \leq \left( \mathcal{R}_T(\widehat{g}) - \mathcal{R}_{T,R}(\widehat{g}) \right) + \mathbb{E}_{\mathcal{Z}}\left[ \sup_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) - 2\widehat{\mathcal{R}}_{T,R,m}(g) \right]$$

$$\text{(G.7)} \qquad + 2\mathbb{E}_{\mathcal{Z}}\left[ \widehat{\mathcal{R}}_{T,R,m}(\widehat{g}) - \widehat{\mathcal{R}}_{T,R,m,K}(\widehat{g}) \right] + 2\mathbb{E}_{\mathcal{Z}}\left[ \widehat{\mathcal{R}}_{T,R,m,K}(\widehat{g}) - \widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) \right]$$

$$+ \mathbb{E}_{\mathcal{Z}}\left[ 2\widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) - \widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(\widehat{g}) \right] + \inf_{g \in \mathscr{G}} \mathbb{E}_{\mathcal{Z}}\left[ \widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(g) \right],$$

where we used the inequality $\sup(a + b) \leq \sup(a) + \sup(b)$.

Up to now, the $L^2$-risk of the estimator $\widehat{g}$ is divided into six terms by (G.7). In the remainder of the proof, we bound six these error terms one by one.

(i) The first term in the right-hand side of (G.7) measures the error caused by the spatial truncation, for which

$$\text{(G.8)} \qquad \mathcal{R}_T(\widehat{g}) - \mathcal{R}_{T,R}(\widehat{g}) \leq \frac{CR^2}{\exp(\theta R^2)},$$

where the constant $\theta$ only depends on $\sigma$, and $C$ only depends on $d$ and $\sigma$.

(ii) The second term in the right-hand side of (G.7) is known as the generalization error, for which

$$\text{(G.9)} \qquad \mathbb{E}_{\mathcal{Z}}\left[ \sup_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) - 2\widehat{\mathcal{R}}_{T,R,m}(g) \right] \leq CR^2 \max_{1 \leq k \leq d} \frac{\mathrm{VCdim}(\Pi_k \mathscr{G})}{m \log^{-1}(m)},$$

where the constant $\theta$ only depends on $\sigma$, and $C$ only depends on $d$ and $\sigma$.

(iii) The third term in the right-hand side of (G.7) is led by the time discretization. It holds for each $g \in \mathscr{G}$ that

$$\text{(G.10)} \qquad \widehat{\mathcal{R}}_{T,R,m}(g) - \widehat{\mathcal{R}}_{T,R,m,K}(g) \leq \frac{CR^2}{K},$$

where the constant $\theta$ only depends on $\sigma$, and $C$ only depends on $d$ and $\sigma$.

(iv) The fourth term in the right-hand side of (G.7) is also a truncation error. By an argument similar to (G.8), it holds that

$$\text{(G.11)} \qquad \mathbb{E}_{\mathcal{Z}}\left[ \widehat{\mathcal{R}}_{T,R,m,K}(\widehat{g}) - \widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) \right] \leq \frac{CR^2}{\exp(\theta R^2)},$$

where the constant $\theta$ only depends on $\sigma$, and $C$ only depends on $d$ and $\sigma$.

(v) The fifth term in the right-hand side of (G.7) is caused by the error of Euler method, for which

$$\text{(G.12)} \qquad \mathbb{E}_{\mathcal{Z}}\left[ \widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) - 2\widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(\widehat{g}) \right] \leq CW_2^2(\widehat{\mu}_K, \mu_T),$$

where the constant $\theta$ only depends on $\sigma$, and $C$ only depends on $d$ and $\sigma$.

(vi) The sixth term in the right-hand side of (G.7) is the empirical risk of the estimator. Using the definition of the empirical risk minimizer, we deduce

$$\text{(G.13)} \qquad \mathbb{E}_{\mathcal{Z}}\left[ \widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(g) \right] \leq \inf_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) + CW_2^2(\widehat{\mu}_K, \mu_T) + \frac{CR^2}{K} + \frac{CR^2}{\exp(\theta R^2)},$$

where the constant $\theta$ only depends on $\sigma$, and $C$ only depends on $d$ and $\sigma$.

Plugging (G.8) to (G.13) into (G.7) obtains the desired result.

*Step 1. Estimate the first term in the right-hand side of* (G.7).

For each hypothesis $g \in \mathscr{G}$, by an argument similar to (E.5), we have

$$\mathbb{E}_{Z_t \sim \mu_t}\Big[\|g^*(t,s,Z_t) - g(t,s,Z_t)\|_2^2 \mathbb{1}\{\|Z_t\|_\infty > R\}\Big]$$

$$\leq \mathbb{E}_{Z_t \sim \mu_t}^{1/2}\Big[\|g^*(t,s,Z_t) - g(t,s,Z_t)\|_2^4\Big] \mathbb{E}_{Z_t \sim \mu_t}^{1/2}\Big[\mathbb{1}\{\|Z_t\|_\infty > R\}\Big]$$

$$(\text{G.14}) \qquad \leq 8\Big(\mathbb{E}_{Z_t \sim \mu_t}^{1/2}\Big[\|Z_s\|_2^4\Big] + \mathbb{E}_{Z_t \sim \mu_t}^{1/2}\Big[\|g(t,s,Z_t)\|_2^4\Big]\Big) \Pr^{1/2}\{\|Z_t\|_\infty > R\},$$

where the first inequality holds from Cauchy-Schwarz inequality, and the second inequality is due to the triangular inequality. By using Assumption 2, we have

$$Z_s \stackrel{\mathrm{d}}{=} X_s \stackrel{\mathrm{d}}{=} \alpha_s X_0 + \beta_s U + \sigma \beta_s \epsilon, \quad U \in \nu, \ \epsilon \sim N(0, I_d),$$

which implies by an argument similar to (E.7) that

$$\mathbb{E}_{Z_s}^{1/2}\Big[\|Z_s\|_2^4\Big] = \mathbb{E}_{X_s}^{1/2}\Big[\|X_s\|_2^4\Big] \leq \mathbb{E}_{(X_0, U, \epsilon)}^{1/2}\Big[(\|\alpha_s X_0\|_2 + \|\beta_s U\|_2 + \|\sigma \beta_s \epsilon\|_2)^4\Big]$$

$$\leq 27\Big(\alpha_s^4 \mathbb{E}_{X_0}\Big[\|X_0\|_2^4\Big] + \beta_s^4 \mathbb{E}_U\Big[\|U\|_2^4\Big] + \sigma^4 \beta_s^4 \mathbb{E}_\epsilon\Big[\|\epsilon\|_2^4\Big]\Big)^{1/2}$$

$$\leq 81 d(\alpha_s^2 + \beta_s^2 + \sigma^2 \beta_s^2),$$

where the last inequality follows from Lemma B.7. As a consequence,

$$(\text{G.15}) \qquad \mathbb{E}_{Z_t}^{1/2}\Big[\|g^*(t,s,Z_t)\|_2^4\Big] = \mathbb{E}_{Z_s}^{1/2}\Big[\|Z_s\|_2^4\Big] \leq C,$$

where the constant $C$ only depends on $d$ and $\sigma$. Additionally, by using the boundedness of $g \in \mathscr{G}$, we have

$$(\text{G.16}) \qquad \mathbb{E}_{Z_t}^{1/2}\Big[\|g(t,s,Z_t)\|_2^4\Big] \leq d B_{\text{flow}}^2 R^2.$$

Further, using (E.8) yields

$$(\text{G.17}) \qquad \sup_{t \in (0,1)} \Pr\{\|Z_t\|_\infty > R\} = \sup_{t \in (0,1)} \Pr\{\|X_t\|_\infty > R\} \leq \frac{2d}{\exp(\theta R^2)},$$

where $\theta$ is a constant only depending on $\sigma$. Substituting inequalities (G.15), (G.16) and (G.17) into (G.14) deduces

$$\mathbb{E}_{Z_t \sim \mu_t}\Big[\|g^*(t,s,Z_t) - \widehat{g}(t,s,Z_t)\|_2^2 \mathbb{1}\{\|Z_t\|_\infty > R\}\Big] \leq \frac{CR^2}{\exp(\theta R^2)},$$

where $C$ is a constant only depending to $d$ and $\sigma$. Finally, combining the above inequality with definitions (G.5) and (G.6) completes the proof of (G.8).

*Step 2. Estimate the second term in the right-hand side of* (G.7).

For simplicity of notation, we define the $k$-th term of $\mathcal{R}_{T,R}$ (G.6) and $\widehat{\mathcal{R}}_{T,R,m}$ (G.4), respectively, as

$$\mathcal{R}_{T,R}^k(b) = \mathbb{E}_{Z_0 \sim \mu_0}\Big[\frac{2}{T^2} \int_0^T \int_t^T (g_k^*(t,s,Z_t) - g_k(t,s,Z_t))^2 \mathbb{1}\{\|Z_t\|_\infty \leq R\} \,\mathrm{d}s\mathrm{d}t\Big],$$

$$\widehat{\mathcal{R}}_{T,R,m}^k(b) = \frac{1}{m} \sum_{i=1}^m \Big\{\frac{2}{T^2} \int_0^T \int_t^T (g_k^*(t,s,Z_t^{(i)}) - g_k(t,s,Z_t^{(i)}))^2 \mathbb{1}\{\|Z_t^{(i)}\|_\infty \leq R\} \,\mathrm{d}s\mathrm{d}t\Big\}.$$

Applying the boundedness of $g \in \mathscr{G}$, Proposition 3.3, and Lemma I.1 yields

$$\mathbb{E}_{\mathcal{Z}}\Big[\sup_{g \in \mathscr{G}} \mathcal{R}^k_{T,R}(g) - 2\widehat{\mathcal{R}}^k_{T,R,m}(g)\Big] \leq CR^2 \frac{\mathrm{VCdim}(\Pi_k \mathscr{G})}{m \log^{-1}(m)}, \quad 1 \leq k \leq d.$$

Summing the above inequalities with respect to $1 \leq k \leq d$ completes the proof of (G.9).

*Step 3. Estimate the third term in the right-hand side of* (G.7).

For each fixed $x \in \mathbb{B}^\infty_R$, we define an auxiliary function

$$u(t,s,x) = \|g^*(t,s,x) - \widehat{g}(t,s,x)\|^2_2, \quad 0 \leq t \leq s \leq T, \ x \in \mathbb{B}^\infty_R.$$

It is apparent that the following inequality holds for each $0 \leq t \leq s \leq T$ and $x \in \mathbb{B}^\infty_R$,

$$(\mathrm{G.18}) \qquad |\partial_t u(t,s,x)| \leq 2\|g^*(t,s,x) - g(t,s,x)\|_2 \|\partial_t g^*(t,s,x) - \partial_t g(t,s,x)\|_2 \leq CR^2,$$

where the first inequality follows from Cauchy-Schwarz inequality, and the last inequality used Corollaries 3.3 and 3.4, and the definition of hypothesis class $\mathscr{G}$. Here the constant $C$ only depends on $d$ and $\sigma$. By the same argument, we have

$$(\mathrm{G.19}) \qquad\qquad |\partial_s u(t,s,x)| \leq CR^2, \quad 0 \leq t \leq s \leq T, \ x \in \mathbb{B}^\infty_R.$$

Substituting (G.18) and (G.19) into Lemma G.3 yields (G.10).

*Step 4. Estimate the forth term in the right-hand side of* (G.7).

We use an argument similar to *Step 1*. For each $0 \leq k \leq \ell \leq K - 1$, it follows that

$$\mathbb{E}_{\mathcal{Z}}\Big[\|g^*(k\tau, \ell\tau, Z^{(i)}_{k\tau}) - \widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|^2_2 \mathbb{1}\{\|Z^{(i)}_{k\tau}\|_\infty > R\}\Big]$$

$$\leq 8\Big(\mathbb{E}^{1/2}_{\mathcal{Z}}\big[\|g^*(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|^4_2\big] + \mathbb{E}^{1/2}_{\mathcal{Z}}\big[\|\widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|^4_2\big]\Big)\mathbb{E}^{1/2}_{\mathcal{Z}}\big[\mathbb{1}\{\|Z^{(i)}_{k\tau}\|_\infty > R\}\big]$$

$$(\mathrm{G.20}) \quad = 8\Big(\mathbb{E}^{1/2}_{\mathcal{Z}}\big[\|Z^{(i)}_{\ell\tau}\|^4_2\big] + \mathbb{E}^{1/2}_{\mathcal{Z}}\big[\|\widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|^4_2\big]\Big)\mathrm{Pr}^{1/2}\{\|Z^{(i)}_{k\tau}\|_\infty > R\},$$

where we used Cauchy-Schwarz inequality and the triangular inequality. Substituting inequalities (G.15), (G.16) and (G.17) into (G.20) yields that for $0 \leq k \leq \ell \leq K - 1$,

$$(\mathrm{G.21}) \qquad \mathbb{E}_{\mathcal{Z}}\Big[\|g^*(k\tau, \ell\tau, Z^{(i)}_{k\tau}) - \widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|^2_2 \mathbb{1}\{\|Z^{(i)}_{k\tau}\|_\infty > R\}\Big] \leq \frac{CR^2}{\exp(\theta R^2)},$$

where $C$ is a constant only depending on $d$ and $\sigma$. Combining (G.21) with definitions (G.2) and (G.3) deduces (G.11).

*Step 5. Estimate the fifth term in the right-hand side of* (G.7).

For each $1 \leq i \leq m$ and $0 \leq k \leq \ell \leq K - 1$, it follows that

$$\|\widehat{Z}^{(i)}_\ell - \widehat{g}(k\tau, \ell\tau, \widehat{Z}^{(i)}_k)\|_2$$

$$\leq \|\widehat{Z}^{(i)}_\ell - Z^{(i)}_{\ell\tau}\|_2 + \|Z^{(i)}_{\ell\tau} - \widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|_2 + \|\widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau}) - \widehat{g}(k\tau, \ell\tau, \widehat{Z}^{(i)}_k)\|_2$$

$$\leq \|\widehat{Z}^{(i)}_\ell - Z^{(i)}_{\ell\tau}\|_2 + \|g^*(k\tau, \ell\tau, Z^{(i)}_{k\tau}) - \widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|_2 + \|\widehat{g}\|_{\mathrm{Lip}}\|\widehat{Z}^{(i)}_k - Z^{(i)}_{k\tau}\|_2,$$

where we used the triangular inequality. Squaring both sides of the inequality yields

$$\|\widehat{Z}^{(i)}_\ell - \widehat{g}(k\tau, \ell\tau, \widehat{Z}^{(i)}_k)\|^2_2$$

$$(\mathrm{G.22}) \quad \leq 4\|\widehat{Z}^{(i)}_\ell - Z^{(i)}_{\ell\tau}\|^2_2 + 4\|\widehat{g}\|^2_{\mathrm{Lip}}\|\widehat{Z}^{(i)}_k - Z^{(i)}_{k\tau}\|^2_2 + 2\|g^*(k\tau, \ell\tau, Z^{(i)}_{k\tau}) - \widehat{g}(k\tau, \ell\tau, Z^{(i)}_{k\tau})\|^2_2.$$

Substituting (G.22) into (G.1) deduces

$$
\begin{aligned}
\widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(\widehat{g}) \leq{} & \frac{2}{mK^2} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \Big\{ \frac{1}{2} \Big( 4\|\widehat{Z}_k^{(i)} - Z_{k\tau}^{(i)}\|_2^2 + 4\|\widehat{g}\|_{\mathrm{Lip}}^2 \|\widehat{Z}_k^{(i)} - Z_{k\tau}^{(i)}\|_2^2 \Big) \\
& + \sum_{\ell=k+1}^{K-1} \Big( 4\|\widehat{Z}_\ell^{(i)} - Z_{\ell\tau}^{(i)}\|_2^2 + 4\|\widehat{g}\|_{\mathrm{Lip}}^2 \|\widehat{Z}_k^{(i)} - Z_{k\tau}^{(i)}\|_2^2 \Big) \Big\} + 2\widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) \\
\leq{} & 4(1 + \|\widehat{g}\|_{\mathrm{Lip}}^2) \frac{1}{m} \sum_{i=1}^{m} \|\widehat{Z}_K^{(i)} - Z_T^{(i)}\|_2^2 + 2\widehat{\mathcal{R}}_{T,m,K}(\widehat{g}),
\end{aligned}
$$

where the last inequality holds from the fact that

$$
\|\widehat{Z}_k^{(i)} - Z_{k\tau}^{(i)}\|_2 \leq \|\widehat{Z}_K^{(i)} - Z_T^{(i)}\|_2, \quad 0 \leq k \leq K-1,
$$

which has been shown in the proof of Theorem 3.12. Consequently,

$$
\widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(\widehat{g}) - 2\widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) \leq 4(1 + \|\widehat{g}\|_{\mathrm{Lip}}^2) \frac{1}{m} \sum_{i=1}^{m} \|\widehat{Z}_K^{(i)} - Z_T^{(i)}\|_2^2.
$$

Taking expectation on both sides of the above inequality with respect to $\mathcal{Z}$ and plugging $\|\widehat{g}\|_{\mathrm{Lip}} \leq 3\exp(GT)$ imply

$$
(\mathrm{G.23}) \qquad \mathbb{E}_{\mathcal{Z}}\Big[ \widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(\widehat{g}) - 2\widehat{\mathcal{R}}_{T,m,K}(\widehat{g}) \Big] \leq C W_2^2(\widehat{\mu}_K, \mu_T),
$$

where $C$ is a constant only depending on $d$ and $\sigma$. By the same argument as inequality (G.23), we can obtain (G.12) immediately.

*Step 6. Estimate the sixth term in the right-hand side of* (G.7).

For each fixed $g \in \mathscr{G}$ independent of $\mathcal{Z}$, it follows that

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Z}}\Big[ \widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(g) \Big] ={} & \mathbb{E}_{\mathcal{Z}}\Big[ \widehat{\mathcal{R}}_{T,m,K}^{\mathrm{Euler}}(g) - 2\widehat{\mathcal{R}}_{T,m,K}(g) \Big] + 2\mathbb{E}_{\mathcal{Z}}\Big[ \widehat{\mathcal{R}}_{T,m,K}(g) - \widehat{\mathcal{R}}_{T,R,m,K}(g) \Big] \\
& + 2\mathbb{E}_{\mathcal{Z}}\Big[ \widehat{\mathcal{R}}_{T,R,m,K}(g) - \widehat{\mathcal{R}}_{T,R,m}(g) \Big] + 2\mathbb{E}_{\mathcal{Z}}\Big[ \widehat{\mathcal{R}}_{T,R,m}(g) \Big],
\end{aligned}
$$

where the first term can be estimated by (G.23), the second and third terms can be bounded by an argument similar to (G.11) and (G.10), respectively. For the last term, we have $\mathbb{E}_{\mathcal{Z}}[\widehat{\mathcal{R}}_{T,R,m}(g)] = \mathcal{R}_{T,R}(g)$. Combining above results yields (G.13). $\qquad\square$

*Proof of Theorem 3.14.* According to Lemma E.2, the following inequality holds

$$
(\mathrm{G.24}) \qquad \inf_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) \leq \frac{CR^2}{N^2},
$$

where $C$ is a constant only depending on $d$ and $\sigma$. On the other hand, by applying Lemma B.12, the VC-dimension of this deep neural network class $\mathscr{G}$ is given as

$$
(\mathrm{G.25}) \qquad \mathrm{VCdim}(\Pi_k \mathscr{G}) \leq C N^{d+2} \log N,
$$

where $C$ is an absolute constant. Plugging (G.24) and (G.25) into Lemma G.2 yields

$$
\mathbb{E}_{\mathcal{Z}}[\mathcal{R}_T(\widehat{g})] \leq \frac{CR^2}{N^2} + C W_2^2(\widehat{\mu}_K, \mu_T) + CR^2 \max_{1 \leq k \leq d} \frac{N^{d+2} \log N}{m \log^{-1}(m)} + \frac{CR^2}{K} + \frac{CR^2}{\exp(\theta R^2)},
$$

where $C$ is a constant only depending to $d$ and $\sigma$. By setting $N = Cm^{\frac{1}{d+4}}$ and $R^2 = \log(m)\theta^{-1}$, we have

$$\mathbb{E}_{\mathcal{Z}}[\mathcal{R}_T(\widehat{g})] \leq Cm^{-\frac{2}{d+4}}\log^2(m) + CW_2^2(\widehat{\mu}_K, \mu_T) + \frac{C\log(m)}{K},$$

Finally, we relate $\mathcal{R}_T(\widehat{g})$ to $\mathcal{D}(\widehat{g})$ by Lemma G.1. Finally, using Theorem 3.12 completes the proof. $\qquad\square$

We conclude this section by giving an error bound for 2-dimensional numerical integral.

**Lemma G.3.** *Let $T > 0$ and $K \in \mathbb{N}_+$. Assume that $u \in W^{1,\infty}([0,T]^2)$. Define the step size as $\tau = T/K$, and define $\{t_\ell = \ell\tau\}_{\ell=0}^{K}$ as the set of time points. Then it follows that*

$$\frac{T^2}{K^2}\sum_{k=1}^{K}\left\{\frac{1}{2}u(t_{k-1}, t_{k-1}) + \sum_{\ell=k+1}^{K}u(t_{k-1}, t_{\ell-1})\right\} - \int_0^T\int_t^T u(s,t)\,\mathrm{d}s\mathrm{d}t$$

$$\leq (\|\partial_t u\|_{L^\infty([0,T]^2)} + \|\partial_s u\|_{L^\infty([0,T]^2)})\frac{T}{K}.$$

*Proof of Lemma G.3.* According to the Taylor expansion of $u(s,t)$ around $(t_{k-1}, t_{\ell-1})$ with $1 \leq k \leq \ell \leq K$, it follows that

(G.26) $\qquad u(t,s) = u(t_{k-1}, t_{\ell-1}) + \partial_t u(\theta_{k-1}^t, t_{\ell-1})(t - t_{k-1}) + \partial_s u(t_{k-1}, \theta_{\ell-1}^s)(s - t_{\ell-1}),$

where $\theta_{k-1}^t \in [t_{k-1}, t]$ and $\theta_{\ell-1}^s \in [t_{\ell-1}, s]$.

For $1 \leq \ell = k \leq K$, integrating both sides of (G.26) on $(t,s) \in [t_{k-1}, t_k] \times [t, t_k]$ yields

$$\int_{t_{k-1}}^{t_k}\int_t^{t_k} u(t,s)\,\mathrm{d}s\mathrm{d}t - u(t_{k-1}, t_{k-1})\frac{\tau^2}{2}$$

$$= \partial_t u(\theta_{k-1}^t, t_{k-1})\int_{t_{k-1}}^{t_k}(t_k - t)(t - t_{k-1})\,\mathrm{d}t + \partial_s u(t_{k-1}, \theta_{k-1}^s)\int_{t_{k-1}}^{t_k}\int_t^{t_k}(s - t_{k-1})\,\mathrm{d}s\mathrm{d}t$$

$$= \partial_t u(\theta_{k-1}^t, t_{k-1})\frac{\tau^3}{6} + \partial_s u(t_{k-1}, \theta_{k-1}^s)\frac{\tau^3}{3},$$

where we used the fact that $t_k - t_{k-1} = \tau$. By summing both sides of the above equality with respect to $k = 1, \ldots, K$, we obtain

$$\frac{T^2}{2K^2}\sum_{k=1}^{K}u(t_{k-1}, t_{k-1}) - \sum_{k=1}^{K}\int_{t_{k-1}}^{t_k}\int_t^{t_k} u(t,s)\,\mathrm{d}s\mathrm{d}t$$

(G.27) $$\leq \sup_{(t,s)\in[0,T]^2}\left\{|\partial_t u(t,s)| + |\partial_s u(t,s)|\right\}\frac{T^3}{3K^2}.$$

For $1 \leq \ell < k \leq K$, integrating both sides of (G.26) on $(t,s) \in [t_{k-1}, t_k] \times [t_{\ell-1}, t_\ell]$ yields

$$\int_{t_{k-1}}^{t_k}\int_{t_{\ell-1}}^{t_\ell} u(t,s)\,\mathrm{d}s\mathrm{d}t - u(t_{k-1}, t_{\ell-1})\tau^2$$

$$= \partial_t u(\theta_{k-1}^t, t_{\ell-1})\tau\int_{t_{k-1}}^{t_k}(t - t_{k-1})\,\mathrm{d}t + \partial_s u(t_{k-1}, \theta_{\ell-1}^s)\tau\int_{t_{\ell-1}}^{t_\ell}(s - t_{\ell-1})\,\mathrm{d}s$$

$$= \partial_t u(\theta_{k-1}^t, t_{\ell-1})\frac{\tau^3}{2} + \partial_s u(t_{k-1}, \theta_{\ell-1}^s)\frac{\tau^3}{2}.$$

By a similar argument to (G.27), it follows that

$$\frac{T^2}{K^2} \sum_{k=1}^{K} \sum_{\ell=k+1}^{K} u(t_{k-1}, t_{\ell-1}) - \sum_{k=1}^{K} \sum_{\ell=k+1}^{K} \int_{t_{k-1}}^{t_k} \int_{t_{\ell-1}}^{t_\ell} u(t,s) \, \mathrm{d}s \mathrm{d}t$$

(G.28)
$$\leq \sup_{(t,s)\in[0,T]^2} \left\{ |\partial_t u(t,s)| + |\partial_s u(t,s)| \right\} \frac{T^3}{4K}.$$

Summing (G.27) and (G.28) completes the proof. □

## APPENDIX H. PROOF OF RESULTS IN SECTION 3.7

First, analogous to the construction of $b^*(t,x)$ and $g^*(t,s,x)$ under Assumptions 1 and 2, we construct a $d^*$-dimensional velocity field $\widetilde{b}^*(t,\widetilde{x})$, and its corresponding ODE flow $\widetilde{g}^*(t,s,\widetilde{x})$.

**Definition H.1.** Let $\widetilde{X}_0 \sim \widetilde{\mu}_0 = N(0, I_{d^*})$ be independent of $\widetilde{X}_1 \sim \widetilde{\mu}_1 = N(0, \sigma^2 I_{d^*}) * \widetilde{\nu}$, where $\widetilde{\nu}$ is from Assumption 3. Let $\widetilde{X}_t = \alpha_t \widetilde{X}_0 + \beta_t \widetilde{X}_1$. Then, we define

$$\widetilde{b}^*(t,\widetilde{x}) := \mathbb{E}[\dot{\alpha}_t \widetilde{X}_0 + \dot{\beta}_t \widetilde{X}_1 | \widetilde{X}_t = \widetilde{x}], \quad \widetilde{x} \in \mathbb{R}^{d^*}.$$

Also, we define the ODE flow corresponding to $\widetilde{b}^*(t,\widetilde{x})$ as $\widetilde{g}^*(t,s,\widetilde{x}), 0 \leq t < s < 1$.

By directly transferring the regularity estimates for $b^*$ and $g^*$ from Section III.B of the main text, we obtain the following regularity characterization for $\widetilde{b}^*$ and $\widetilde{g}^*$. Here, we denote by $\widetilde{\mathbb{B}}_R^\infty$ the $\ell_\infty$ ball in $\mathbb{R}^{d^*}$ with radius $R$.

**Lemma H.2.** Let $T \in (1/2, 1)$, $R \in (1, +\infty)$ and $(t, \widetilde{x}) \in [0,T] \times \widetilde{\mathbb{B}}_R^\infty$. Then, it holds that

$$\max_{1 \leq k \leq d^*} |\widetilde{b}_k^*(t,\widetilde{x})| \leq \widetilde{B}_{\mathrm{vel}} R; \quad \|\nabla \widetilde{b}^*(t,x)\|_{\mathrm{op}} \leq \widetilde{G}; \quad \|\partial_t \widetilde{b}^*(t,\widetilde{x})\|_2 \leq \widetilde{D}\kappa(T)R,$$

where the constants $\widetilde{B}_{\mathrm{vel}}$, $\widetilde{G}$ and $\widetilde{D}$ only depend on $d^*$ and $\sigma$, and $\kappa(T)$ is from Proposition 3.8.

**Lemma H.3.** Let $T \in (1/2, 1)$, $R \in (1, +\infty)$, $0 \leq t \leq s \leq T$ and $x \in \widetilde{\mathbb{B}}_R^\infty$. Then, it holds that

$$\max_{1 \leq k \leq d^*} |\widetilde{g}_k^*(t,s,\widetilde{x})| \leq \widetilde{B}_{\mathrm{flow}} R; \qquad \|\nabla \widetilde{g}^*(t,s,\widetilde{x})\|_{\mathrm{op}} \leq \exp(\widetilde{G});$$

$$\max \left\{ \|\partial_t \widetilde{g}^*(t,s,\widetilde{x})\|_2, \|\partial_s \widetilde{g}^*(t,s,\widetilde{x})\|_2 \right\} \leq \widetilde{B}'_{\mathrm{flow}} R,$$

where the constants $\widetilde{B}_{\mathrm{flow}}$ and $\widetilde{B}'_{\mathrm{flow}}$ only depend on $d^*$ and $\sigma$, and $\widetilde{G}$ is from Lemma H.2.

Under Assumptions 1 and 3, with the aid of $\widetilde{b}^*$ and $\widetilde{g}^*$, we can decompose $b^*$ as $b^* = b_1^* + b_2^*$ and $g^*$ as $g^* = g_1^* + g_2^*$, where $b_1^*$ and $g_1^*$ are directly related to $\widetilde{b}^*$ and $\widetilde{g}^*$, respectively, and can be regarded as the components of $b^*$ and $g^*$ in the space $W$, which are essentially functions of dimension $d^*$. On the other hand, $b_2^*$ and $g_2^*$ can be considered as the components of $b^*$ and $g^*$ in the orthogonal space $W^\perp$, and can be approximated by small-scale neural networks.

**Lemma H.4.** *Under Assumptions 1 and 3, the velocity field $b^*(t, x)$ and the probabilistic ODE flow $g^*(t, s, x)$ have the following decomposition:*

$$(H.1) \qquad b^*(t, x) = P\widetilde{b}^*(t, P^\top x) + \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2}(I_d - PP^\top)x \,,$$

$$(H.2) \qquad g^*(t, s, x) = P\widetilde{g}^*(t, s, P^\top x) + \sqrt{\frac{\alpha_s^2 + \sigma^2 \beta_s^2}{\alpha_t^2 + \sigma^2 \beta_t^2}}(I_d - PP^\top)x \,,$$

*where $P \in \mathbb{R}^{d \times d^*}$ is from Assumption 3, $\widetilde{b}^*(t, \widetilde{x})$ and $\widetilde{g}^*(t, s, \widetilde{x})$ are from Definition H.1.*

*Proof of Lemma H.4.* By Lemma D.2 and (D.1), it holds that

$$(H.3) \qquad b^*(t, x) = \frac{\dot{\beta}_t}{\beta_t}x + \alpha_t^2 \Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\nabla_x \log \rho_t(x) \,,$$

where $X_t \sim \rho_t(x)$. Letting $\widetilde{X}_t \sim \widetilde{\rho}_t(\widetilde{x})$, we analogously get that

$$(H.4) \qquad \widetilde{b}^*(t, \widetilde{x}) = \frac{\dot{\beta}_t}{\beta_t}\widetilde{x} + \alpha_t^2 \Big(\frac{\dot{\beta}_t}{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\Big)\nabla_{\widetilde{x}} \log \widetilde{\rho}_t(\widetilde{x}) \,, \quad \widetilde{x} \in \widetilde{\mathbb{R}} \,.$$

By Assumption 3,

$$(H.5) \qquad X_t = \alpha_t X_0 + \beta_t X_1 \stackrel{d}{=} \alpha_t X_0 + \beta_t(U + \sigma Z')$$

$$\stackrel{d}{=} \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}Z + \beta_t U = \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}Z + \beta_t P\widetilde{U} \,.$$

where $\widetilde{U} \sim \widetilde{\nu}$, $U = P\widetilde{U} \sim \nu$, and $Z, Z' \sim N(0, I_d)$ are independent of $U$, $\widetilde{U}$ and $X_0$. Note that we have applied the sum property of independent Gaussian random variables in the third equality above. By (H.5), we expand $\rho_t(x)$ as follows:

$$\rho_t(x) = \frac{1}{\sqrt{2\pi}^d \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^d} \int \exp\Big\{\frac{-\|x - \beta_t u\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)}\Big\}p(u)\,\mathrm{d}u$$

$$= \frac{1}{\sqrt{2\pi}^d \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^d} \int \exp\Big\{\frac{-\|x - \beta_t u\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)}\Big\}\Big(\int \rho_{u|\widetilde{u}}(u|\widetilde{u})q(\widetilde{u})\,\mathrm{d}\widetilde{u}\Big)\,\mathrm{d}u$$

$$= \frac{1}{\sqrt{2\pi}^d \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^d} \int \Big(\int \exp\Big\{\frac{-\|x - \beta_t u\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)}\Big\}\delta_{P\widetilde{u}}(u)\,\mathrm{d}u\Big)q(\widetilde{u})\,\mathrm{d}\widetilde{u}$$

$$= \frac{1}{\sqrt{2\pi}^d \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^d} \int \exp\Big\{\frac{-\|x - \beta_t P\widetilde{u}\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)}\Big\}q(\widetilde{u})\,\mathrm{d}\widetilde{u} \,,$$

where we have used the conditional density of $U$ given $\widetilde{U} = \widetilde{u}$ is $\delta_{P\widetilde{u}}(u)$. Here, $\delta_a(x)$ denotes the $d$-dimensional Dirac delta density concentrated at $a$. Since we have $x = (I_d - PP^\top)x +$

$PP^\top x$, and $\|Py\|_2 = \|y\|_2$ for any $y \in \mathbb{R}^{d^*}$, it further holds that

$$\rho_t(x) = \frac{1}{\sqrt{2\pi}^d \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^d} \int \exp\left\{ \frac{-\|(I_d - PP^\top)x + PP^\top x - \beta_t P\widetilde{u}\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)} \right\} q(\widetilde{u}) \, \mathrm{d}\widetilde{u}$$

$$= \frac{1}{\sqrt{2\pi}^d \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^d} \int \exp\left\{ \frac{-\|(I_d - PP^\top)x\|_2^2 - \|P^\top x - \beta_t \widetilde{u}\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)} \right\} q(\widetilde{u}) \, \mathrm{d}\widetilde{u}$$

$$= \widetilde{\rho}_t(P^\top x) \cdot \frac{1}{\sqrt{2\pi}^{d-d^*} \sqrt{\alpha_t^2 + \sigma^2 \beta_t^2}^{d-d^*}} \exp\left\{ \frac{-\|(I_d - PP^\top)x\|_2^2}{2(\alpha_t^2 + \sigma^2 \beta_t^2)} \right\}.$$

Thus, we have

(H.6) $$\nabla_x \log \rho_t(x) = P \nabla_{\widetilde{x}} \log \widetilde{\rho}_t(P^\top x) - \frac{(I_d - PP^\top)x}{\alpha_t^2 + \sigma^2 \beta_t^2}.$$

Combining (H.3), (H.4) and (H.6), we conclude that

(H.7) $$b^*(t, x) = P\widetilde{b}^*(t, P^\top x) + \frac{\alpha_t \dot\alpha_t + \sigma^2 \beta_t \dot\beta_t}{\alpha_t^2 + \sigma^2 \beta_t^2}(I_d - PP^\top)x.$$

Now, Let

$$b_1(t, x) = P\widetilde{b}^*(t, P^\top x), \quad b_2(t, x) = \frac{\alpha_t \dot\alpha_t + \sigma^2 \beta_t \dot\beta_t}{\alpha_t^2 + \sigma^2 \beta_t^2} x.$$

Denote by $g_1(t, s, x)$ and $g_2(t, s, x)$ the ODE flows corresponding to $b_1(t, x)$ and $b_2(t, x)$, respectively. Note that $g_2(t, s, x)$ has an explicit expression, namely:

$$g_2(t, s, x) = x \cdot \exp\left\{ \int_t^s \frac{\alpha_\tau \dot\alpha_\tau + \sigma^2 \beta_\tau \dot\beta_\tau}{\alpha_\tau^2 + \sigma^2 \beta_\tau^2} \, \mathrm{d}\tau \right\} = \sqrt{\frac{\alpha_s^2 + \sigma^2 \beta_s^2}{\alpha_t^2 + \sigma^2 \beta_t^2}} x.$$

Let $x(\tau)$ be any solution trajectory of the ODE $\mathrm{d}x(\tau) = b^*(\tau, x(\tau)) \, \mathrm{d}\tau$, where $0 \le \tau \le 1$. For any given $0 \le t \le s \le 1$, it holds that

(H.8) $$g^*(t, s, x(t)) = x(s) = PP^\top x(s) + (I_d - PP^\top)x(s).$$

By (H.7) and $P^\top P = I_{d^*}$, we have

$$PP^\top x(s) - PP^\top x(t) = PP^\top \int_t^s b^*(\tau, x(\tau)) \, \mathrm{d}\tau$$

$$= PP^\top \int_t^s \left[ P\widetilde{b}^*(\tau, P^\top x(\tau)) + \frac{\alpha_\tau \dot\alpha_\tau + \sigma^2 \beta_\tau \dot\beta_\tau}{\alpha_\tau^2 + \sigma^2 \beta_\tau^2}(I_d - PP^\top)x(\tau) \right] \mathrm{d}\tau$$

$$= \int_t^s P\widetilde{b}^*(\tau, P^\top(PP^\top x(\tau))) \, \mathrm{d}\tau = \int_t^s b_1(\tau, PP^\top x(\tau)) \, \mathrm{d}\tau,$$

which means $PP^\top x(\tau)$ is a solution trajectory of the ODE $\mathrm{d}\bar{x}(\tau) = b_1(\tau, \bar{x}(\tau)) \, \mathrm{d}\tau$. Thus, it holds that

$$PP^\top x(s) = g_1(t, s, PP^\top x(t)).$$

Similarly, we can obtain

$$(I_d - PP^\top)x(s) = g_2(t, s, (I_d - PP^\top)x(t)) = \sqrt{\frac{\alpha_s^2 + \sigma^2 \beta_s^2}{\alpha_t^2 + \sigma^2 \beta_t^2}}(I_d - PP^\top)x(t).$$

Combining these with (H.8), and due to the arbitrariness of $x(t)$, we can infer that

$$(\text{H.9}) \qquad g^*(t, s, x) = g_1(t, s, PP^\top x) + \sqrt{\frac{\alpha_s^2 + \sigma^2 \beta_s^2}{\alpha_t^2 + \sigma^2 \beta_t^2}}(I_d - PP^\top)x.$$

Further, let $\widetilde{x}(\tau)$ be any solution trajectory of the ODE $\mathrm{d}\widetilde{x}(\tau) = \widetilde{b}^*(\tau, \widetilde{x}(\tau))\,\mathrm{d}\tau, 0 \leq \tau \leq 1$. For any given $0 \leq t \leq s \leq 1$, it holds that

$$P\widetilde{x}(s) - P\widetilde{x}(t) = P\int_t^s \widetilde{b}^*(\tau, \widetilde{x}(\tau))\,\mathrm{d}\tau = \int_t^s P\widetilde{b}^*(\tau, P^\top(P\widetilde{x}(\tau)))\,\mathrm{d}\tau = \int_t^s b_1(\tau, P\widetilde{x}(\tau))\,\mathrm{d}\tau,$$

which means $P\widetilde{x}(\tau)$ is a solution trajectory of the ODE $\mathrm{d}\bar{x}(\tau) = b_1(\tau, \bar{x}(\tau))\,\mathrm{d}\tau$. Thus,

$$g_1(t, s, P\widetilde{x}(t)) = P\widetilde{x}(s) = P\widetilde{g}^*(t, s, \widetilde{x}(t)) \implies g_1(t, s, P\widetilde{x}) = P\widetilde{g}^*(t, s, \widetilde{x}), \quad \forall \widetilde{x} \in \mathbb{R}^{d^*}.$$

For any $x \in \mathbb{R}^d$, let $\widetilde{x} = P^\top x$. Then, we have

$$g_1(t, s, PP^\top x) = P\widetilde{g}^*(t, s, P^\top x).$$

Combing this with (H.9) concludes the proof. $\qquad\square$

Lemma H.4 reveals a low-dimensional structure in $b^*$ and $g^*$, enabling neural network approximations with complexity depending on the intrinsic dimension $d^*$ rather than the ambient dimension $d$. We first establish the approximation result for the manifold-structured velocity $b^*$.

**Lemma H.5** (Approximation under manifold-structured velocity). *Let $T \in (1/2, 1)$ and $R \in (1, +\infty)$. Suppose $b^*$ admits the decomposition in* (H.1):

$$(\text{H.10}) \qquad b^*(t, x) = P\widetilde{b}^*(t, P^\top x) + \gamma(t)(I_d - PP^\top)x,$$

*where $P \in \mathbb{R}^{d \times d^*}$ has orthonormal columns, $\widetilde{b}^*$ satisfies Lemma H.2, and $\gamma(t) := \frac{\alpha_t \dot{\alpha}_t + \sigma^2 \beta_t \dot{\beta}_t}{\alpha_t^2 + \sigma^2 \beta_t^2}$. Set the hypothesis class $\mathscr{B}$ as a deep neural network class, which is defined as*

$$\mathscr{B} = \left\{ b \in N(L, S) : \begin{array}{l} \|b(t, x)\|_\infty \leq B_{\mathrm{man}}R, \ \|\partial_t b(t, x)\|_2 \leq D_{\mathrm{man}}R, \\ \|\nabla b(t, x)\|_{\mathrm{op}} \leq G_{\mathrm{man}}, \ (t, x) \in [0, T] \times \mathbb{R}^d \end{array} \right\},$$

*where the depth and the number of nonzero weights of the neural network are given, respectively, as $L = C$ and $S = CN^{d^*+1}$. Then the following inequality holds for each $N \in \mathbb{N}_+$,*

$$(\text{H.11}) \qquad \inf_{b \in \mathscr{B}} \mathcal{E}_{T,R}(b) \ \leq \ C\,\kappa^2(T)\,R^2\,N^{-2}.$$

*Using Lemma H.2, we have admissible choices for $B_{\mathrm{man}}$, $D_{\mathrm{man}}$ and $G_{\mathrm{man}}$ as follows:*

$$B_{\mathrm{man}} = \sqrt{d}(\widetilde{B}_{\mathrm{vel}}\sqrt{d^*} + \gamma_{\max}), \quad D_{\mathrm{man}} = 3\sqrt{d}(\widetilde{D}\,\kappa(T) + \gamma_{\max}), \quad G_{\mathrm{man}} = 3(\widetilde{G} + \gamma_{\max})$$

*with $\gamma_{\max} := \max\{\|\gamma\|_{L^\infty([0,T])}, \|\dot{\gamma}\|_{L^\infty([0,T])}\}$ finite by Condition 1. And $C$ denotes a constant may only depend on $d^*$, $d$ and $\sigma$.*

*Proof of Lemma H.5.* Write $b_1^*(t, x) := P\widetilde{b}^*(t, P^\top x)$, $b_2^*(t, x) := \gamma(t)(I_d - PP^\top)x$ and $b^*(t, x) = b_1^*(t, x) + b_2^*(t, x)$. We complete the proof through following three steps. Note that for a vector function $f : \mathbb{R}^d \to \mathbb{R}^d$ and a bounded set $K \in \mathbb{R}^d$, its $L^\infty(K)$-norm is denoted as $\|f\|_{L^\infty(K)} := \|\|f\|_2\|_{L^\infty(K)}$.

*Step 1 (Approximate $b_1^*$).* Set $R' := \sqrt{d}\,R$. Denote by $\widetilde{\mathbb{B}}_{R',T} := [0,T] \times \widetilde{\mathbb{B}}_{R'}^\infty$. Under Lemma H.2, we can apply Lemma E.2 to construct a vector-valued neural network $\widetilde{b} \in N(L,S)$, $\widetilde{b} : \mathbb{R}^{d^*} \to \mathbb{R}^{d^*}$ with depth $L = C_1$ and number of nonzero weights $S = C_2\,N^{d^*+1}$, such that

$$\|\widetilde{b} - \widetilde{b}^*\|_{L^\infty(\widetilde{\mathbb{B}}_{R',T})} \le C_3\kappa(T)R'N^{-1} = C_3\sqrt{d}\kappa(T)RN^{-1}.$$

Now, set $b_1(t,x) := P\widetilde{b}(t, P^\top x)$. The lifts $x \mapsto P^\top x$ and $u \mapsto Pu$ are affine with $O(dd^*)$ weights. Since $\|P^\top x\|_\infty \le R'$ for $\|x\|_\infty \le R$ and $\|P\widetilde{x}\|_2 = \|\widetilde{x}\|_2$, we have

$$\|b_1 - b_1^*\|_{L^\infty(\mathbb{B}_{R,T})} = \|P\widetilde{b}(\cdot, P^\top\cdot) - P\widetilde{b}^*(\cdot, P^\top\cdot)\|_{L^\infty(\mathbb{B}_{R,T})} \le \|\widetilde{b} - \widetilde{b}^*\|_{L^\infty(\widetilde{\mathbb{B}}_{R',T})}.$$

Here, $C_1$, $C_2$ and $C_3$ are constants only depending on $d^*$ and $\sigma$. Moreover, since $\|P\widetilde{x}\|_\infty \le R\sqrt{d^*}$ for $\|\widetilde{x}\| \le R$, by Corollary J.6 and Lemma H.2 we have

$$\|\nabla b_1\|_{\mathrm{op}} \le 3\widetilde{G}, \qquad \|\partial_t b_1\|_2 \le 3\widetilde{D}\,\kappa(T)\,R\sqrt{d}, \qquad \|b_1\|_\infty \le \widetilde{B}_{\mathrm{vel}}\,R\sqrt{dd^*}.$$

*Step 2 (Approximate $b_2^*$).* By Corollary J.6, we first approximate $\gamma(t)$ on $[0,T]$ by an 1-D network $\gamma_\theta$ of $L = 3$ and $S = 28(N+1)$ with $\|\gamma_\theta - \gamma\|_{L^\infty([0,T])} \le C_4N^{-1}$ and $\|\partial_t\gamma_\theta\|_{L^\infty([0,T])} \le 3\gamma_{\max}$. Set $b_2(t,x) := \gamma_\theta(t)(I - PP^\top)x$. The linear map $x \mapsto (I - PP^\top)x$ is affine with $O(d^2)$ weights. The coordinate-wise products $\gamma_\theta(t) \cdot ((I - PP^\top)x)_j$, $j = 1, \ldots, d$ are realized via Lemma J.2, adding one ReQU hidden layer and $12d$ parameters. Then

$$\|b_2 - b_2^*\|_{L^\infty(\mathbb{B}_{R,T})} \le \|\gamma_\theta - \gamma\|_{L^\infty([0,T])}R\sqrt{d} \le C_4R\sqrt{d}N^{-1},$$

while $\|\nabla b_2\|_{\mathrm{op}} \le \gamma_{\max}$, $\|\partial_t b_2\|_2 \le 3\gamma_{\max}R\sqrt{d}$ and $\|b_2\|_\infty \le \gamma_{\max}R\sqrt{d}$. Here $C_4$ is a universal constant.

*Step 3 (Aggregation).* We set $b = b_1 + b_2$ as the final network, which implements add operation in the output affine layer with $O(d)$ weights. The total extra overhead is $O(dd^*) + O(d^2) + O(N)$, which is absorbed into $L = C_5$, $S = C_6\,N^{d^*+1}$. Combining Step 1 and 2 gives

$$\|b - b^*\|_{L^\infty(\mathbb{B}_{R,T})} \le \|b_1 - b_1^*\|_{L^\infty(\mathbb{B}_{R,T})} + \|b_2 - b_2^*\|_{L^\infty(\mathbb{B}_{R,T})} \le C_7\kappa(T)R\,N^{-1},$$

which yields $\mathcal{E}_{T,R}(b) \le C_8\kappa^2(T)R^2N^{-2}$. Furthermore,

$$\|\nabla b\|_{\mathrm{op}} \le 3\widetilde{G} + \gamma_{\max}, \quad \|\partial_t b\|_2 \le 3R\sqrt{d}(\widetilde{D}\kappa(T) + \gamma_{\max}), \quad \|b\|_\infty \le R\sqrt{d}(\widetilde{B}_{\mathrm{vel}}\sqrt{d^*} + \gamma_{\max}).$$

Here, $C_5$, $C_6$, $C_7$ and $C_8$ are constants which may only depend on $d$, $d^*$, and $\sigma$. This completes the proof. $\square$

With Lemma H.5 established, we are now ready to derive the convergence rate for velocity matching under the manifold hypothesis. The following theorem is the manifold counterpart of Theorem 3.10.

**Theorem H.6** (Convergence rate for velocity matching under manifold hypothesis). *Suppose Assumptions 1 and 3 hold. Let $T \in (1/2, 1)$. Set the hypothesis class $\mathscr{B}$ as a deep neural network class, which is defined as*

$$\mathscr{B} = \left\{ b \in N(L,S) : \begin{array}{l} \|b(t,x)\|_\infty \le B_{\mathrm{man}}\log^{1/2}n, \\ \|\partial_t b(t,x)\|_2 \le D_{\mathrm{man}}\log^{1/2}n, \\ \|\nabla b(t,x)\|_{\mathrm{op}} \le G_{\mathrm{man}} \end{array} \right\},$$

*where the Lipschitz control parameters $B_{\mathrm{man}}$, $D_{\mathrm{man}}$, and $G_{\mathrm{man}}$ are given in Lemma H.5 and depend only on $d$, $d^*$, and $\sigma$. The depth and the number of non-zero weights of the neural network are given, respectively, as $L = C$ and $S = Cn^{\frac{d^*+1}{d^*+3}}$. Then the following inequality holds*

$$\mathbb{E}_{\mathbb{S}}\big[\mathcal{E}_T(\widehat{b})\big] \leq C\kappa^2(T)n^{-\frac{2}{d^*+3}}\log^2 n,$$

*where $C$ is a constant only depending on $d$, $d^*$ and $\sigma$.*

*Proof of Theorem H.6.* Under the manifold hypothesis, Lemma H.5 establishes that there exists a neural network class $\mathscr{B}$ with $S = CN^{d^*+1}$ nonzero weights such that $\inf_{b \in \mathscr{B}} \mathcal{E}_{T,R}(b) \leq C\kappa^2(T)R^2N^{-2}$. By Lemma B.12, the VC-dimension satisfies $\mathrm{VCdim}(\Pi_k\mathscr{B}) \leq CN^{d^*+1}\log N$. Following the same proof strategy as Theorem 3.10 (Appendix E) and setting $N = Cn^{\frac{1}{d^*+3}}$ yields the desired result. $\qquad\square$

Following the same line of reasoning, the manifold counterparts of Corollary 3.11, Theorem 3.12, and Corollary 3.13 can be readily obtained by replacing $d$ with $d^*$ in the corresponding convergence rates. We omit these results here for brevity and proceed directly to the analysis of the characteristic generator under the manifold hypothesis, for which we require an approximation result for the manifold-structured flow map $g^*$. The following lemma parallels Lemma H.5.

**Lemma H.7** (Approximation under manifold-structured flow). *Let $T \in (1/2, 1)$ and $R \in (1, +\infty)$. Suppose $g^*$ admits the decomposition in (H.9):*

(H.12) $$g^*(t,s,x) = P\,\widetilde{g}^*(t,s,P^\top x) + \iota(t,s)(I_d - PP^\top)x,$$

*where $P \in \mathbb{R}^{d \times d^*}$ has orthonormal columns, $\widetilde{g}^*$ satisfies Lemma H.3, and $\iota(t,s) := \sqrt{\frac{\alpha_s^2 + \sigma^2\beta_s^2}{\alpha_t^2 + \sigma^2\beta_t^2}}$. Set the hypothesis class $\mathscr{G}$ as a deep neural network class, which is defined as*

$$\mathscr{G} = \left\{ g \in N(L,S) : \begin{array}{l} \|g(t,s,x)\|_\infty \leq B_{\mathrm{man}}^{\mathrm{flow}}R, \ \max\{\|\partial_t g\|_2, \|\partial_s g\|_2\} \leq D_{\mathrm{man}}^{\mathrm{flow}}R, \\ \|\nabla g(t,s,x)\|_{\mathrm{op}} \leq G_{\mathrm{man}}^{\mathrm{flow}}, \ 0 \leq t \leq s \leq T, \ x \in \mathbb{R}^d \end{array} \right\},$$

*where the depth and the number of nonzero weights of the neural network are given, respectively, as $L = C$ and $S = CN^{d^*+2}$. Then the following inequality holds for each $N \in \mathbb{N}_+$,*

(H.13) $$\inf_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) \leq CR^2N^{-2}.$$

*Using Lemma H.3, we have admissible choices for $B_{\mathrm{man}}^{\mathrm{flow}}$, $D_{\mathrm{man}}^{\mathrm{flow}}$ and $G_{\mathrm{man}}^{\mathrm{flow}}$ as follows:*

$$B_{\mathrm{man}}^{\mathrm{flow}} = \sqrt{d}(\widetilde{B}_{\mathrm{flow}}\sqrt{d^*} + \iota_{\max}), \quad D_{\mathrm{man}}^{\mathrm{flow}} = 3\sqrt{d}(\widetilde{B}'_{\mathrm{flow}} + \iota_{\max}), \quad G_{\mathrm{man}}^{\mathrm{flow}} = 3\exp(\widetilde{G}) + \iota_{\max}$$

*with $\iota_{\max} := \max\left\{\|\iota\|_{L^\infty}, \|\partial_t\iota\|_{L^\infty}, \|\partial_s\iota\|_{L^\infty}\right\}$ finite by Condition 1. And $C$ denotes a constant may only depend on $d^*$, $d$ and $\sigma$.*

With Lemma H.7, we establish the convergence result for the characteristic generator under the manifold hypothesis. This theorem is the manifold counterpart of Theorem 3.14, demonstrating that the characteristic generator can achieve a convergence rate depending on the intrinsic dimension $d^*$ rather than the ambient dimension $d$.

**Theorem H.8** (Error analysis for characteristic generator under manifold hypothesis). *Suppose Assumptions 1 and 3 hold. Under the same conditions as Theorem H.6. Further, set the hypothesis class $\mathscr{G}$ as a deep neural network class, which is defined as*

$$
\mathscr{G} = \left\{ g \in N(L, S) : \begin{array}{l} \|g(t, s, x)\|_\infty \leq B_{\mathrm{man}}^{\mathrm{flow}} \log^{1/2} m, \\[4pt] \|\partial_s g(t, s, x)\|_2 \leq D_{\mathrm{man}}^{\mathrm{flow}} \log^{1/2} m, \\[4pt] \|\partial_t g(t, s, x)\|_2 \leq D_{\mathrm{man}}^{\mathrm{flow}} \log^{1/2} m, \\[4pt] \|\nabla g(t, s, x)\|_{\mathrm{op}} \leq G_{\mathrm{man}}^{\mathrm{flow}} \end{array} \right\},
$$

*where the Lipschitz control parameters $B_{\mathrm{man}}^{\mathrm{flow}}$, $D_{\mathrm{man}}^{\mathrm{flow}}$, and $G_{\mathrm{man}}^{\mathrm{flow}}$ are given in Lemma H.7 and depend only on $d$, $d^*$, and $\sigma$. The depth and the number of non-zero weights of the neural network are given, respectively, as $L = C$ and $S = Cm^{\frac{d^*+2}{d^*+4}}$. Then it follows that*

$$
\mathbb{E}_\mathscr{S} \mathbb{E}_\mathscr{Z} [\mathcal{D}(\widehat{g})] \leq C\kappa^2(T) \Big\{ n^{-\frac{2}{d^*+3}} \log^2 n + \frac{\log n}{K^2} \Big\} + C \Big\{ m^{-\frac{2}{d^*+4}} \log^2 m + \frac{\log m}{K} \Big\},
$$

*where the constant $C$ only depends on $d$, $d^*$ and $\sigma$. Furthermore, if the number of time steps $K$ for Euler method and the number of samples $m$ for characteristic fitting satisfy*

$$
K \geq \max \big\{ Cn^{\frac{1}{d^*+3}}, C\kappa^{-2}(T)n^{-\frac{2}{d^*+3}} \big\}, \quad m \geq C\kappa^{-(d^*+4)}(T)n^{\frac{d^*+4}{d^*+3}},
$$

*respectively, then the following inequality holds*

$$
\mathbb{E}_\mathscr{S} \mathbb{E}_\mathscr{Z} [\mathcal{D}(\widehat{g})] \leq C\kappa^2(T)n^{-\frac{2}{d^*+3}} \log^2 n.
$$

*Proof of Theorem H.8.* Under the manifold hypothesis, Lemma H.7 establishes that there exists a neural network class $\mathscr{G}$ with $S = CN^{d^*+2}$ nonzero weights such that $\inf_{g \in \mathscr{G}} \mathcal{R}_{T,R}(g) \leq C\kappa^2(T)R^2 N^{-2}$. By Lemma B.12, the VC-dimension satisfies $\mathrm{VCdim}(\Pi_k \mathscr{G}) \leq CN^{d^*+2} \log N$. Following the same proof strategy as Theorem 3.14 (Appendix G) and setting $N = Cm^{\frac{1}{d^*+4}}$ yields the desired result. $\qquad\square$

Theorems H.6 and H.8 demonstrate that when the target distribution has a low dimensional manifold structure, both the velocity matching and the characteristic generator can achieve convergence rates that depend on the intrinsic dimension $d^*$ instead of the ambient dimension $d$, thereby effectively mitigating the curse of dimensionality. Applying these results to the analysis of linear interpolant and Föllmer flow yields Corollaries 3.19 and 3.20 in the main text.

## Appendix I. Generalization Error Analysis for Least-Squares Regression

In this section, we provide a generalization error analysis for nonparametric regression, which is used in establishing oracle inequalities for velocity matching (Lemma E.1) and characteristic fitting (Lemma G.2).

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a bounded domain, and let $X \in \mathcal{X}$ be a random variable obeying the probability distribution $\mu$. Let $f^* : \mathcal{X} \to \mathbb{R}$ be a measurable function. Define the population $L^2(\mu)$-risk for a function $f : \mathcal{X} \to \mathbb{R}$ as

$$
R(f) = \|f - f^*\|_{L^2(\mu)}^2 = \mathbb{E}_{X \sim \mu}[(f(X) - f^*(X))^2].
$$

Let $\mathcal{D} = \{X^{(i)}\}_{i=1}^n$ be a set of i.i.d. copies of $X \sim \mu$. Then define the empirical risk of $f$ by

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n (f(X^{(i)}) - f^*(X^{(i)}))^2.$$

The following lemma relates the population risk to its empirical counterpart.

**Lemma I.1.** *Suppose that $|f^*(x)| \leq B$ for each $x \in \mathcal{X}$. Let $\mathscr{F}$ be a set of functions mapping from $\mathcal{X}$ to $[-B, B]$. Then it follows that for each $n \geq \mathrm{VCdim}(\mathscr{F})$,*

$$\mathbb{E}_{\mathcal{D}}\Big[ \sup_{f \in \mathscr{F}} R(f) - 2\widehat{R}_n(f) \Big] \leq CB^2 \frac{\mathrm{VCdim}(\mathscr{F})}{n \log^{-1} n},$$

*where $C$ is an absolute constant.*

This lemma provides a generalization error bound with fast rate via the technique of the offset Rademacher complexity, which was first proposed by Liang et al. (2015). In recent years, this technique has been applied to the convergence rate analysis for deep nonparametric regression, such as (Duan et al., 2023, Lemma 14) and (Ding et al., 2024, Lemma 4.1).

*Proof of Lemma I.1.* We define an auxiliary function class

$$\mathscr{H} = \Big\{ x \mapsto h(x) = (f(x) - f^*(x))^2 : f \in \mathscr{F} \Big\}.$$

It is apparent that $0 \leq h(x) \leq 4B^2$ for each $x \in \mathcal{X}$ and $h \in \mathscr{H}$. Then it is easy to show that

$$\mathbb{E}_{\mathcal{D}}\Big[ \sup_{f \in \mathscr{F}} R(\widehat{f}) - 2\widehat{R}_n(\widehat{f}) \Big] \leq \mathbb{E}_{\mathcal{D}}\Big[ \sup_{h \in \mathscr{H}} \mathbb{E}[h(X)] - \frac{2}{n} \sum_{i=1}^n h(X^{(i)}) \Big]$$

$$\leq \mathbb{E}_{\mathcal{D}}\Big[ \sup_{h \in \mathscr{H}} \mathbb{E}\Big[\frac{3}{2}h(X) - \frac{1}{8B^2}h^2(X)\Big] - \frac{1}{n} \sum_{i=1}^n \Big( \frac{3}{2}h(X^{(i)}) + \frac{1}{8B^2}h^2(X^{(i)}) \Big) \Big],$$

where we used the fact that $h^2(x) \leq 4B^2 h(x)$ for each $x \in \mathcal{X}$ and $h \in \mathscr{H}$.

Let us introduce a ghost sample $\mathcal{D}' = \{X^{(i),'}\}_{i=1}^n$, which is a set of $n$ i.i.d. random copies of $X \sim \mu$. Here the ghost sample $\mathcal{D}'$ is independent of $\mathcal{D} = \{X^{(i)}\}_{i=1}^n$. Let $\xi = \{\xi^{(i)}\}_{i=1}^n$ be a set of i.i.d. Rademacher variables. Then replacing the expectation by the empirical mean based on the ghost sample $\mathcal{D}'$ yields

$$\mathbb{E}_{\mathcal{D}}\Big[ \sup_{h \in \mathscr{H}} \mathbb{E}_X\Big[\frac{3}{2}h(X) - \frac{1}{8B^2}h^2(X)\Big] - \frac{1}{n} \sum_{i=1}^n \Big( \frac{3}{2}h(X^{(i)}) + \frac{1}{8B^2}h^2(X^{(i)}) \Big) \Big]$$

$$= \mathbb{E}_{\mathcal{D}}\Big[ \sup_{h \in \mathscr{H}} \mathbb{E}_{\mathcal{D}'}\Big[\frac{1}{n} \sum_{i=1}^n \frac{3}{2}h(X^{(i),'}) - \frac{1}{8B^2}h^2(X^{(i),'})\Big] - \frac{1}{n} \sum_{i=1}^n \Big( \frac{3}{2}h(X^{(i)}) + \frac{1}{8B^2}h^2(X^{(i)}) \Big) \Big]$$

$$\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\Big[ \sup_{h \in \mathscr{H}} \frac{3}{2n} \sum_{i=1}^n (h(X^{(i),'}) - h(X^{(i)})) - \frac{1}{8B^2 n} \sum_{i=1}^n (h^2(X^{(i),'}) + h^2(X^{(i)})) \Big]$$

$$= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\mathbb{E}_{\xi}\Big[ \sup_{h \in \mathscr{H}} \frac{3}{2n} \sum_{i=1}^n \xi^{(i)}(h(X^{(i),'}) - h(X^{(i)})) - \frac{1}{8B^2 n} \sum_{i=1}^n (h^2(X^{(i),'}) + h^2(X^{(i)})) \Big]$$

$$= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\xi}\Big[ \sup_{h \in \mathscr{H}} \frac{3}{n} \sum_{i=1}^n \xi^{(i)}h(X^{(i)}) - \frac{1}{4B^2 n} \sum_{i=1}^n h^2(X^{(i)}) \Big],$$

where the inequality holds from Jensen's inequality. Combining the above results, we have

$$\text{(I.1)} \qquad \mathbb{E}_{\mathcal{D}}\Big[\sup_{f\in\mathscr{F}} R(\widehat{f}) - 2\widehat{R}_n(\widehat{f})\Big] \le \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\xi}\Big[\sup_{h\in\mathscr{H}} \frac{3}{n}\sum_{i=1}^{n} \xi^{(i)} h(X^{(i)}) - \frac{1}{4B^2 n}\sum_{i=1}^{n} h^2(X^{(i)})\Big].$$

We next estimate the expectation in the right-hand side of (I.1). Let $\delta \in (0, 4B^2)$ and $\mathscr{H}_\delta$ be a $L^\infty(\mathcal{D})$ $\delta$-cover of $\mathscr{H}$ satisfying $|\mathscr{H}_\delta| = N(\delta, \mathscr{H}, L^\infty(\mathcal{D}))$. Then for each $h \in \mathscr{H}$, there exists $h_\delta \in \mathscr{H}_\delta$ such that

$$\max_{1\le i\le n} |h(X^{(i)}) - h_\delta(X^{(i)})| \le \delta.$$

Without loss of generality, we assume $|h_\delta(x)| \le 4B^2$ for each $h_\delta \in \mathscr{H}_\delta$. Consequently, it follows from Hölder's inequality that

$$\frac{1}{n}\sum_{i=1}^{n} \xi^{(i)} h(X^{(i)}) - \frac{1}{n}\sum_{i=1}^{n} \xi^{(i)} h_\delta(X^{(i)}) \le \frac{1}{n}\sum_{i=1}^{n} |\xi^{(i)}||h(X^{(i)}) - h_\delta(X^{(i)})| \le \delta.$$

By the same argument, it holds that

$$-\frac{1}{n}\sum_{i=1}^{n} h^2(X^{(i)}) + \frac{1}{n}\sum_{i=1}^{n} h_\delta^2(X^{(i)}) \le 8B^2\delta.$$

With the help of the above two inequalities, we have

$$\mathbb{E}_{\xi}\Big[\sup_{h\in\mathscr{H}} \frac{3}{n}\sum_{i=1}^{n} \xi^{(i)} h(X^{(i)}) - \frac{1}{4B^2 n}\sum_{i=1}^{n} h^2(X^{(i)})\Big]$$

$$\text{(I.2)} \qquad \le \mathbb{E}_{\xi}\Big[\max_{h_\delta\in\mathscr{H}_\delta} \frac{3}{n}\sum_{i=1}^{n} \xi^{(i)} h_\delta(X^{(i)}) - \frac{1}{4B^2 n}\sum_{i=1}^{n} h_\delta^2(X^{(i)})\Big] + 5\delta.$$

Observe that $\{\xi^{(i)} h_\delta(X^{(i)})\}_{i=1}^{n}$ is a set of $n$ i.i.d. random variables with

$$-h_\delta(X^{(i)}) \le \xi^{(i)} h_\delta(X^{(i)}) \le h_\delta(X^{(i)}), \quad 1 \le i \le n.$$

Then it follows Hoeffding's inequality (Mohri et al., 2018, Theorem D.2) that

$$\text{Pr}_{\xi}\Big\{\frac{3}{n}\sum_{i=1}^{n} \xi^{(i)} h_\delta(X^{(i)}) > t + \frac{1}{4B^2 n}\sum_{i=1}^{n} h_\delta^2(X^{(i)})\Big\}$$

$$= \text{Pr}_{\xi}\Big\{\sum_{i=1}^{n} \xi^{(i)} h_\delta(X^{(i)}) > \frac{nt}{3} + \frac{1}{12B^2}\sum_{i=1}^{n} h_\delta^2(t^{(i)}, X_t^{(i)})\Big\}$$

$$\text{(I.3)} \qquad \le \exp\Big(-\frac{(\frac{nt}{3} + \frac{1}{12B^2}\sum_{i=1}^{n} h_\delta^2(t^{(i)}, X_t^{(i)}))^2}{2\sum_{i=1}^{n} h_\delta^2(t^{(i)}, X_t^{(i)})}\Big) \le \exp\Big(-\frac{nt}{18B^2}\Big),$$

where the first inequality follows from Hoeffding's inequality (Mohri et al., 2018, Theorem D.2), and the second inequality is due to $(a + b)^2/b \le 4a$ for each $a > 0$ and $b \in \mathbb{R}$. As a

consequence, for each $A > 0$,

$$
\mathbb{E}_\xi \Big[ \max_{h_\delta \in \mathscr{H}_\delta} \frac{3}{n} \sum_{i=1}^n \xi^{(i)} h_\delta(X^{(i)}) - \frac{1}{4B^2 n} \sum_{i=1}^n h_\delta^2(X^{(i)}) \Big]
$$

$$
= \int_0^\infty \mathrm{Pr}_\xi \Big\{ \max_{h_\delta \in \mathscr{H}_\delta} \frac{3}{n} \sum_{i=1}^n \xi^{(i)} h_\delta(X^{(i)}) - \frac{1}{4B^2 n} \sum_{i=1}^n h_\delta^2(X^{(i)}) > t \Big\} \mathrm{d}t
$$

$$
\leq A + |\mathscr{H}_\delta| \int_T^\infty \exp\Big(-\frac{nt}{18B^2}\Big) \mathrm{d}t = A + \frac{18B^2}{n} |\mathscr{H}_\delta| \exp\Big(-\frac{nA}{18B^2}\Big),
$$

where the inequality is owing to (I.3). Letting $A = 18B^2 \log |\mathscr{H}_\delta| n^{-1}$ gives that

$$
\text{(I.4)} \qquad \mathbb{E}_\xi \Big[ \max_{h_\delta \in \mathscr{H}_\delta} \frac{3}{n} \sum_{i=1}^n \xi^{(i)} h_\delta(X^{(i)}) - \frac{1}{4B^2 n} \sum_{i=1}^n h_\delta^2(X^{(i)}) \Big] \leq 18B^2 \frac{\log |\mathscr{H}_\delta| + 1}{n}.
$$

It remains to estimate the covering number $|\mathscr{H}_\delta| = N(\delta, \mathscr{H}, L^\infty(\mathcal{D}))$. Noticing that

$$
|h(x) - h'(x)| = |(f(x) - f^*(x))^2 - (f'(x) - f^*(x))^2| \leq 4B|f(x) - f'(x)|,
$$

we obtain that for $n \geq \mathrm{VCdim}(\mathscr{F})$,

$$
\text{(I.5)} \qquad \log N(\delta, \mathscr{H}, L^\infty(\mathcal{D})) \leq \log N\Big(\frac{\delta}{4B}, \mathscr{F}, L^\infty(\mathcal{D})\Big) \leq \mathrm{VCdim}(\mathscr{F}) \log\Big(\frac{4eB^2 n}{\delta}\Big),
$$

where the first and last inequalities follows from Lemmas B.10 and B.11, respectively. Combining (I.1), (I.2), (I.4) and (I.5), we have

$$
\mathbb{E}_{\mathcal{D}} \Big[ \sup_{f \in \mathscr{F}} R(\widehat{f}) - 2\widehat{R}_n(\widehat{f}) \Big] \leq \inf_{\delta > 0} \Big\{ 36B^2 \frac{\mathrm{VCdim}(\mathscr{F})}{n} \log\Big(\frac{4eB^2 n}{\delta}\Big) + 5\delta \Big\}.
$$

Substituting $\delta = 4B^2/n$ into the above inequality completes the proof. $\qquad\square$

**Lemma I.2.** *Suppose that $|f^*(x)| \leq B$ for each $x \in \mathcal{X}$. Let $\mathscr{F}$ be a set of functions mapping from $\mathcal{X}$ to $[-B, B]$. Let $\{\varepsilon^{(i)}\}_{i=1}^n$ be a set of independent $\sigma^2$-sub-Gaussian random variables. Then it follows that for each $n \geq \mathrm{VCdim}(\mathscr{F})$,*

$$
\mathbb{E}_{(\mathcal{D}, \varepsilon)} \Big[ \frac{1}{n} \sum_{i=1}^n \varepsilon^{(i)} \widehat{f}(X^{(i)}) \Big] \leq \frac{1}{4} \mathbb{E}_{(\mathcal{D}, \varepsilon)}[\widehat{R}(\widehat{f})] + C(B^2 + \sigma^2) \frac{\mathrm{VCdim}(\mathscr{F})}{n \log^{-1} n},
$$

*where $C$ is an absolute constant.*

This proof uses a technique similar to the proof of (Schmidt-Hieber, 2020, Lemma 4).

*Proof of Lemma I.2.* Let $\delta \in (0, B)$ and let $\mathscr{F}_\delta$ be a $L^\infty(\mathcal{D})$ $\delta$-cover of $\mathscr{F}$ with $|\mathscr{F}_\delta| = N(\delta, \mathscr{F}, L^\infty(\mathcal{D}))$. Then there exists $\widehat{f}_\delta \in \mathscr{F}_\delta$, such that

$$
\max_{1 \leq i \leq n} |\widehat{f}(X^{(i)}) - \widehat{f}_\delta(X^{(i)})| \leq \delta.
$$

Then it follows from Hölder's inequality that

$$
\text{(I.6)} \qquad \mathbb{E}_{(\mathcal{D}, \varepsilon)} \Big[ \frac{1}{n} \sum_{i=1}^n \varepsilon^{(i)} (\widehat{f}(X^{(i)}) - \widehat{f}_\delta(X^{(i)})) \Big] \leq \delta \mathbb{E}_{(\mathcal{D}, \varepsilon)} \Big[ \frac{1}{n} \sum_{i=1}^n |\varepsilon^{(i)}| \Big] \leq \delta\sigma,
$$

where we used the fact that $\{\varepsilon^{(i)}\}_{i=1}^n$ are a set of $\sigma^2$-sub-Gaussian random variables. Additionally, according to the triangular inequality, we have

$$(I.7) \qquad \widehat{R}_n^{1/2}(\widehat{f}_\delta) - \widehat{R}_n^{1/2}(\widehat{f}) \le \Big(\frac{1}{n}\sum_{i=1}^n (\widehat{f}_\delta(X^{(i)}) - \widehat{f}(X^{(i)}))^2\Big)^{1/2} \le \delta.$$

Consequently, we have

$$\mathbb{E}_{(\mathcal{D},\varepsilon)}\Big[\frac{1}{n}\sum_{i=1}^n \varepsilon^{(i)}\widehat{f}(X^{(i)})\Big] = \mathbb{E}_{(\mathcal{D},\varepsilon)}\Big[\frac{1}{n}\sum_{i=1}^n \varepsilon^{(i)}(\widehat{f}(X^{(i)}) - \widehat{f}_\delta(X^{(i)}) + \widehat{f}_\delta(X^{(i)}) - f^*(X^{(i)}))\Big]$$

$$\le \mathbb{E}_{(\mathcal{D},\varepsilon)}\Big[\frac{1}{n}\sum_{i=1}^n \varepsilon^{(i)}(\widehat{f}_\delta(X^{(i)}) - f^*(X^{(i)}))\Big] + \delta\sigma$$

$$\le \frac{1}{\sqrt{n}}\mathbb{E}_{(\mathcal{D},\varepsilon)}\Big[\Big(\widehat{R}_n^{1/2}(\widehat{f}) + \delta\Big)\sum_{i=1}^n \frac{\varepsilon^{(i)}(\widehat{f}_\delta(X^{(i)}) - f^*(X^{(i)}))}{\sqrt{n}\widehat{R}_n^{1/2}(\widehat{f}_\delta)}\Big] + \delta\sigma$$

$$\le \frac{1}{\sqrt{n}}\Big(\mathbb{E}_{(\mathcal{D},\varepsilon)}^{1/2}\Big[\widehat{R}_n(\widehat{f})\Big] + \delta\Big)\mathbb{E}_{(\mathcal{D},\varepsilon)}^{1/2}\Big[\psi^2(\widehat{f}_\delta)\Big] + \delta\sigma$$

$$(I.8) \qquad \le \frac{1}{4}\mathbb{E}_{(\mathcal{D},\varepsilon)}\Big[\widehat{R}_n(\widehat{f})\Big] + \frac{2}{n}\mathbb{E}_{(\mathcal{D},\varepsilon)}[\psi^2(\widehat{f}_\delta)] + \frac{1}{4}\delta^2 + \delta\sigma,$$

where $\psi(\widehat{f}_\delta)$ is defined as

$$\psi(\widehat{f}_\delta) = \sum_{i=1}^n \frac{\widehat{f}_\delta(X^{(i)}) - f^*(X^{(i)})}{\sqrt{n}\widehat{R}_n^{1/2}(\widehat{f}_\delta)}\varepsilon^{(i)}.$$

Here the first inequality follows from (I.6), the second inequality holds from (I.7), the third inequality is due to Cauchy-Schwarz inequality, and the last inequality is owing to the weighted AM-GM inequality $ab \le a/4 + b$ for each $a, b \in \mathbb{R}$.

Observe that for each fixed function $f : \mathbb{R}^d \to \mathbb{R}$, the random variable $\psi(f)$ is sub-Gaussian with variance proxy $\sigma^2$ conditioning on $\mathcal{D} = \{X^{(i)}\}_{i=1}^n$. Then it follows that

$$(I.9) \qquad \mathbb{E}_\varepsilon[\psi^2(\widehat{f}_\delta)] \le \mathbb{E}_\varepsilon\Big[\max_{f_\delta \in \mathscr{F}_\delta} \psi^2(f_\delta)\Big] \le 4\sigma^2(\log|\mathscr{F}_\delta| + 1).$$

We now estimate the covering number $|\mathscr{F}_\delta| = N(\delta, \mathscr{F}, L^\infty(\mathcal{D}))$. It follows from Lemma B.11 that for $n \ge \mathrm{VCdim}(\mathscr{F})$,

$$(I.10) \qquad \log N(\delta, \mathscr{F}, L^\infty(\mathcal{D})) \le \mathrm{VCdim}(\mathscr{F})\log\Big(\frac{eBn}{\delta}\Big).$$

Combining (I.8), (I.9) and (I.10), and setting $\delta = B/n$ complete the proof. $\qquad\square$

## APPENDIX J. APPROXIMATION BY DEEP NEURAL NETWORKS WITH LIPSCHITZ CONSTRAINT

The approximation error analysis for deep neural networks has been investigated by Yarotsky (2017, 2018), Yarotsky and Zhevnerchuk (2020), Shen et al. (2019), Shen (2020), Lu et al. (2021), Petersen and Voigtlaender (2018), Jiao et al. (2023a), Duan et al. (2022). However, limited work has been done for deep neural networks with Lipschitz constraint Huang et al. (2022), Chen et al. (2022), Jiao et al. (2023b), Ding et al. (2024).

This proof is based on the proof of (Yarotsky, 2017, Theorem 1). The ReLU activation function is defined as $\mathrm{ReLU}(x) = \max\{0, x\}$, and the ReQU activation function is defined as the squared ReLU function $\mathrm{ReQU}(x) = (\max\{0, x\})^2$.

**Lemma J.1.** *The maximum or minimum of two inputs can be implemented by a ReLU neural network with $1$ hidden layer and $7$ non-zero parameters.*

*Proof of Lemma J.1.* According to the equality $a = \mathrm{ReLU}(a) - \mathrm{ReLU}(-a)$, the identity mapping can be implemented by a ReLU neural network with $1$ hidden layer and $4$ non-zero parameters. We also notice that

$$\max\{a, b\} = a + \mathrm{ReLU}(b - a) = \mathrm{ReLU}(a) - \mathrm{ReLU}(-a) + \mathrm{ReLU}(b - a),$$

which means that the maximum of two inputs can be implemented by $7$ non-zero parameters. By a same argument, with the aid of equality $\min\{a, b\} = a - \mathrm{ReLU}(a - b)$, we can obtain a same result for the minimum of two inputs. This completes the proof. $\square$

**Lemma J.2.** *The product of two inputs can be implement by a ReQU neural network with $1$ hidden layer and $12$ non-zero parameters.*

*Proof of Lemma J.2.* According to (Li et al., 2019, Lemma 2.1), the following identities hold:

$$x_1 x_2 = \frac{1}{4} w_3^T \mathrm{ReQU}(w_1 x_1 + w_2 x_2),$$

where $w_1 = (1, -1, 1, -1)^T$, $w_2 = (1, -1, -1, 1)^T$ and $w_3 = (1, 1, -1, -1)^T$, which completes the proof. $\square$

**Lemma J.3.** *The product of $p$ inputs can be implement by a ReQU neural network with $\lceil \log_2 p \rceil$ hidden layers and $6(p + 1)$ non-zero parameters.*

*Proof of Lemma J.3.* Define the augmented input vector $(x_1, \ldots, x_p, x_{p+1}, \ldots, x_n)$ where $n = 2^{\lceil \log_2 p \rceil}$ and $x_i = 1$ for $p + 1 \le i \le n$. Observe that $\prod_{i=1}^p x_i = \prod_{i=1}^n x_i$. According to Lemma J.2, the mapping $(x_1, \ldots, x_n) \to (x_1 x_2, \ldots, x_{n-1} x_n) \in \mathbb{R}^{n/2}$ can be implemented by a ReQU neural network with $1$ hidden layer and $6n$ non-zero parameters. By a same argument, we can construct a ReQU neural network with $1$ hidden layer and $3n$ non-zero parameters, which maps $(x_1 x_2, \ldots, x_{n-1} x_n)$ to $(x_1 x_2 x_3 x_4, \ldots, x_{n-3} x_{n-2} x_{n-1} x_n) \in \mathbb{R}^{n/4}$. By induction on $n$, the number of layers is given as $\lceil \log_2 p \rceil$, and the total number of non-zero parameters is given by

$$12 \times (1 + 2 + 2^2 + \ldots + 2^{\lceil \log_2 p \rceil - 1}) \le 6(p + 1).$$

This completes the proof. $\square$

**Lemma J.4.** *The univariate trapezoid function*

(J.1)
$$\psi(z) = \begin{cases} 1, & |z| < 1, \\ 2 - |z|, & 1 \le |z| \le 2, \\ 0, & 2 < |z|, \end{cases}$$

*can be implement by a ReLU neural network with $3$ hidden layers and $14$ non-zero parameters.*

*Proof of Lemma J.4.* We first implement the following hat-function by ReLU neural network

$$\widetilde{\psi}(z) = \begin{cases} 2 - |z|, & |z| \leq 2, \\ 0, & 2 < |z|. \end{cases}$$

Noticing that $\widetilde{\psi}(z) = \min\{\text{ReLU}(z+2), \text{ReLU}(-z+2)\}$, by applying Lemma J.1, we find that $\widetilde{\psi}$ can be implemented by a ReLU neural network with 2 hidden layers and 11 non-zero parameters. Further, according to the equality $\psi(z) = \min\{1, \widetilde{\psi}(z)\}$, the univariate trapezoid function $\psi$ can be implemented by a ReLU neural network with 3 hidden layers and 14 non-zero parameters. This completes the proof. $\square$

**Lemma J.5** (Approximation error). *Let $p \in \mathbb{N}_+$, and let $\{N_k\}_{k=1}^p$ be a set of positive integer. Set the deep neural network class $N(L, S)$ as $L = \lceil \log_2 p \rceil + 3$ and $S = (22p + 6) \prod_{k=1}^p (N_k + 1)$. Then for each $u^* \in W^{1,\infty}([0,1]^p)$, there exists a deep neural network $u \in N(L, S)$ such that*

$$\|u - u^*\|_{L^\infty([0,1]^p)} \leq 2^p \sum_{k=1}^p \frac{1}{N_k} \|\partial_k u^*\|_{L^\infty(\mathcal{X})}.$$

*Further, it holds that the following holds for each $1 \leq k \leq p$:*

$$\|u\|_{L^\infty([0,1]^p)} = \|u^*\|_{L^\infty([0,1]^p)} \quad and \quad \|\partial_k u\|_{L^\infty([0,1]^p)} \leq 3\|\partial_k u^*\|_{L^\infty([0,1]^p)}.$$

*Proof of Lemma J.5.* The proof is divided into three steps. In the first step, we approximate the target function based on a partition of unity and the degree-0 Taylor expansion. Then we implement this piece-wise linear function using deep neural network exactly in the second step. Finally, in the last step, we estimate the Lipschitz constant of the deep neural network.

*Step 1. Approximate the target function by a piecewise linear function.*

Consider a partition of unity formed by a grid of $\prod_{k=1}^p (N_k + 1)$ functions $\phi_m$ on the domain $[0,1]^p$:

$$\text{(J.2)} \qquad \sum_m \phi_m(x) \equiv 1, \quad x \in [0,1]^p,$$

where the multi-index $m$ is defined as $m = (m_1, \ldots, m_p)^T$ with $m_k \in \{0, \ldots, N_k\}$, and the function $\phi_m$ is defined as the product

$$\text{(J.3)} \qquad \phi_m(x) = \prod_{k=1}^p \psi\Big(3N_k\Big(x_k - \frac{m_k}{N_k}\Big)\Big).$$

Here $\psi$ is the univariate trapezoid function defined as (J.1). It is noticeable that for each $m$,

$$\text{(J.4)} \qquad \sup_{z \in [0,1]} |\psi(z)| = 1, \qquad \sup_{x \in [0,1]^p} |\phi_m(x)| = 1,$$

and

$$\text{(J.5)} \qquad \text{supp}(\phi_m) \subseteq \Big\{x \in [0,1]^p : \Big|x_k - \frac{m_k}{N_k}\Big| \leq \frac{2}{3N_k}, \ 1 \leq k \leq p\Big\}.$$

Now define a piecewise linear approximation to $u^*$ by

$$\text{(J.6)} \qquad u(x) = \sum_m \phi_m(x) u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\Big).$$

Then it follows that

$$
\begin{aligned}
|u^*(x) - u(x)| &\le \sum_m \left| \phi_m(x)\left(u^*(x) - u^*\left(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\right)\right)\right| \\
&\le \sum_m \left|u^*(x) - u^*\left(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\right)\right| \mathbb{1}\left\{m : \left|x_k - \frac{m_k}{N_k}\right| \le \frac{2}{3N_k}, \ k \in [p]\right\} \\
&\le 2^p \max_m \left|u^*(x) - u^*\left(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\right)\right| \mathbb{1}\left\{m : \left|x_k - \frac{m_k}{N_k}\right| \le \frac{2}{3N_k}, \ k \in [p]\right\} \\
&\le 2^p \max_m \sum_{k=1}^p \operatorname*{ess\,sup}_{x\in[0,1]^p} |\partial_k u^*(x)| \left|x_k - \frac{m_k}{N_k}\right| \mathbb{1}\left\{m : \left|x_k - \frac{m_k}{N_k}\right| \le \frac{2}{3N_k}, \ k \in [p]\right\} \\
&\le 2^p \sum_{k=1}^p \operatorname*{ess\,sup}_{x\in[0,1]^p} |\partial_k u^*(x)| \frac{2}{3N_k},
\end{aligned}
$$

where the first inequality follows from the triangular inequality, the second inequality is due to (J.4) and (J.5), the third inequality used the observation that any $x \in [0,1]^p$ belongs to the support of at most $2^d$ functions $\phi_m$, the forth inequality used Taylor's theorem of degree-0, and the last inequality holds from Hölder's inequality. Consequently, we obtain the following inequality

$$
\|u - u^*\|_{L^\infty([0,1]^p)} \le 2^p \sum_{k=1}^p \frac{1}{N_k} \|\partial_k u^*\|_{L^\infty([0,1]^p)}.
$$

*Step 2. Implement the piecewise linear function by a deep neural network.*

In this step, we implement the piececwise linear approximation (J.6) by a deep neural network. Using Lemmas J.3 and J.4, for each $m$, the function $\phi_m$ defined as (J.3) can be implemented by a deep neural network with $\lceil \log_2 p \rceil + 3$ layers and $16p + 6(p+1) = 22p + 6$ non-zero parameters. Since $u$ defined in (J.6) is a linear combination of $\prod_{k=1}^p (N_k + 1)$ functions $\phi_m$, it can be implemented by a deep neural network with $\lceil \log_2 p \rceil + 3$ layers and $(22p + 6)\prod_{k=1}^p (N_k + 1)$ non-zero parameters.

*Step 3. Compute the Lipschitz constant of the deep neural network.*

According to *Step 2*, the piecewise linear approximation $u$ can be implemented by a deep neural network with no error. Therefore, it suffices to compute the Lipschitz constant of $u$ in (J.6). Taking derivative on both sides of (J.6) with respect to $x_k$ yields

$$
\begin{aligned}
\partial_k u(x) &= \sum_m u^*\left(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\right) \partial_k \phi_m(x) \\
&= \sum_m u^*\left(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\right) A_k(m) \partial_k \psi\left(3N_k\left(x_k - \frac{m_k}{N_k}\right)\right) \\
\text{(J.7)} \qquad &= \sum_m u^*\left(\frac{m_1}{N_1}, \ldots, \frac{m_p}{N_p}\right) A_k(m) \partial_k \psi\left(3N_k\left(x_k - \frac{m_k}{N_k}\right)\right) \mathbb{1}\left\{m : \left|x_k - \frac{m_k}{N_k}\right| \le \frac{2}{3N_k}\right\},
\end{aligned}
$$

where the constant is given as

$$
A_k(m) = \prod_{\ell \ne k} \psi\left(3N_\ell\left(x_\ell - \frac{m_\ell}{N_\ell}\right)\right) \mathbb{1}\left\{m : \left|x_\ell - \frac{m_\ell}{N_\ell}\right| \le \frac{2}{3N_\ell}, \ \ell \ne k\right\}.
$$

It is evident that $0 \le A_k(m) \le 1$ for each $1 \le k \le p$ and $m$.

We next estimate (J.7) in the following cases.

(i) If there exists $m_k^* \in \{1, \ldots, N_k\}$ such that $|x_k - \frac{m_k^*}{N_k}| \leq \frac{1}{3N_k}$, then

$$\partial_k u(x) = u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*}{N_k}, \ldots, \frac{m_p}{N_p}\Big) A_k(m^*) \partial_k \psi\Big(3N_k\Big(x_k - \frac{m_k^*}{N_k}\Big)\Big) = 0,$$

where $m^* = (m_1, \ldots, m_k^*, \ldots, m_p)^T$.

(ii) If there exists $m_k^* \in \{1, \ldots, N_k\}$ such that $\frac{m_k^*}{N_k} + \frac{1}{3N_k} \leq x_k \leq \frac{m_k^*}{N_k} + \frac{2}{3N_k}$, then

$$\partial_k u(x) = u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*}{N_k}, \ldots, \frac{m_p}{N_p}\Big) A_k(m^*) \partial_k \psi\Big(3N_k\Big(x_k - \frac{m_k^*}{N_k}\Big)\Big)$$

$$+ u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*+1}{N_k}, \ldots, \frac{m_p}{N_p}\Big) A_k(m_+^*) \partial_k \psi\Big(3N_k\Big(x_k - \frac{m_k^*+1}{N_k}\Big)\Big)$$

$$= -3u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*}{N_k}, \ldots, \frac{m_p}{N_p}\Big) A_k(m^*) N_k$$

$$+ 3u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*+1}{N_k}, \ldots, \frac{m_p}{N_p}\Big) A_k(m_+^*) N_k$$

$$\leq 3N_k \Big| u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*+1}{N_k}, \ldots, \frac{m_p}{N_p}\Big) - u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*}{N_k}, \ldots, \frac{m_p}{N_p}\Big) \Big|$$

$$\leq 3 \operatorname*{ess\,sup}_{x \in [0,1]^p} |\partial_k u^*(x)|,$$

where $m_+^* = (m_1, \ldots, m_k^*+1, \ldots, m_p)^T$, the first inequality follows from the fact that $|A_k(m)| \leq 1$, and the last inequality is due to Taylor's theorem.

(iii) If there exists $m_k^* \in \{1, \ldots, N_k\}$ such that $\frac{m_k^*}{N_k} - \frac{2}{3N_k} \leq x_k \leq \frac{m_k^*}{N_k} - \frac{1}{3N_k}$, then by a same argument, we have

$$\partial_k u(x) \leq 3N_k \Big| u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*}{N_k}, \ldots, \frac{m_p}{N_p}\Big) - u^*\Big(\frac{m_1}{N_1}, \ldots, \frac{m_k^*-1}{N_k}, \ldots, \frac{m_p}{N_p}\Big) \Big|$$

$$\leq 3 \operatorname*{ess\,sup}_{x \in [0,1]^p} |\partial_k u^*(x)|.$$

Combining the three cases above, we obtain the following inequality

$$\|\partial_k u\|_{L^\infty([0,1]^p)} \leq 3\|\partial_k u^*\|_{L^\infty([0,1]^p)}, \quad 1 \leq k \leq p.$$

This completes the proof. $\qquad \square$

We have investigated the approximation error of a target function on the hypercube $[0,1]^p$ in Lemma J.5. In the following corollary, we extend our analysis to target functions on general bounded domain $[0, T] \times [-R, R]^d$.

**Corollary J.6.** *Let $p \in \mathbb{N}_+$, $\mathcal{X} = \prod_{k=1}^p [a_i, b_i]$, and let $\{N_k\}_{k=1}^p$ be a set of positive integer. Set the deep neural network class $N(L, S)$ as $L = \lceil \log_2 p \rceil + 3$ and $S = (22p + 6) \prod_{k=1}^p (N_k + 1)$. Then for each $u^* \in W^{1,\infty}(\mathcal{X})$, there exists a deep neural network $u \in N(L, S)$ such that*

$$\|u - u^*\|_{L^\infty(\mathcal{X})} \leq 2p \sum_{k=1}^p \frac{b_k - a_k}{N_k} \|\partial_k u^*\|_{L^\infty(\mathcal{X})}.$$

*Further, it holds that for each $1 \leq k \leq p$:*

$$\|u\|_{L^\infty(\mathcal{X})} = \|u^*\|_{L^\infty(\mathcal{X})} \quad and \quad \|\partial_k u\|_{L^\infty(\mathcal{X})} \leq 3\|\partial_k u^*\|_{L^\infty(\mathcal{X})}.$$

*Proof of Corollary J.6.* We first define a variable transformation on $u^* \in W^{1,\infty}(\mathcal{X})$ as

$$\phi : W^{1,\infty}(\mathcal{X}) \to W^{1,\infty}([0,1]^p)$$

$$u^*(x) \mapsto (\phi \circ u^*)(x') = u^*(a + (b-a)x'),$$

where $(b-a)x' = ((b_k - a_k)x'_k)_{k=1}^p \in \mathbb{R}^p$. Then it is noticeable that

$$\operatorname*{ess\,sup}_{x' \in [0,1]^p} \left| \frac{\partial(\phi \circ u^*)}{\partial x'_k}(x') \right| = (b_k - a_k) \operatorname*{ess\,sup}_{x \in \mathcal{X}} \left| \frac{\partial u^*}{\partial x_k}(x) \right|.$$

Set the deep neural network class $N(L, S)$ as $L = \lceil \log_2 p \rceil + 3$ and $S = (22p + 6) \prod_{k=1}^p (N_k + 1)$. According to Lemma J.5, there exists a neural network $u \in N(L, S)$ such that

$$|u(x') - (\phi \circ u^*)(x')| \le 2^p \sum_{k=1}^p \operatorname*{ess\,sup}_{x' \in [0,1]^p} \left| \frac{\partial(\phi \circ u^*)}{\partial x'_k}(x') \right| \frac{1}{N_k}$$

(J.8)
$$\le 2^p \sum_{k=1}^p \frac{b_k - a_k}{N_k} \operatorname*{ess\,sup}_{x \in \mathcal{X}} \left| \frac{\partial u^*}{\partial x_k}(x) \right|,$$

and for each $1 \le k \le p$, the following inequality holds:

(J.9)
$$\operatorname*{ess\,sup}_{x' \in [0,1]^p} \left| \frac{\partial u}{\partial x'_k}(x') \right| \le 3 \operatorname*{ess\,sup}_{x' \in [0,1]^p} \left| \frac{\partial(\phi \circ u^*)}{\partial x'_k}(x') \right| = 3(b_k - a_k) \operatorname*{ess\,sup}_{x \in \mathcal{X}} \left| \frac{\partial u^*}{\partial x_k}(x) \right|.$$

We next define the inverse transform on $u \in W^{1,\infty}([0,1]^p)$ as

$$\psi : W^{1,\infty}([0,1]^p) \to W^{1,\infty}(\mathcal{X})$$

$$u(x') \mapsto (\psi \circ u)(x) = u\left( \frac{x' - a}{b - a} \right),$$

where $(x' - a)/(b - a) = ((x'_k - a_k)/(b_k - a_k))_{k=1}^p \in \mathbb{R}^p$. It follows from (J.9) that

$$\operatorname*{ess\,sup}_{x \in \mathcal{X}} \left| \frac{\partial(\psi \circ u)}{\partial x_k}(x) \right| = \frac{1}{b_k - a_k} \operatorname*{ess\,sup}_{x' \in [0,1]^p} \left| \frac{\partial u}{\partial x'_k}(x') \right| \le 3 \operatorname*{ess\,sup}_{x \in \mathcal{X}} \left| \frac{\partial u^*}{\partial x_k}(x) \right|,$$

for each $1 \le k \le p$. Then composing $\psi$ on both sides of (J.8) yields the desired inequality. $\square$

## APPENDIX K. DENOISER PARAMETERIZATION

In practice, we parameterize the network $D_\theta(t, x)$ following Karras et al. (2022):

(K.1)
$$D_\theta(t, x) = c_{\text{skip}}(t) x + c_{\text{out}}(t) F_\theta \left( c_{\text{noise}}(t), c_{\text{in}}(t) x \right),$$

where $F_\theta$ is the neural network to be trained, $c_{\text{skip}}(t)$ scale the skip connection, $c_{\text{in}}(t)$ and $c_{\text{out}}(t)$ scale the input and output of $F_\theta$, and $c_{\text{noise}}(t)$ scales time $t$.

Now (4.1) becomes

(K.2)
$$\mathcal{L}(F) = \int_0^1 \mathbb{E}_{X_0} \mathbb{E}_{X_1} \left[ \omega(t) \| c_{\text{skip}}(t) X_t + c_{\text{out}}(t) F \left( c_{\text{noise}}(t), c_{\text{in}}(t) X_t \right) - x_1 \|^2 \right] \, dt$$

$$= \int_0^1 \mathbb{E}_{X_0} \mathbb{E}_{X_1} \left[ \omega(t) c_{\text{out}}^2(t) \| F \left( c_{\text{noise}}(t), c_{\text{in}}(t) X_t \right) - \frac{X_1 - c_{\text{skip}}(t) X_t}{c_{\text{out}}(t)} \|^2 \right] \, dt$$

$$= \int_0^1 \mathbb{E}_{X_0} \mathbb{E}_z \left[ \lambda(t) \| F_{\text{pred}} - F_{\text{target}} \|^2 \right] \, dt,$$

where $\lambda(t) = \omega(t) c_{\text{out}}^2(t)$, $F_{\text{pred}} = F \left( c_{\text{noise}}(t), c_{\text{in}}(t) X_t \right)$ and $F_{\text{target}} = \frac{X_1 - c_{\text{skip}}(t) X_t}{c_{\text{out}}(t)}$.

Let $\sigma$ denote the standard deviation of $\mu_1$. We now design $c_{\text{in}}$ so that the spatial inputs of $F$ has unit variance.

$$\text{Var}[c_{\text{in}}(t)x_t] = 1,$$

$$\Leftrightarrow c_{\text{in}}(t) = \sqrt{\frac{1}{\text{Var}[X_t]}} = \sqrt{\frac{1}{\alpha_t^2\sigma^2 + \beta_t^2}}.$$

We then design $c_{\text{out}}$ so that the $F_{\text{target}}$ has unit variance.

$$\text{Var}[\frac{X_1 - c_{\text{skip}}(t)X_t}{c_{\text{out}}(t)}] = 1,$$

$$\Leftrightarrow c_{\text{out}}(t) = \sqrt{\text{Var}[(1 - \alpha_t c_{\text{skip}}(t))X_1 - c_{\text{skip}}(t)\beta_t z]},$$

(K.3)
$$\Leftrightarrow c_{\text{out}}(t) = \sqrt{(1 - \alpha_t c_{\text{skip}}(t))^2\sigma^2 + c_{\text{skip}}(t)^2\beta_t^2}.$$

We then design $c_{\text{skip}}$ that minimizes $c_{\text{out}}$ so that the errors of $F$ are amplified as little as possible. Let $\frac{\partial c_{\text{out}}^2(t)}{\partial c_{\text{skip}}(t)} = 0$, we obtain

$$-\alpha_t(1 - \alpha_t c_{\text{skip}}(t))\sigma^2 + \beta_t^2 c_{\text{skip}}(t) = 0,$$

(K.4)
$$\Leftrightarrow c_{\text{skip}}(t) = \frac{\alpha_t\sigma^2}{\alpha_t^2\sigma^2 + \beta_t^2}.$$

We can check that (K.4) is indeed the minima of $c_{\text{out}}$. Meanwhile, (K.3) yields

$$c_{\text{out}}(t) = \frac{\beta_t\sigma}{\sqrt{\alpha_t^2\sigma^2 + \beta_t^2}}.$$

We can then design $\omega(t)$ so that $\lambda(t) = 1$ uniformly on $[0, 1]$.

$$\omega(t) = \frac{1}{c_{\text{out}}^2(t)} = \frac{\alpha_t^2\sigma^2 + \beta_t^2}{\beta_t^2\sigma^2}.$$

We conclude the form of coefficients in Table 8.

TABLE 8. Denoiser parameterization.

| function | requirements | form |
|---|---|---|
| $c_{\text{noise}}(t)$ | - | free choice |
| $c_{\text{in}}(t)$ | $\text{Var}[c_{\text{in}}(t)X_t] = 1$ | $\sqrt{\frac{1}{\alpha_t^2\sigma^2 + \beta_t^2}}$ |
| $c_{\text{out}}(t)$ | $\text{Var}[\frac{X_1 - c_{\text{skip}}(t)X_t}{c_{\text{out}}(t)}] = 1$ | $\frac{\beta_t\sigma}{\sqrt{\alpha_t^2\sigma^2 + \beta_t^2}}$ |
| $c_{\text{skip}}(t)$ | $\frac{\partial c_{\text{out}}^2(t)}{\partial c_{\text{skip}}(t)} = 0$ | $\frac{\alpha_t\sigma^2}{\alpha_t^2\sigma^2 + \beta_t^2}$ |
| $\omega(t)$ | $\lambda(t) = 1$ | $\frac{\alpha_t^2\sigma^2 + \beta_t^2}{\beta_t^2\sigma^2}$ |

Now (K.2) can be used as the working denoiser matching loss.

TABLE 9. Hyperparameters.

|  | MNIST | CIFAR-10 | CelebAHQ-256 | CelebAHQ-512 |
|---|---|---|---|---|
| encoder/decoder | ✗ | ✗ | ✓ | ✓ |
| number of GPUs | $4 \times$ A800 | $4 \times$ A800 | $12 \times$ A100 | $8 \times$ A100 |
| architecture | DDPM++ | DDPM++ | DiT-B/2 | DiT-XL/2 |
| channel multiplier | 32 | 128 | - | - |
| channels per resolution | 1, 2, 2 | 1, 2, 2, 2 | - | - |
| model size (GB) | 0.02 | 0.5 | 0.25 | 1.3 |
| optimizer | RAdam | RAdam | AdamW | AdamW |
| learning rate | 1e-3 | 1e-4 | 1e-4 | 1e-4 |
| LR ramp-up (Mimg) | 0.5 | 0.5 | 10 | 10 |
| EMA rate | 0.999 | 0.9999 | - | - |
| EMA half-life (Mimg) | - | - | 0.5 | 0.5 |
| teacher duration (Mimg) | 10 | 100 | - | - |
| student duration (Mimg) | 2 | 2 | 100 | 15 |
| batch size | 1024 | 1024 | 192 | 128 |
| stochastic interploants | Föllmer | Föllmer | Linear | Linear |
| global loss metric | LPIPS [1] | LPIPS | $L^2$ | $L^2$ |
| adaptive weightning [2] | ✓ | ✓ | ✗ | ✗ |
| NFE of teacher (max) | 19 | 19 | 2 | 2 |
| dropout | 0.1 | 0.1 | 0.13 | 0.13 |

APPENDIX L. EXTRA EXPERIMENT DETAILS

This section reports the detailed settings of the experiments conducted on image datasets. The details are shown in Table 9. In practice, when there is an available teacher network $D_{\mathcal{T}}$, we initialize $D_{\mathcal{S}}$ from the pre-trained $D_{\mathcal{T}}$, and then embed another temporal input into it to construct the characteristic generator as Kim et al. (2024) did. For the choice of $u$ in (4.8), we apply a weighted random strategy which is more likely to choose a long range interval $[t, u]$ to ensure the precision of the teacher solver, as Kim et al. (2024) did. Inspired by Esser et al. (2021), Kim et al. (2024), we adaptively balance the global characteristic matching loss and the local denoiser matching loss on MNIST and CIFAR-10. We calculate FID in original pixel space for MNIST, and in the feature space (dimension=2048) extracted by a pre-trained Inceptionv3 network for other image datasets.

---

[1]Zhang et al. (2018)

[2]Esser et al. (2021), Kim et al. (2024)

## Appendix M.  Remarks on Time Consumption

For numerical experiments on the MNIST and CIFAR-10 dataset, a pre-trained denoiser network (teacher model) is explicitly involved, to generate the reference trajectories, from which the characteristic generator is distilled.

Take the CIFAR-10 dataset for example. The total time consumption is summarized as follows: we first spend around 40 NVIDIA A800 hours during the pre-train stage of the denoiser network. During the training of the characteristic generator, the maximum number of steps to generate each reference trajectory is set to 19. On average, the reference trajectory is generated on-the-fly with speed 0.007 sec per image. In total, the training time of the characteristic generator is around 3 NVIDIA A800 hours, 38% of which is spent on generating the reference trajectory from the teacher model.

For the CelebA 256 × 256 dataset, we first spend around 78 NVIDIA A800 hours during the pre-train stage of the denoiser network. Then we use the pre-trained result from the first stage as the source of parameter initialization and teacher model. The training time of characteristic generator is around 480 NVIDIA A800 hours, around 45% of which are spent on generating a reference trajectory, by evaluating the characteristic generator itself twice.

For the CelebA 512 × 512 dataset, there is no pre-training stage. The training time of characteristic generator is around 480 NVIDIA A800 hours.

It is important to note that these training times are one-time offline costs. Once trained, the characteristic generator can significantly reduce the number of function evaluations (NFE) during inference. As demonstrated in Table 5 of Section 4.2, traditional sampling methods for diffusion models require hundreds or thousands of NFE: DDPM requires 1000 NFE, DDIM requires 100 NFE, and Score SDE requires 2000 NFE to achieve their reported FID scores. In contrast, our characteristic generator achieves competitive performance with substantially fewer NFE. For instance, CG achieves FID of 4.59 with only NFE=1, and FID of 2.83 with NFE=4, which is already comparable to Score SDE (FID=2.20) that requires 2000 NFE. This demonstrates that our approach enables significantly faster inference by reducing the number of function evaluations by orders of magnitude. Thus, the offline training cost is an one-time investment that enables efficient inference once and for all.

## Appendix N.  Comparison with Consistency Trajectory Models

In this section, we discuss the implementation details compared with the Consistency Trajectory Models (CTM) (Kim et al., 2024).

N.1. **Similarities.**

- Our work and the CTM share similar neural network initialization strategy (from a pre-trained teacher model, if any).
- On MNIST and CIFAR-10, where the teacher and student model are two different networks, our work and CTM shares similar strategies for generating reference solutions from the teacher model: running some ODE solver on the given time interval, with a threshold on the maximum number of steps allowed.

- Also on MNIST and CIFAR-10, we and CTM both use the Learned Perceptual Image Patch Similarity (LPIPS) metric instead of $L^2$-norm to measure distance in the feature space instead of in the original pixel space (Zhang et al., 2018). We also share similar loss balance techniques to balance between the global loss and the local loss.

N.2. **Differences.**

- We extend the distillation framework to the latent space on CelebAHQ-256 and CelebA-512, significantly reducing the training and generating cost, while CTM only experiments on the original pixel space. We also manage to shift from the U-Net architecture to the DiT architecture.

- On CelebAHQ, We extend the distillation framework to self-distillation, and offer a candidate design for the self-guided teacher model, which takes only 2 NFE to generate a reference solution, while CTM still needs dozens of NFEs. Under this setting, the training process of our proposed method would be significantly faster than CTM.

- The incorporation of GAN contributes a lot to the low FID of CTM, but we find it introduces extra training instability and requires some manual tuning. Our work is more like a clean and minimal template for the distillation-based diffusion models, and the experiments results indicates that our implementation works well enough and could be easily adapted to downstream tasks. We believe that less is more when it comes to high-level framework.

APPENDIX O. COMPARISONS BETWEEN ONE-STEP EULER AND CHARACTERISTIC GENERATOR

- Error bound for one-step Euler generation. Recall the error bound for the Euler sampling in Theorem III.11:

$$\text{Error of Euler} \leq C\kappa^2(T)\Big\{n^{-\frac{2}{d+3}}\log^2 n + \frac{\log n}{K_{\text{Euler}}^2}\Big\},$$

where $K_{\text{Euler}}$ is the number of step of the Euler sampling. One-step Euler generation means $K_{\text{Euler}} = 1$, thus

$$\text{Error of one-step Euler} \leq C\kappa^2(T)\Big\{n^{-\frac{2}{d+3}}\log^2 n + \log n\Big\}.$$

The error is dominated by the large $\log n$ term, and does not converge even for sufficiently large number of samples $n$. This large inherent error from using a single, coarse step explains the poor image quality seen in Figure 4.

- Error bound for one-step characteristic generation. The characteristic generator is fundamentally different. It is a one-step model that is trained to approximate the result of a high-quality, multi-step numerical sampler. In the characteristic learning, the generator are learned from the trajectories generated using an Euler sampler with $K_{\text{CG}}$ steps. Recall the error bound for the characteristic generator in Theorem III.13:

$$\text{Error of CG} \leq C\kappa^2(T)\Big\{n^{-\frac{2}{d+3}}\log^2 n + \frac{\log n}{K_{\text{CG}}^2}\Big\} + C\Big\{m^{-\frac{2}{d+4}}\log^2 m + \frac{\log m}{K_{\text{CG}}}\Big\}.$$

In practical, we can choose sufficiently large $K_{\mathrm{CG}}$. As a result, the final error of the one-step CG is dominated by the sample complexity terms:

$$\text{Error of CG} \leq C\kappa^2(T) n^{-\frac{2}{d+3}} \log^2 n + Cm^{-\frac{2}{d+4}} \log^2 m,$$

which converge to zero as the dataset sizes ($n$ and $m$) increase.

In essence, the CG distills the knowledge of a precise, multi-step solver into a fast, single-step network. Our theory correctly predicts that its error should be much lower than that of a naive one-step Euler sampler, which aligns perfectly with our empirical findings in Figure 4.