WHEN DIFFERENTIAL EQUATIONS MEET GENERATIVE MODELING: REGULARITY, APPROXIMATION, AND CONVERGENCE

YUAN GAO

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Data Science and Artificial Intelligence

When Differential Equations Meet Generative Modeling: Regularity, Approximation, and Convergence

Yuan Gao

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

April 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Yuan Gao

Abstract

In recent years, continuous generative models based on ordinary differential equations (ODEs) and stochastic differential equations (SDEs) have played a central role in the rapidly expanding field of generative AIs. These generative AIs have shown remarkable empirical success across various applications, including large-scale image synthesis, protein structure prediction, and molecule generation. In this thesis, we aim to investigate the theoretical properties of these continuous generative models by considering the regularity of the differential equations, the ability to approximate them with deep neural networks, and the non-asymptotic convergence rate of these continuous generative models.

In the first part, we address the regularity of a class of simulation-free continuous normalizing flows (CNFs) constructed with ODEs. Through a unified framework of the flow models termed Gaussian interpolation flows, we establish the Lipschitz regularity of the flow velocity field, the existence and uniqueness of the flow, and the Lipschitz continuity of the flow map and the time-reversed flow map for several rich classes of target distributions. This analysis also sheds light on the auto-encoding and cycle consistency properties of Gaussian interpolation flows. Our findings offer valuable insights into the learning techniques and accumulations of errors when employing Gaussian interpolation flows for generative modeling.

In the second part, we study the theoretical properties of continuous normalizing flows with linear interpolation in learning probability distributions from a finite random sample, using a flow-matching objective function. We establish non-asymptotic error bounds for the distribution estimator based on CNFs, in terms of the Wasserstein-2 distance. We present a convergence analysis framework that encompasses the error due to velocity estimation, the discretization error, and the early stopping error. A key step in our analysis involves establishing the regularity properties of the velocity field and its estimator for CNFs constructed with linear interpolation. This necessitates the development of uniform error bounds with Lipschitz regularity control of deep ReLU networks that approximate the Lipschitz function class. Our nonparametric convergence analysis offers theoretical guarantees for using CNFs to learn probability distributions from a finite random sample.

The last part of the thesis addresses the convergence properties of a Bayesian fine-tuning approach for large diffusion models. Diffusion models are a class of continuous generative models built with SDEs whose generation ability has been largely reinforced by various fine-tuning procedures. However, the mystery of fine-tuning has seldom been uncovered from a statistical perspective. In this part, we address the gap in the systematic understanding of the advantages of fine-tuning mechanisms from a statistical perspective. We prove that a pre-trained large diffusion model can gain a faster convergence rate from the Bayesian fine-tuning procedure when adapted to perform conditional generation tasks. This improvement in the convergence rate justifies that a pre-trained large diffusion model would perform better on a downstream conditional generation task than a standard conditional diffusion model, whenever an appropriate fine-tuning procedure is implemented.

Acknowledgments

Throughout my PhD study, I have received invaluable support and company from people within and outside of PolyU. If possible, this thesis is dedicated to all of them.

First of all, I would like to express my deep and sincere gratitude to my supervisor, Professor Jian Huang, for all the consistent support along the way and through my ups and downs. Professor Huang has provided me with patient guidance and excellent expertise, which has shaped my thinking, my learning, and my research taste.

Then, I wish to thank Professor Binyan Jiang, Professor Junhui Wang, and Professor Xueqin Wang for their valuable and constructive comments, which have significantly improved the quality of this thesis.

Furthermore, I am very grateful to my collaborators – Ding Huang, Professor Yuling Jiao, Professor Ting Li, and Professor Shurong Zheng, without whom this thesis would not have such accomplishment.

In addition, I feel fortunate to meet our talented and hardworking group members. Their stimulating and enjoyable discussions have been fruitful to me during these years. I also would like to thank my friends around the university and even the world for enriching my life during these years.

Last but not least, I would like to express my sincere gratitude to my family for their unconditional love and support throughout my life. Without them, I would have never made it this far.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Let $\{X_i\}_{i=1}^n$ be independent and identically distributed (i.i.d.) random variables drawn from an underlying probability distribution ν with support in $\mathbb{R}^d$. The task of generative learning is to learn ν from the data $\{X_i\}_{i=1}^n$ by generating new samples [Salakhutdinov, 2015]. Several generative learning methods have been developed during the recent decade, including generative adversarial networks [Goodfellow et al., 2014], variational auto-encoders [Kingma and Welling, 2014], diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Song et al., 2021b], and normalizing flows [Tabak and Turner, 2013, Rezende and Mohamed, 2015, Chen et al., 2018]. Deep neural networks [LeCun et al., 2015], as a powerful modeling tool, have played an important role in the development of these methods.

Among the generative learning methods, continuous normalizing flows (CNFs) use ordinary differential equations (ODEs) to determine a stochastic process for transporting a Gaussian distribution to the target distribution, achieving the goal of generative learning. CNFs have achieved impressive empirical performance across various applications. These applications include large-scale image synthesis [Ma et al., 2024], protein structure prediction [Jing et al., 2023], and 3D molecule generation [Song et al., 2023b]. Rectified flow [Liu et al., 2023], a CNF model that linearly interpolates Gaussian noise and data, has been recently implemented in the large image model Stable Diffusion 3 [Esser et al., 2024]. Simulation-free CNFs that use flow matching to learn probability distributions have been the focus of much recent attention [Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023, Liu et al., 2023, Neklyudov et al., 2023].

An early model of CNFs was proposed by Chen et al. [2018]. This model is based on neural ODEs and employs a simulation-based maximum likelihood method to estimate velocity fields. However, simulation-based CNFs are computationally demanding in large-scale applications. To address the computational challenges of simulation-based CNFs, significant efforts have been made to develop simulation-free CNFs, where ve-

locity fields can be represented in terms of conditional expectations. Noteworthy examples include the probability flows of diffusion models [Song et al., 2021b] and denoising diffusion implicit models [Song et al., 2021a], which are trained using a denoising score matching objective function. In contrast, the flow matching method solves a least squares problem to estimate the conditional expectation that represents the velocity field [Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023, Liu et al., 2023].

The essence of CNFs lies in defining ODEs that govern the evolution of CNFs in terms of continuous trajectories. Inspired by the Gaussian denoising approach, which learns a target distribution by denoising its Gaussian smoothed counterpart, many authors have considered simulation-free estimation methods that have shown great potential in large-scale applications [Song et al., 2021a, Liu et al., 2023, Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023, Neklyudov et al., 2023, Tong et al., 2023, Chen and Lipman, 2023, Albergo et al., 2023b, Shaul et al., 2023, Pooladian et al., 2023]. However, despite the empirical success of simulation-free CNFs based on Gaussian denoising, a rigorous theoretical analysis of these CNFs has received limited attention thus far.

One target of this thesis is to explore an ODE flow-based approach for generative modeling, which we refer to as Gaussian interpolation flows (GIFs). This method is derived from the Gaussian stochastic interpolation. GIFs represent a straightforward extension of the stochastic interpolation method [Albergo and Vanden-Eijnden, 2023, Liu et al., 2023, Lipman et al., 2023]. They can be considered a class of CNFs and encompass various ODE flows as special cases. According to the classical Cauchy-Lipschitz theorem, also known as the Picard-Lindelöf theorem [Hartman, 2002b, Theorem 1.1], a unique solution to the initial value problem for an ODE flow exists if the velocity field is continuous in the time variable and uniformly Lipschitz continuous in the space variable. In the case of GIFs, the velocity field depends on the score function of the push-forward measure. Therefore, it remains to be shown that this velocity field satisfies the regularity conditions stipulated by the Cauchy-Lipschitz theorem. These regularity conditions are commonly assumed in the literature when analyzing the convergence properties of CNFs or general neural ODEs [Chen et al., 2018, Biloš et al., 2021, Marion et al., 2023, Marion, 2023, Marzouk et al., 2023]. However, there is a theoretical gap in understanding how to translate these regularity conditions on velocity fields into conditions on target distributions.

When a random sample from the target distribution is available, the process of learning simulation-free CNFs involves statistically and numerically solving a class of ODE-based initial value problems (IVPs). Let ν denote an easy-to-sample source distribution. We consider the IVP on the unit time interval

$$\frac{\mathrm{d}X_t}{\mathrm{d}t}(x) = v(t, X_t(x)), \quad X_0(x) = x \sim \mu, \quad (t, x) \in [0, 1] \times \mathbb{R}^d, \tag{1.1}$$

where v represents the velocity field, which can be estimated based on data. The solution to the IVP (1.1) is a family of flow maps $(X_t)_{t\in[0,1]}$ indexed by the time variable, which generates a smoothing path between the source and target distributions. Gao et al. [2024a] have studied the mathematical properties of these ODE-based IVPs under the framework of Gaussian interpolation flow. This defines a Gaussian smoothing path as a transport map that pushes forward a Gaussian distribution onto the target distribution in terms of measure transport.

Simulation-free CNFs adopt a two-step "estimation-then-simulation" approach to learning the desired transport map based on a random sample. In the estimation stage, a deep learning model is trained to estimate the velocity field without simulating the ODE that defines the CNF. During the simulation stage, numerical solvers simulate the numerical solution of the ODE associated with the estimated velocity field, and the generated data is collected at the end time point. Another target of this thesis is to establish statistical convergence guarantees for these simulation-free CNFs in terms of error bounds of distribution learning. These convergence guarantees are necessary to broaden applications of simulation-free CNFs in statistical and machine learning methods, such as transfer learning, statistical hypothesis testing, and semi-supervised learning.

In addition to the ODE-based flow models, a lot of efforts have been made to the development of diffusion models that are built on stochastic differential equations (SDEs). Diffusion models are a promising approach to deep generative modeling that has evolved rapidly since its emergence [Song and Ermon, 2019, 2020, Ho et al., 2020, Song et al., 2021b,a]. The basis of diffusion models lies in the notion of the score function, which characterizes the gradient of the log-density function of a given distribution. Compared with learning the law of a random vector with the principle of generative modeling, conditional generative modeling, which learns the law of a random vector given another

one, has gained more interest among practical generation tasks. To tackle a conditional generation task, we need to resort to conditional diffusion models that are defined by conditional score functions. The required conditional score is directly linked with the unconditional score of the unconditional diffusion model due to the classical Bayes' rule. Such a relation between the score and the conditional score has inspired the proposal of a Bayesian fine-tuning approach [Ho and Salimans, 2022, Huang et al., 2024]. This fine-tuning approach involves taking an unconditional diffusion model that has already been trained on a broad dataset and refining it using a smaller, task-specific dataset for conditional generation. The goal is to use the general knowledge embedded in the large diffusion model while tailoring its capabilities to meet particular needs. This kind of fine-tuning not only enhances performance on specialized conditional generation tasks but also reduces the computational resources and time required compared to training a model from scratch. Due to such observations, we focus on investigating the benefits of the Bayesian fine-tuning approach from a statistical perspective. These theoretical investigations form the third part of the thesis.

1.1 Main contributions

This thesis is based on our recent work [Gao et al., 2024a,b] and an ongoing work jointly with Ding Huang, Jian Huang, and Ting Li. We summarize the main contributions into three parts.

1.1.1 Regularity analysis of CNFs constructed with stochastic interpolation

The nature of CNFs is an ODE with a random starting point. To establish the well-posedness properties of these CNFs, we resort to the classical Cauchy-Lipschitz theorem. We first show that the regularity conditions of the Cauchy-Lipschitz theorem are satisfied for several rich classes of probability distributions using variance inequalities. Based on the obtained regularity results, we further expose the well-posedness of GIFs, the Lipschitz continuity of flow mappings, and applications to generative modeling. The well-posedness results are crucial for studying the approximation and convergence properties of GIFs learned with the flow or score matching method. When applied to gen-

erative modeling, our results further elucidate the auto-encoding and cycle consistency properties exhibited by GIFs.

Related work. There is a series of papers exploring the idea of Gaussian denoising for constructing continuous normalizing flows for generative modeling [Song et al., 2021b,a, Liu et al., 2023, Albergo and Vanden-Eijnden, 2023, Albergo et al., 2023b, Neklyudov et al., 2023, Tong et al., 2023, Chen and Lipman, 2023, Albergo et al., 2023b, Shaul et al., 2023, Pooladian et al., 2023, Albergo et al., 2023a,c]. Most of them focus on the modeling and computation aspects of the flow models. For the target of analyzing the regularity properties of flow models, we find a substantial body of research on the Lipschitz properties of transport maps is closely related to ours. The celebrated Caffarelli's contraction theorem [Caffarelli, 2000, Theorem 2] establishes the Lipschitz continuity of optimal transport maps that push the standard Gaussian measure onto a log-concave measure. Colombo et al. [2017] study a Lipschitz transport map between perturbations of log-concave measures using optimal transport theory. Mikulincer and Shenfeld [2024] demonstrate that the Brownian transport map, defined by the Föllmer process, is Lipschitz continuous when it pushes forward the Wiener measure on the Wiener space to the target measure on the Euclidean space. Additionally, Neeman [2022] and Mikulincer and Shenfeld [2023] prove that the transport map along the reverse heat flow of certain target measures is Lipschitz continuous. Our analysis is based on establishing similar regularity properties of the GIFs. We show that GIFs share similar Lipschitz continuity properties using the techniques developed in the literature on Lipschitz properties of transport maps.

1.1.2 Convergence analysis of flow matching for learning CNFs

We contribute to conducting a non-asymptotic convergence analysis of CNFs learned with the simulation-free flow matching approach. We develop a general framework for error analyses of CNFs with flow matching for learning probability distributions based on a random sample. Central to simulation-free CNFs, deep ReLU networks are employed for function approximation and nonparametric estimation of the velocity field. We establish the approximation properties of deep ReLU networks with Lipschitz regularity control, which is essential for analyzing the impact of the estimated velocity field on the distribution of the data generated through the flow. In particular, it is cru-

cial to control the Lipschitz regularity of the estimated velocity field to ensure that the associated IVP is well-posed.

Related work. In existing literature, it is typical to assume strong regularity conditions directly on the velocity field (or score function) and its estimator. Moreover, current studies often only consider certain sources of errors, neglecting either the discretization error or the estimation error of the velocity field (or score function). For example, Albergo and Vanden-Eijnden [2023] used a Lipschitz assumption for the estimated velocity field. Chen et al. [2023e] considered second-order smoothness in the space variable and Hölder-type regularity in the time variable for the score function, and their analysis ignored the score estimation error. Chen et al. [2023c] assumed that the score function and the score estimator both have Lipschitz regularity in the space variable and that the score estimation error is sufficiently small in the L^2 distance.

In our study, we conduct an end-to-end convergence analysis of the CNF distribution estimator with flow matching. Furthermore, we only stipulate general assumptions on the target distribution, rather than making assumptions on the velocity field (or score function) and its estimator.

1.1.3 Statistical analysis of a Bayesian fine-tuning approach

From a statistical perspective, we provide a systematic investigation of the pre-training and fine-tuning mechanisms for diffusion models. We consider Stable Diffusion – a cutting-edge open-source large image model, and the Bayesian fine-tuning approach [Ho and Salimans, 2022, Huang et al., 2024] that is widely used in diffusion models and has demonstrated effectiveness in numerous experiments.

We prove that, under some regularity conditions, the Bayesian fine-tuning approach achieves the convergence rate $m^{-\frac{2\beta}{d+2\beta}} \vee n^{-\frac{2\alpha}{d+k+2\alpha}}$, where m is the sample size of pre-training, n is the labeled data size for fine-tuning, and β, α are smoothness indices. Then, if we train a conditional diffusion model from scratch using only the labeled data, the convergence rate is $n^{-\frac{2\delta}{d+k+2\delta}}$ with $\delta \leq \min(\alpha, \beta)$. Our result rigorously shows the benefit of pre-training when we have abundant data ($m >> n$) from the prior data space.

Related work. The idea of fine-tuning diffusion models can be dated back to the approaches termed classifier guidance [Dhariwal and Nichol, 2021] and classifier-free

guidance [Ho and Salimans, 2022]. More generally, the guidance plays a central role in steering the samples generated by diffusion models toward a desired property. Subsequently, a series of work [Mou et al., 2024, Zhang et al., 2023, Huang et al., 2024] explore flexible deep learning models to add guidance to the generation process of a pre-trained unconditional model. Meanwhile, several fine-tuning approaches based on reinforcement learning are proposed to achieve the goal of guidance [Black et al., 2024, Fan et al., 2023].

The target of our statistical analysis is to justify that conditional diffusion models can perform better with the proper usage of additional unlabelled data. In the literature, there are several papers studying the convergence of conditional diffusion models learned with classifier or classifier-free guidance [Fu et al., 2024, Wu et al., 2024]. However, their works do not focus on justifying the efficiency of fine-tuning diffusion models.

The rest of this thesis is organized as follows. Chapter 2 introduces Gaussian interpolation flows, and we conduct regularity studies of the flows throughout this chapter. In Chapter 3, we study the convergence properties of CNFs based on flow matching in learning probability distributions from a finite random sample. In Chapter 4, we provide a statistical investigation into the Bayesian fine-tuning approach. Finally, Chapter 5 concludes the thesis and presents a further discussion on it.

1.2 Notations

Here we summarize the notations used throughout this thesis.

Number. For two numbers $X, Y \in \mathbb{R}$, we use $X \lesssim Y$ and $Y \gtrsim X$ to denote $X \leq CY$ for some constant $C > 0$. The notation $X \asymp Y$ indicates that $X \lesssim Y \lesssim X$. For $X, Y \in \mathbb{R}$, we denote $X \vee Y := \max\{X, Y\}$.

Vector. We use $\|x\|_p$ to denote the ℓ_p-norm of a vector $x \in \mathbb{R}^d$ for $p \in [1, \infty]$. Especially, we use $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ to denote the Euclidean metric and the corresponding inner product. For two vectors $x, y \in \mathbb{R}^d$, we denote $x \otimes y := xy^\top$.

Matrix. For a matrix $A \in \mathbb{R}^{k \times d}$, we use $A^\top$ for the transpose, and the spectral norm is denoted by $\|A\|_{2,2} := \sup_{x \in \mathbb{S}^{d-1}} \|Ax\|$. For a square matrix $A \in \mathbb{R}^{d \times d}$, we use $\det(A)$

for the determinant and $\mathrm{Tr}(A)$ for the trace. We use I_d to denote the $d \times d$ identity matrix. For two symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, we denote $A \succeq B$ or $B \preceq A$ if $A - B$ is positive semi-definite.

Set. Let $\mathbb{N}_0$ and $\mathbb{N}$ denote the set of non-negative integers and the set of positive integers, respectively, that is, $\mathbb{N} = \{1, 2, 3, \cdots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For an integer $N \in \mathbb{N}_0$, let $[N] := \{0, 1, ..., N\}$. Let $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, $\mathbb{B}^d(x_0, r, \|\cdot\|_p) := \{x \in \mathbb{R}^d : \|x - x_0\|_p < r\}$, and $\bar{\mathbb{B}}^d(x_0, r, \|\cdot\|_p) := \{x \in \mathbb{R}^d : \|x - x_0\|_p \le r\}$. For a set $\Omega \subset \mathbb{R}^d$, let $\Omega^c := \{x \in \mathbb{R}^d : x \notin \Omega\}$, and we use $\mathrm{Id}_\Omega : \mathbb{R}^d \to \{0, 1\}$ to denote the indicator function of Ω.

Function. For $\Omega_1 \subset \mathbb{R}^k, \Omega_2 \subset \mathbb{R}^d, n \ge 1$, we denote by $C^n(\Omega_1; \Omega_2)$ the space of continuous functions $f : \Omega_1 \to \Omega_2$ that are n times differentiable and whose partial derivatives of order n are continuous. If $\Omega_2 \subset \mathbb{R}$, we simply write $C^n(\Omega_1)$. For any $f(x) \in C^2(\mathbb{R}^d)$, let $\nabla_x f$ and $\dot{f}$ denote its gradient and let $\nabla_x^2 f, \nabla_x \cdot f$, and $\Delta_x f$ denote its Hessian, divergence, and Laplacian, respectively. The function composition operation is marked as $g \circ f := g(f(x))$ for functions f and g.

Measure. The Borel σ-algebra of $\mathbb{R}^d$ is denoted by $\mathcal{B}(\mathbb{R}^d)$. The space of probability measures defined on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is denoted as $\mathcal{P}(\mathbb{R}^d)$. For any $\mathbb{R}^d$-valued random variable X, we use $\mathbb{E}[\mathsf{X}]$ and $\mathrm{Cov}(\mathsf{X})$ to denote its expectation and covariance matrix, respectively. We use $\mu * \nu$ to denote the convolution for any two probability measures μ and ν. For a random variable X, let $\mathrm{Law}(\mathsf{X})$ denote its probability distribution. For two random variables X and Y, let $\mathsf{X} \overset{d}{=} \mathsf{Y}$ mean that X and Y have the same distribution. Let $g : \mathbb{R}^k \to \mathbb{R}^d$ be a measurable map and μ be a probability measure on $\mathbb{R}^k$. The push-forward measure $f_\# \mu$ of a measurable set A is defined as $f_\# \mu := \mu(f^{-1}(A))$.

Let $N(m, \Sigma)$ denote a d-dimensional Gaussian random variable with mean vector $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. For simplicity, let $\gamma_{d,\sigma^2} := N(0, \sigma^2 \mathrm{I}_d)$, and let $\varphi_{m,\sigma^2}(x)$ denote the probability density function of $N(m, \sigma^2 \mathrm{I}_d)$ with respect to the Lebesgue measure. If $m = 0, \sigma = 1$, we abbreviate these as γ_d and $\varphi(x)$.

Lebesgue space. Let $L^p(\mathbb{R}^d; \mathbb{R}^\ell, \mu)$ denote the L^p space with the L^p norm for $p \in [1, \infty]$ w.r.t. a measure μ. To simplify the notation, we write $L^p(\mathbb{R}^d, \mu)$ if $\ell = 1$, $L^p(\mathbb{R}^d; \mathbb{R}^\ell)$ if the Lebesgue measure is used, and $L^p(\mathbb{R}^d)$ if both hold.

Chapter 2

Gaussian Interpolation Flows

Gaussian denoising has emerged as a powerful method for constructing simulation-free continuous normalizing flows for generative modeling. Despite their empirical successes, theoretical properties of these flows and the regularizing effect of Gaussian denoising have remained largely unexplored. In this chapter, we aim to address this gap by investigating the well-posedness of simulation-free continuous normalizing flows built on Gaussian denoising. Through a unified framework termed Gaussian interpolation flow, we establish the Lipschitz regularity of the flow velocity field, the existence and uniqueness of the flow, and the Lipschitz continuity of the flow map and the time-reversed flow map for several rich classes of target distributions. This analysis also sheds light on the auto-encoding and cycle consistency properties of Gaussian interpolation flows. Additionally, we study the stability of these flows in source distributions and perturbations of the velocity field, using the quadratic Wasserstein distance as a metric. Our findings offer valuable insights into the learning techniques employed in Gaussian interpolation flows for generative modeling, providing a solid theoretical foundation for end-to-end error analyses of learning Gaussian interpolation flows with empirical observations.

2.1 Main results

The main focus of this chapter is to study and establish the theoretical properties of Gaussian interpolation flow and its corresponding flow map. We show that the regularity conditions of the Cauchy-Lipschitz theorem are satisfied for several rich classes of probability distributions using variance inequalities. Based on the obtained regularity results, we further expose the well-posedness of GIFs, the Lipschitz continuity of flow mappings, and applications to generative modeling. The well-posedness results are crucial for studying the approximation and convergence properties of GIFs learned with

the flow or score matching method. When applied to generative modeling, our results further elucidate the auto-encoding and cycle consistency properties exhibited by GIFs.

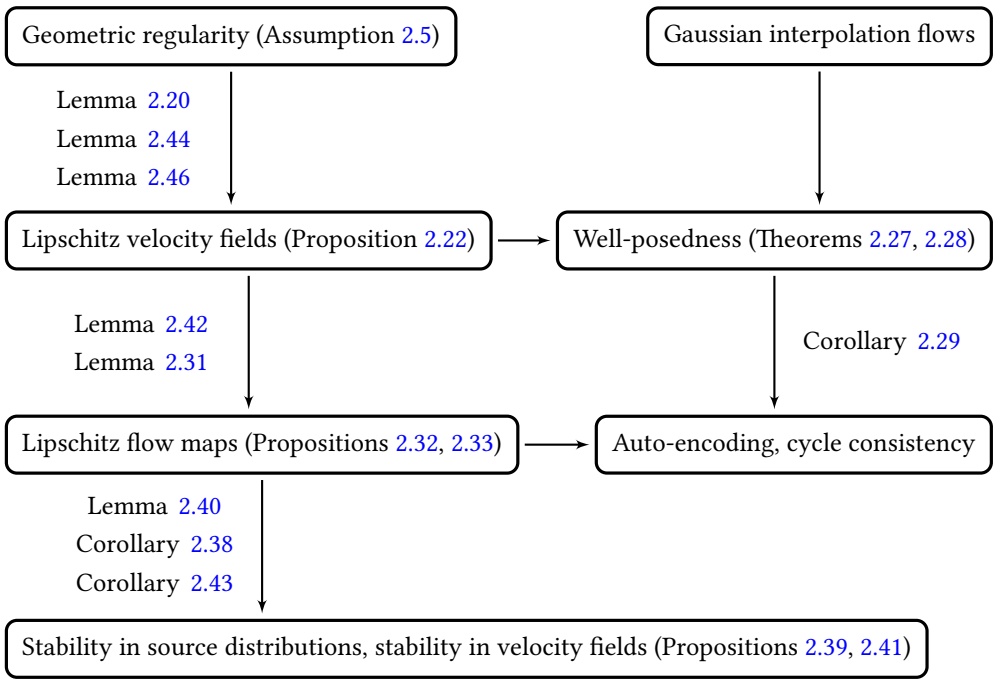

Figure 2.1. Roadmap of the main results.

We provide an overview of the main results in Figure 2.1, in which we indicate the assumptions used in our analysis and the relationship between the results. We also summarize our main contributions below.

- In Section 2.3, we extend the framework of stochastic interpolation proposed in Albergo and Vanden-Eijnden [2023]. Various ODE flows can be considered special cases of the extended framework. We prove that the marginal distributions of GIFs satisfy the continuity equation converging to the target distribution in the weak sense. Several explicit formulas of the velocity field and its derivatives are derived, which can facilitate computation and regularity estimation.

- In Sections 2.4 and 2.5, we establish the spatial Lipschitz regularity of the velocity field for a range of target measures with rich structures, which is sufficient to guarantee the well-posedness of GIFs. Additionally, we deduce the Lipschitz regularity of both the flow map and its time-reversed counterpart. The well-posedness

of GIFs is an essential attribute, serving as a foundational requirement for investigating numerical solutions of GIFs. It is important to note that while the flow maps are demonstrated to be Lipschitz continuous transport maps for generative modeling, the Lipschitz regularity for optimal transport maps has only been partially established to date.

- In Section 2.6, we show that the auto-encoding and cycle consistency properties of GIFs are inherently satisfied when the flow maps exhibit Lipschitz continuity with respect to the spatial variable. This demonstrates that exact auto-encoding and cycle consistency are intrinsic characteristics of GIFs. Our findings lend theoretical support to the findings made by Su et al. [2023], as illustrated in Figures 2.3 and 2.4.

- In Section 2.6, we conduct the stability analysis of GIFs, examining how they respond to changes in source distributions and to perturbations in the velocity field. This analysis, conducted in terms of the quadratic Wasserstein distance, provides valuable insights that justify the use of learning techniques such as Gaussian initialization and flow or score matching.

2.2 Preliminaries

In this section, we include several preliminary setups to show basic assumptions and several useful variance inequalities.

2.2.1 Assumptions

We focus on the probability distributions satisfying several types of assumptions of weak convexity, which offer a geometric notion of regularity in the study of high-dimensional distributions [Klartag, 2010]. The index of these regularity conditions would not explicitly depend on the dimension. On the one hand, weak-convexity regularity conditions are useful in deriving dimension-free guarantees for generative modeling and sampling from high-dimensional distributions. On the other hand, they accommodate distributions with complex shapes, including those with multiple modes.

Definition 2.1 (Cattiaux and Guillin, 2014)**.** A probability measure $\mu(\mathrm{d}x) = \exp(-U)\mathrm{d}x$ is κ-semi-log-concave for some $\kappa \in \mathbb{R}$ if its support $\Omega \subseteq \mathbb{R}^d$ is convex and its potential function $U \in C^2(\Omega)$ satisfies

$$\nabla_x^2 U(x) \geq \kappa \mathrm{I}_d, \quad \forall x \in \Omega.$$

The κ-semi-log-concavity condition is a relaxed notion of log-concavity, since here $\kappa < 0$ is allowed. When $\kappa \geq 0$, we are considering a log-concave probability measure that is proved to be unimodal [Saumard and Wellner, 2014]. However, when $\kappa < 0$, a κ-semi-log-concave probability measure can be multimodal.

Definition 2.2 (Eldan and Lee, 2018)**.** A probability measure $\mu(\mathrm{d}x) = \exp(-U)\mathrm{d}x$ is β-semi-log-convex for some $\beta > 0$ if its support $\Omega \subseteq \mathbb{R}^d$ is convex and its potential function $U \in C^2(\Omega)$ satisfies

$$\nabla_x^2 U(x) \leq \beta \mathrm{I}_d, \quad \forall x \in \Omega.$$

The following definition of L-log-Lipschitz continuity is a variant of L-Lipschitz continuity. It characterizes a first-order condition on the target function rather than a second-order condition such as κ-semi-log-concavity and β-semi-log-convexity in Definitions 2.1 and 2.2.

Definition 2.3. A function $f : \mathbb{R}^d \to \mathbb{R}_+$ is L-log-Lipschitz continuous if its logarithm is L-Lipschitz continuous for some $L \geq 0$.

Based on the definitions, we present two assumptions on the target distribution. Assumption 2.4 concerns the absolute continuity and the moment condition. Assumption 2.5 imposes geometric regularity conditions.

Assumption 2.4. The probability measure ν is absolutely continuous with respect to the Lebesgue measure and has a finite second moment.

Assumption 2.5. Let $D := (1/\sqrt{2})\mathrm{diam}(\mathrm{supp}(\nu))$. The probability measure ν satisfies one or more of the following conditions:

(i) ν is β-semi-log-convex for some $\beta > 0$ and κ-semi-log-concave for some $\kappa > 0$ with $\mathrm{supp}(\nu) = \mathbb{R}^d$;

(ii) ν is κ-semi-log-concave for some $\kappa \in \mathbb{R}$ with $D \in (0, \infty)$;

(iii) $\nu = \gamma_{d,\sigma^2} * \rho$ where ρ is a probability measure supported on a Euclidean ball of radius R on $\mathbb{R}^d$;

(iv) ν is β-semi-log-convex for some $\beta > 0$, κ-semi-log-concave for some $\kappa \leq 0$, and $\frac{\mathrm{d}\nu}{\mathrm{d}\gamma_d}(x)$ is L-log-Lipschitz in x for some $L \geq 0$ with $\mathrm{supp}(\nu) = \mathbb{R}^d$.

Multimodal distributions. Assumption 2.5 enumerates scenarios where probability distributions are endowed with geometric regularity. We examine the scenarios and clarify whether they cover multimodal distributions. Scenario (i) is referred to as the classical strong log-concavity case ($\kappa > 0$), and thus, describes unimodal distributions. Scenario (ii) allows $\kappa \leq 0$ and requires that the support is bounded. Mixtures of Gaussian distributions are considered in Scenario (iii), and typically are multimodal distributions. Scenario (iv) also allows $\kappa \leq 0$ when considering a log-Lipschitz perturbation of the standard Gaussian distribution. Both Scenario (ii) and Scenario (iv) incorporate multimodal distributions due to the potential negative lower bound κ.

Lipschitz score. Lipschitz continuity of the score function is a basic regularity assumption on target distributions in the study of sampling algorithms based on Langevin and Hamiltonian dynamics. Even for high-dimensional distributions, this assumption endows a great source of regularity. For an L-Lipschitz score function, its corresponding distribution is both L-semi-log-convex and $(-L)$-semi-log-concave for some $L \geq 0$.

2.2.2 Variance inequalities

Variance inequalities like the Brascamp-Lieb inequality and the Cramér-Rao inequality are fundamental inequalities for explaining the regularizing effect of Gaussian denoising. Combined with κ-semi-log-concavity and β-semi-log-convexity, these inequalities are crucial for deducing the Lipschitz regularity of the velocity fields of GIFs in Proposition 2.22-(b) and (c).

Lemma 2.6 (Brascamp-Lieb inequality)**.** *Let $\mu(\mathrm{d}x) = \exp(-U(x))\mathrm{d}x$ be a probability measure on a convex set $\Omega \subseteq \mathbb{R}^d$ whose potential function $U : \Omega \to \mathbb{R}$ is of class C^2 and strictly convex. Then for every locally Lipschitz function $f \in L^2(\Omega, \mu)$,*

$$\mathrm{Var}_\mu(f) \leq \mathbb{E}_\mu\Big[\langle \nabla_x f, (\nabla_x^2 U)^{-1} \nabla_x f \rangle\Big]. \tag{2.1}$$

When applied to functions of the form $f : x \mapsto \langle x, e\rangle$ for any $e \in \mathbb{S}^{d-1}$, the Brascamp-Lieb inequality yields an upper bound of the covariance matrix

$$\mathrm{Cov}_\mu(\mathsf{X}) \preceq \mathbb{E}_\mu\left[(\nabla_x^2 U(x))^{-1}\right] \tag{2.2}$$

with equality if $\mathsf{X} \sim N(m, \Sigma)$ with Σ positive definite.

Under the strong log-concavity condition, that is, μ is κ-semi-log-concave with $\kappa > 0$, and if the Euclidean Bakry-Émery criterion is satisfied [Bakry and Émery, 1985], the Brascamp-Lieb inequality instantly recovers the Poincaré inequality (see Definition 2.50).

The Brascamp-Lieb inequality originally appeared in [Brascamp and Lieb, 1976, Theorem 4.1]. Alternative proofs are provided in Bobkov and Ledoux [2000], Bakry et al. [2014], Cordero-Erausquin [2017]. The dimension-free inequality (2.1) can be further strengthened to obtain several variants with dimensional improvement.

Lemma 2.7 (Cramér-Rao inequality)**.** *Let $\mu(\mathrm{d}x) = \exp(-U(x))\mathrm{d}x$ be a probability measure on $\mathbb{R}^d$ whose potential function $U : \mathbb{R}^d \to \mathbb{R}$ is of class C^2. Then for every $f \in C^1(\mathbb{R}^d)$,*

$$\mathrm{Var}_\mu(f) \geq \langle \mathbb{E}_\mu[\nabla_x f], \left(\mathbb{E}_\mu[\nabla_x^2 U]\right)^{-1} \mathbb{E}_\mu[\nabla_x f]\rangle. \tag{2.3}$$

When applied to functions of the form $f : x \mapsto \langle x, e\rangle$ for any $e \in \mathbb{S}^{d-1}$, the Cramér-Rao inequality yields a lower bound of the covariance matrix

$$\mathrm{Cov}_\mu(\mathsf{X}) \geq \left(\mathbb{E}_\mu[\nabla_x^2 U(x)]\right)^{-1} \tag{2.4}$$

with equality as well if $\mathsf{X} \sim N(m, \Sigma)$ with Σ positive definite.

The Cramér-Rao inequality plays a central role in asymptotic statistics as well as in information theory. The inequality (2.4) has an alternative derivation from the Cramér-Rao bound for the location parameter. For detailed proofs of the Cramér-Rao inequality, readers are referred to Chewi and Pooladian [2022], Dai et al. [2023], and the references therein.

2.3 Gaussian interpolation flows

Simulation-free CNFs represent a potent class of generative models based on ODE flows. Albergo and Vanden-Eijnden [2023] and Albergo et al. [2023b] introduce an innovative

CNF that is constructed using stochastic interpolation techniques, such as Gaussian denoising. They conduct a thorough investigation of this flow, particularly examining its applications and effectiveness in generative modeling.

We study the ODE flow and its associated flow map as defined by the Gaussian denoising process. This process has been explored from various perspectives, including diffusion models and stochastic interpolants. Building upon the work of Albergo and Vanden-Eijnden [2023] and Albergo et al. [2023b], we expand the stochastic interpolant framework by relaxing certain conditions on the functions a_t and b_t, offering a more comprehensive perspective on the Gaussian denoising process.

In our generalization, we introduce an adaptive starting point to the stochastic interpolation framework, which allows for greater flexibility in the modeling process. By examining this modified framework, we aim to demonstrate that the Gaussian denoising principle is effectively implemented within the context of stochastic interpolation.

Definition 2.8 (Vector interpolation)**.** Let $z \in \mathbb{R}^d$, $x_1 \in \mathbb{R}^d$ be two vectors in the Euclidean space and let $x_0 := a_0 z + b_0 x_1$ with $a_0 > 0, b_0 \geq 0$. Then we construct an interpolant between x_0 and x_1 over time $t \in [0,1]$ through $I_t(x_0, x_1)$, defined by

$$I_t(x_0, x_1) = a_t z + b_t x_1, \tag{2.5}$$

where a_t, b_t satisfy

$$\begin{aligned} &\dot{a}_t \leq 0, \quad \dot{b}_t \geq 0, \quad a_0 > 0, \quad b_0 \geq 0, \quad a_1 = 0, \quad b_1 = 1, \\ &a_t > 0 \ \text{ for any } t \in (0,1), \quad b_t > 0 \ \text{ for any } t \in (0,1), \\ &a_t, b_t \in C^2([0,1)), \quad a_t^2 \in C^1([0,1]), \quad b_t \in C^1([0,1]). \end{aligned} \tag{2.6}$$

Remark 2.9. Compared with the vector interpolant defined by Albergo and Vanden-Eijnden [2023] (a.k.a. one-sided interpolant in Albergo et al. [2023b]), we extend its definition by relaxing the requirements that $a_0 = 1, b_0 = 0$ with $a_0 > 0, b_0 \geq 0$. This consideration is largely motivated by analyzing the probability flow ODEs of the variance-exploding (VE) SDE and the variance-preserving (VP) SDE [Song et al., 2021b]. We illustrate examples of interpolants incorporated by Definition 2.8 in Table 2.1. In this table, we consider the VE interpolant [Song et al., 2021b], VP interpolant [Song et al., 2021b], linear interpolant [Liu et al., 2023], Föllmer interpolant [Dai et al., 2023], and trigonometric interpolant [Albergo and Vanden-Eijnden, 2023]. There are two types of

source measures including a standard Gaussian distribution γ_d and a convoluted distribution consisting of the target distribution and γ_d.

Remark 2.10. We have eased the smoothness conditions for the functions a_t and b_t required in Albergo and Vanden-Eijnden [2023]. Specifically, we consider the case where $a_t, b_t \in C^2([0,1))$, $a_t^2 \in C^1([0,1])$, and $b_t \in C^1([0,1])$. This relaxation enables us to include the Föllmer flow into our framework, characterized by $a_t = \sqrt{1-t^2}$ and $b_t = t$. It is evident that $a_t = \sqrt{1-t^2}$ does not fulfill the condition $a_t \in C^2([0,1])$, but it does meet the requirements $a_t \in C^2([0,1))$ and $a_t^2 \in C^1([0,1])$.

Remark 2.11. The C^2 regularity of a_t, b_t is necessary to derive the regularity of the velocity field $v(t,x)$ in Eq. (2.9) concerning the time variable t. In addition, the C^1 regularity of a_t^2, b_t is sufficient to ensure the Lipschitz regularity of the velocity field $v(t,x)$ in Eq. (2.9) concerning the space variable x.

A natural generalization of the vector interpolant (2.5) is to construct a set interpolant between two convex sets through Minkowski sum, which is common in convex geometry. A set interpolant stimulates the construction of a measure interpolant between a structured source measure and a target measure.

As noted, we can construct a measure interpolation using a Gaussian convolution path. The measure interpolation is particularly relevant to Gaussian denoising and Gaussian channels in information theory as elucidated in Remark 2.18. Because of this connection with Gaussian denoising, we call the measure interpolation a Gaussian stochastic interpolation. The Gaussian stochastic interpolation can be understood as a collection of linear combinations of a standard Gaussian random variable and the target random variable. The coefficients of the linear combinations vary with time $t \in [0,1]$ as shown in Definition 2.8. Later in this section, we will show this Gaussian stochastic interpolation can be transformed into a deterministic ODE flow.

Gaussian stochastic interpolation has been investigated from several perspectives in the literature. The rectified flow has been proposed in Liu et al. [2023], and its theoretical connection with optimal transport has been investigated in Liu [2022]. The formulation of the rectified flow is to learn the ODE flow defined by stochastic interpolation with linear time coefficients. In Appendix C of Liu et al. [2023], there is a nonlinear extension of the rectified flow in which the linear coefficients are replaced by

general nonlinear coefficients. Albergo et al. [2023b] extends the stochastic interpolant framework proposed in [Albergo and Vanden-Eijnden, 2023] by considering a linear combination among three random variables. In Section 3 of Albergo et al. [2023b], the original stochastic interpolant framework is recovered as a one-sided interpolant between the Gaussian distribution and the target distribution. Moreover, Lipman et al. [2023] propose a flow matching method which directly learns a Gaussian conditional probability path with a neural ODE. In Section 4.1 of Lipman et al. [2023], the velocity fields of the variance exploding and variance preserving probability flows are shown as special instances of the flow matching framework. We summarize these formulations as Gaussian stochastic interpolation by slightly extending the original stochastic interpolant framework.

Type	VE	VP	Linear	Föllmer	Trigonometric
a_t	α_t	α_t	$1-t$	$\sqrt{1-t^2}$	$\cos(\frac{\pi}{2}t)$
b_t	1	$\sqrt{1-\alpha_t^2}$	t	t	$\sin(\frac{\pi}{2}t)$
a_0	α_0	α_0	1	1	1
b_0	1	$\sqrt{1-\alpha_0^2}$	0	0	0
Source	Convolution	Convolution	γ_d	γ_d	γ_d

Table 2.1. Summary of various measure interpolants.

Definition 2.12 (Measure interpolation)**.** Let $\mu = \mathrm{Law}(\mathsf{X}_0)$ and $\nu = \mathrm{Law}(\mathsf{X}_1)$ be two probability measures satisfying $\mathsf{X}_0 = a_0\mathsf{Z} + b_0\mathsf{X}_1$ where $\mathsf{Z} \sim \gamma_d := N(0, \mathrm{I}_d)$ is independent from X_1. We call $(\mathsf{X}_t)_{t\in[0,1]}$ a Gaussian stochastic interpolation from the source measure μ to the target measure ν, which is defined through I_t over time interval $[0,1]$ as follows

$$\mathsf{X}_t = I_t(\mathsf{X}_0, \mathsf{X}_1), \quad \mathsf{X}_0 = a_0\mathsf{Z} + b_0\mathsf{X}_1, \quad \mathsf{Z} \sim \gamma_d, \quad \mathsf{X}_1 \sim \nu. \tag{2.7}$$

Remark 2.13. It is obvious that the marginal distribution of X_t satisfies $\mathsf{X}_t \overset{d}{=} a_t\mathsf{Z} + b_t\mathsf{X}_1$ with $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$.

Motivated by the time-varying properties of the Gaussian stochastic interpolation, we derive that its marginal flow satisfies the continuity equation. This result characterizes the dynamics of the marginal density flow of the Gaussian stochastic interpolation.

Theorem 2.14. *Suppose that Assumption 2.4 holds. Then the marginal flow $(p_t)_{t\in[0,1]}$ of the Gaussian stochastic interpolation $(\mathsf{X}_t)_{t\in[0,1]}$ between μ and ν satisfies the continuity equation*

$$\partial_t p_t + \nabla_x \cdot (p_t v(t,x)) = 0, \quad (t,x) \in [0,1]\times\mathbb{R}^d, \quad p_0(x) = \frac{\mathrm{d}\mu}{\mathrm{d}x}(x), \quad p_1(x) = \frac{\mathrm{d}\nu}{\mathrm{d}x}(x) \tag{2.8}$$

in the weak sense with the velocity field

$$v(t,x) := \mathbb{E}[\dot{a}_t \mathsf{Z} + \dot{b}_t \mathsf{X}_1 | \mathsf{X}_t = x], \quad t \in (0,1), \tag{2.9}$$

$$v(0,x) := \lim_{t\downarrow 0} v(t,x), \quad v(1,x) := \lim_{t\uparrow 1} v(t,x). \tag{2.10}$$

Remark 2.15. We notice that $x = a_t\mathbb{E}[\mathsf{Z}|\mathsf{X}_t = x] + b_t\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x]$ due to Eq. (2.7). Then it holds that

$$v(t,x) = \tfrac{\dot{a}_t}{a_t}x + \left(\dot{b}_t - \tfrac{\dot{a}_t}{a_t}b_t\right)\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x], \quad t \in (0,1). \tag{2.11}$$

We also notice that, according to Tweedie's formula (cf. Lemma 2.51), it holds that

$$s(t,x) = \tfrac{b_t}{a_t^2}\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x] - \tfrac{1}{a_t^2}x, \quad t \in (0,1), \tag{2.12}$$

where $s(t,x)$ is the score function of the marginal distribution of $\mathsf{X}_t \sim p_t$.

Combining (2.11) and (2.12), it follows that the velocity field is a gradient field and its nonlinear term is the score function $s(t,x)$, namely, for any $t \in (0,1)$,

$$v(t,x) = \tfrac{\dot{b}_t}{b_t}x + \left(\tfrac{\dot{b}_t}{b_t}a_t^2 - \dot{a}_t a_t\right)s(t,x). \tag{2.13}$$

Remark 2.16. A relevant result has been provided in the proof of [Albergo and Vanden-Eijnden, 2023, Proposition 4] in a restricted case that $a_0 = 1, b_0 = 0$. In this case, if $\dot{a}_0, \dot{a}_1, \dot{b}_0, \dot{b}_1$ are well-defined, the velocity field reads

$$v(0,x) = \dot{a}_0 x + \dot{b}_0\mathbb{E}_\nu[\mathsf{X}_1], \quad v(1,x) = \dot{b}_1 x + \dot{a}_1\mathbb{E}_{\gamma_d}[\mathsf{Z}]$$

at time 0 and 1. Otherwise, if any one of $\dot{a}_0, \dot{a}_1, \dot{b}_0, \dot{b}_1$ is not well-defined, the velocity field $v(0,x)$ or $v(1,x)$ should be considered on a case-by-case basis. In addition, we provide an alternative viewpoint of the relationship between the velocity field associated with stochastic interpolation and the score function of its marginal flow using Tweedie's formula in Lemma 2.51.

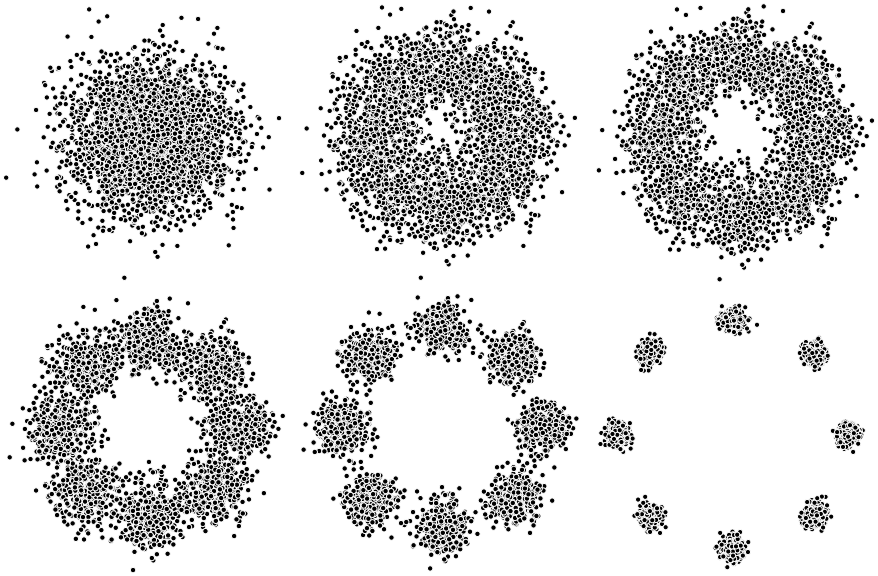

Figure 2.2. Snapshots of a Gaussian interpolation flow based on the Föllmer interpolant.

Remark 2.17 (Diffusion process)**.** The marginal flow of the Gaussian stochastic interpolation (2.7) coincides with the time-reversed marginal flow of a diffusion process $(\overline{X}_t)_{t\in[0,1)}$ [Albergo et al., 2023b, Theorem 3.5] defined by

$$\mathrm{d}\overline{X}_t = -\frac{\dot{b}_{1-t}}{b_{1-t}}\overline{X}_t + \sqrt{2\left(\frac{\dot{b}_{1-t}}{b_{1-t}}a_{1-t}^2 - \dot{a}_{1-t}a_{1-t}\right)}\mathrm{d}\overline{W}_t.$$

Remark 2.18 (Gaussian denoising)**.** The Gaussian stochastic interpolation has an information-theoretic interpretation as a time-varying Gaussian channel. Here a_t^2 and b_t^2/a_t^2 stand for the noise level and signal-to-noise ratio (SNR) for time $t \in [0,1]$, respectively. As time $t \to 1$, we are approaching the high-SNR regime, that is, the SNR b_t^2/a_t^2 grows to ∞. Moreover, the SNR b_t^2/a_t^2 is monotonically increasing in time t over $[0,1]$. The Gaussian noise level gets reduced through this Gaussian denoising process.

We are now ready to define Gaussian interpolation flows by representing the continuity equation (2.8) with Lagrangian coordinates [Ambrosio and Crippa, 2014]. A basic observation is that GIFs share the same marginal density flow with Gaussian stochastic interpolations. The continuity equation (2.8) plays a central role in the derandomization procedure from Gaussian stochastic interpolations to GIFs. We additionally illustrate GIFs using a two-dimensional example as in Figure 2.2. The source distribution is the standard two-dimensional Gaussian distribution γ_2, and the target distribution is a mix-

ture of six two-dimensional Gaussian distributions as the shape of a circle. The image panels are placed sequentially from time $t = 0$ to time $t = 1$.

Definition 2.19 (Gaussian interpolation flow)**.** Suppose that probability measure ν satisfies Assumption 2.4. If $(X_t)_{t\in[0,1]}$ solves the initial value problem (IVP)

$$\frac{\mathrm{d}X_t}{\mathrm{d}t}(x) = v(t, X_t(x)), \quad X_0(x) \sim \mu, \quad t \in [0,1], \tag{2.14}$$

where μ is defined in Definition 2.12 and the velocity field v is given by Eq. (2.9) and (2.10), we call $(X_t)_{t\in[0,1]}$ a Gaussian interpolation flow associated with the target measure ν.

2.4 Spatial Lipschitz estimates for the velocity field

We have explicated the idea of Gaussian denoising with the procedure of Gaussian stochastic interpolation or a Gaussian channel with increasing SNR w.r.t. time. By interpreting the process as an ODE flow, we derive the framework of Gaussian interpolation flows. First and foremost, an intuition is that the regularizing effect of Gaussian denoising would ensure the Lipschitz smoothness of the velocity field. Since the standard Gaussian distribution is both 1-semi-log-concave and 1-semi-log-convex, its convolution with a target distribution will maintain its high regularity as long as the target distribution satisfies the regularity conditions. We rigorously justify this intuition by establishing spatial Lipschitz estimates for the velocity field. These estimates are established based on the upper bounds and lower bounds regarding the Jacobian matrix of the velocity field $v(t,x)$ according to the Cauchy-Lipschitz theorem, which are given in Proposition 2.22 below. To deal with the Jacobian matrix $\nabla_x v(t,x)$, we introduce a covariance expression of it and present the associated upper bounds and lower bounds.

The velocity field $v(t,x)$ is decomposed into a linear term and a nonlinear term, the score function $s(t,x)$. To analyze the Jacobian $\nabla_x v(t,x)$, we only need to focus on $\nabla_x s(t,x)$, that is, $\nabla_x^2 \log p_t(x)$. To ease the notation, we would henceforth use Y for X_1. Correspondingly, we replace $p_1(x)$ with $p_1(y)$ for the density function of Y.

According to Bayes' rule, the marginal density p_t of X_t satisfies

$$p_t(x) = \int p(t,x|y)p_1(y)\mathrm{d}y$$

where $\mathsf{Y} \sim p_1(y)$ and $p(t,x|y) = \varphi_{b_t y, a_t^2}(x)$ is a conditional distribution induced by the Gaussian noise. Due to the factorization $p_t(x)p(y|t,x) = p(t,x|y)p_1(y)$, the score function $s(t,x)$ and its derivative $\nabla_x s(t,x)$ have the following expressions

$$s(t,x) = -\nabla_x \log p(y|t,x) - \frac{x - b_t y}{a_t^2}, \quad \nabla_x s(t,x) = -\nabla_x^2 \log p(y|t,x) - \frac{1}{a_t^2}\mathrm{I}_d.$$

Thanks to the expressions above, a covariance matrix expression of $\nabla_x s(t,x)$ is endowed by the exponential family property of $p(y|t,x)$.

Lemma 2.20. *The conditional distribution $p(y|t,x)$ is an exponential family distribution and a covariance matrix expression of the log-Hessian matrix $\nabla_x^2 \log p(y|t,x)$ for any $t \in (0,1)$ is given by*

$$\nabla_x^2 \log p(y|t,x) = -\frac{b_t^2}{a_t^4}\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x), \tag{2.15}$$

where $\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x)$ is the covariance matrix of $\mathsf{Y}|\mathsf{X}_t = x \sim p(y|t,x)$. Moreover, for any $t \in (0,1)$, it holds that

$$\nabla_x s(t,x) = \frac{b_t^2}{a_t^4}\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x) - \frac{1}{a_t^2}\mathrm{I}_d, \tag{2.16}$$

and that

$$\nabla_x v(t,x) = \frac{b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x) + \frac{\dot{a}_t}{a_t}\mathrm{I}_d. \tag{2.17}$$

Remark 2.21. Since $\partial_t\left(\frac{b_t^2}{a_t^2}\right) = \frac{2b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)$, it follows from (2.17) that the derivative of the SNR with respect to time t controls the dependence of $\nabla_x v(t,x)$ on $\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x)$.

The representation (2.17) can be used to upper bound and lower bound $\nabla_x v(t,x)$. This technique has been widely used to deduce the regularity of the score function concerning the space variable [Mikulincer and Shenfeld, 2024, 2023, Chen et al., 2023d, Lee et al., 2023, Chen et al., 2023a]. The covariance matrix expression (2.16) of the score function has a close connection with the Hatsell-Nolte identity in information theory [Hatsell and Nolte, 1971, Palomar and Verdú, 2005, Wu and Verdú, 2011, Cai and Wu, 2014, Wibisono et al., 2017, Wibisono and Jog, 2018a,b, Dytso et al., 2023a,b].

Employing the covariance expression in Lemma 2.20, we establish several bounds on $\nabla_x v(t,x)$ in the following proposition.

Proposition 2.22. *Let $\nu(\mathrm{d}y) = p_1(y)\mathrm{d}y$ be a probability measure on $\mathbb{R}^d$ with $D := (1/\sqrt{2})\mathrm{diam}(\mathrm{supp}(\nu))$.*

(a) For any $t \in (0,1)$,

$$\frac{\dot{a}_t}{a_t}\mathrm{I}_d \preceq \nabla_x v(t,x) \preceq \left\{\frac{b_t(a_t\dot{b}_t - \dot{a}_t b_t)}{a_t^3}D^2 + \frac{\dot{a}_t}{a_t}\right\}\mathrm{I}_d.$$

(b) Suppose that p_1 is β-semi-log-convex with $\beta > 0$ and $\mathrm{supp}(p_1) = \mathbb{R}^d$. Then for any $t \in (0,1]$,

$$\nabla_x v(t,x) \succeq \frac{\beta a_t\dot{a}_t + b_t\dot{b}_t}{\beta a_t^2 + b_t^2}\mathrm{I}_d.$$

(c) Suppose that p_1 is κ-semi-log-concave with $\kappa \in \mathbb{R}$. Then for any $t \in (t_0, 1]$,

$$\nabla_x v(t,x) \preceq \frac{\kappa a_t\dot{a}_t + b_t\dot{b}_t}{\kappa a_t^2 + b_t^2}\mathrm{I}_d,$$

where t_0 is the root of the equation $\kappa + \frac{b_t^2}{a_t^2} = 0$ over $t \in (0,1)$ if $\kappa < 0$ and $t_0 = 0$ if $\kappa \geq 0$.

(d) Fix a probability measure ρ on $\mathbb{R}^d$ supported on a Euclidean ball of radius R, and let $\nu := \gamma_{d,\sigma^2} * \rho$ with $\sigma > 0$. Then for any $t \in (0,1)$,

$$\frac{\dot{a}_t a_t + \sigma^2\dot{b}_t b_t}{a_t^2 + \sigma^2 b_t^2}\mathrm{I}_d \preceq \nabla_x v(t,x) \preceq \left\{\frac{a_t b_t(a_t\dot{b}_t - \dot{a}_t b_t)}{(a_t^2 + \sigma^2 b_t^2)^2}R^2 + \frac{\dot{a}_t a_t + \sigma^2\dot{b}_t b_t}{a_t^2 + \sigma^2 b_t^2}\right\}\mathrm{I}_d.$$

(e) Suppose that $\frac{\mathrm{d}\nu}{\mathrm{d}\gamma_d}(x)$ is L-log-Lipschitz for some $L \geq 0$. Then for any $t \in (0,1)$,

$$\left\{\left(\frac{\dot{b}_t}{b_t}a_t^2 - \dot{a}_t a_t\right)\left(-B_t - L^2\left(\frac{b_t}{a_t^2+b_t^2}\right)^2\right) + \frac{\dot{a}_t a_t + \dot{b}_t b_t}{a_t^2+b_t^2}\right\}\mathrm{I}_d$$
$$\preceq \nabla_x v(t,x) \preceq \left\{\left(\frac{\dot{b}_t}{b_t}a_t^2 - \dot{a}_t a_t\right)B_t + \frac{\dot{a}_t a_t + \dot{b}_t b_t}{a_t^2+b_t^2}\right\}\mathrm{I}_d,$$

where $B_t := 5Lb_t(a_t^2 + b_t^2)^{-\frac{3}{2}}(L + (\log(\sqrt{a_t^2 + b_t^2}/b_t))^{-\frac{1}{2}})$.

Comparing part (a) with part (d) in Proposition 2.22, we can see that the bounds in (a) are consistent with those in (d) in the sense that (a) is a limiting case of part (d) as $\sigma \to 0$.

The lower bound in part (a) blows up at time $t = 1$ owing to $a_1 = 0$, while in part (d) it behaves well since the lower bound in part (d) coincides with a lower bound indicated by the $\frac{1}{\sigma^2}$-semi-log-convex property. It reveals that the regularity of the velocity field $v(t,x)$ with respect to the space variable x improves when the target random variable is bounded and is subject to Gaussian perturbation.

The lower bound in part (b) and the upper bound in part (c) are tight in the sense that both of them are attainable for a Gaussian target distribution, that is,

$$\nabla_x v(t,x) = \frac{\beta a_t \dot{a}_t + b_t \dot{b}_t}{\beta a_t^2 + b_t^2} \mathrm{I}_d \quad \text{if } \nu = \gamma_{d,1/\beta}.$$

The upper and lower bounds in Proposition 2.22-(a) and (e) become vacuous as they both blow up at time $t = 1$. The intuition behind is that the Jacobian matrix of the velocity field can be both lower and upper bounded at time $t = 1$ only if the score function of the target measure is Lipschitz continuous in the space variable x. Under an additional Lipschitz score assumption (equivalently, β-semi-log-convex and κ-semi-log-concave for some $\beta = -\kappa \geq 0$), the upper and lower bounds in part (a) and part (e) can be strengthened at time $t = 1$ based on the lower bound in (b) and the upper bound in part (c).

According to Proposition 2.22-(a) and (c), there are two upper bounds available that shall be compared with each other. One is the D^2-based bound in part (a), and the other is the κ-based bound in part (c). According to the proof of Proposition 2.22, these two upper bounds are equal if and only if the corresponding upper bounds on $\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x)$ are equal, that is,

$$D^2 = \left(\kappa + \frac{b_t^2}{a_t^2}\right)^{-1}. \tag{2.18}$$

Then the critical case is $\kappa D^2 = 1$ since simplifying Eq. (2.18) reveals that

$$D^{-2} - \kappa = \frac{b_t^2}{a_t^2}. \tag{2.19}$$

We note that b_t^2/a_t^2, ranging over $(0,\infty)$, is monotonically increasing w.r.t. $t \in (0,1)$. Suppose that $\kappa D^2 > 1$. Then (2.19) has no root over $t \in (0,1)$, which implies that the

κ-based bound is tighter over $[0,1)$, i.e.,

$$D^2 > \left(\kappa + \frac{b_t^2}{a_t^2}\right)^{-1}, \quad \forall t \in [0,1).$$

Otherwise, suppose that $\kappa D^2 < 1$. Then (2.19) has a root $t_1 \in (0,1)$, which implies that the D^2-based bound is tighter over $[0,t_1)$, i.e.,

$$D^2 < \left(\kappa + \frac{b_t^2}{a_t^2}\right)^{-1}, \quad \forall t \in [0,t_1),$$

and that the κ-based bound is tighter over $[t_1,1)$, i.e.,

$$D^2 \geq \left(\kappa + \frac{b_t^2}{a_t^2}\right)^{-1}, \quad \forall t \in [t_1,1).$$

Next, we present several upper bounds on the maximum eigenvalue of the Jacobian matrix of the velocity field $\lambda_{\max}(\nabla_x v(t,x))$ and its exponential estimates for studying the Lipschitz regularity of the flow maps as noted in Lemma 2.31.

Corollary 2.23. *Let ν be a probability measure on $\mathbb{R}^d$ with $D := (1/\sqrt{2})\mathrm{diam}(\mathrm{supp}(\nu))$ and suppose that ν is κ-semi-log-concave with $\kappa \geq 0$.*

(a) If $\kappa D^2 \geq 1$, then

$$\lambda_{\max}(\nabla_x v(t,x)) \leq \theta_t := \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2}, \ t \in [0,1]. \tag{2.20}$$

(b) If $\kappa D^2 < 1$, then

$$\lambda_{\max}(\nabla_x v(t,x)) \leq \theta_t := \begin{cases} \frac{b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)D^2 + \frac{\dot{a}_t}{a_t}, & t \in [0,t_1), \\ \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2}, & t \in [t_1,1], \end{cases} \tag{2.21}$$

where t_1 solves (2.19).

Corollary 2.24. *Let ν be a probability measure on $\mathbb{R}^d$ with $D := (1/\sqrt{2})\mathrm{diam}(\mathrm{supp}(\nu)) < \infty$ and suppose that ν is κ-semi-log-concave with $\kappa < 0$. Then*

$$\lambda_{\max}(\nabla_x v(t,x)) \leq \theta_t := \begin{cases} \frac{b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)D^2 + \frac{\dot{a}_t}{a_t}, & t \in [0,t_1), \\ \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2}, & t \in [t_1,1], \end{cases} \tag{2.22}$$

where t_1 solves (2.19).

Corollary 2.25. *Fix a probability measure ρ on $\mathbb{R}^d$ supported on a Euclidean ball of radius R and let $\nu := \gamma_{d,\sigma^2} * \rho$ with $\sigma > 0$. Then*

$$\lambda_{\max}(\nabla_x v(t,x)) \le \theta_t := \frac{\dot{a}_t a_t + \sigma^2 \dot{b}_t b_t}{a_t^2 + \sigma^2 b_t^2} + \frac{a_t b_t (a_t \dot{b}_t - \dot{a}_t b_t)}{(a_t^2 + \sigma^2 b_t^2)^2} R^2. \tag{2.23}$$

Corollary 2.26. *Suppose that ν is κ-semi-log-concave for some $\kappa \le 0$, and $\frac{\mathrm{d}\nu}{\mathrm{d}\gamma_d}(x)$ is L-log-Lipschitz for some $L \ge 0$. Then*

$$\lambda_{\max}(\nabla_x v(t,x)) \le \theta_t := \begin{cases} \left(\frac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right) B_t + \frac{\dot{a}_t a_t + \dot{b}_t b_t}{a_t^2 + b_t^2}, & t \in [0, t_2), \\ \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2}, & t \in [t_2, 1], \end{cases} \tag{2.24}$$

where $B_t := 5L b_t (a_t^2 + b_t^2)^{-\frac{3}{2}} (L + (\log(\sqrt{a_t^2 + b_t^2}/b_t))^{-\frac{1}{2}})$ and $t_2 \in (t_0, 1)$.

2.5 Well-posedness and Lipschtiz flow maps

In this section, we study the well-posedness of GIFs and the Lipschitz properties of their flow maps. We also show that the marginal distributions of GIFs satisfy the log-Sobolev inequality and the Poincaré inequality if Assumptions 2.4 and 2.5 are satisfied.

Theorem 2.27 (Well-posedness)**.** *Suppose Assumptions 2.4 and 2.5-(i), (iii), or (iv) are satisfied. Then there exists a unique solution $(X_t)_{t\in[0,1]}$ to the IVP (2.14). Moreover, the push-forward measure satisfies $X_{t\#}\mu = \mathrm{Law}(a_t \mathsf{Z} + b_t \mathsf{X}_1)$ with $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$.*

Theorem 2.28. *Suppose Assumptions 2.4 and 2.5-(ii) are satisfied. For any $\underline{t} \in (0,1)$, there exists a unique solution $(X_t)_{t\in[0,1-\underline{t}]}$ to the IVP (2.14). Moreover, the push-forward measure satisfies $X_{t\#}\mu = \mathrm{Law}(a_t \mathsf{Z} + b_t \mathsf{X}_1)$ with $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$.*

Corollary 2.29 (Time-reversed flow)**.** *Suppose Assumptions 2.4 and 2.5-(i), (iii), or (iv) are satisfied. Then the time-reversed flow $(X_t^*)_{t\in[0,1]}$ associated with ν is a unique solution to the IVP:*

$$\frac{\mathrm{d}X_t^*}{\mathrm{d}t}(x) = -v(1-t, X_t^*(x)), \quad X_0^*(x) \sim \nu, \quad t \in [0,1]. \tag{2.25}$$

The push-forward measure satisfies $X_{t\#}^*\nu = \mathrm{Law}(a_{1-t}\mathsf{Z} + b_{1-t}\mathsf{X}_1)$ where $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$. Moreover, the flow map satisfies $X_t^*(x) = X_t^{-1}(x)$.

Corollary 2.30. *Suppose Assumptions 2.4 and 2.5-(ii) are satisfied. For any $\underline{t} \in (0,1)$, the time-reversed flow $(X_t^*)_{t\in[\underline{t},1]}$ associated with ν is a unique solution to the IVP:*

$$\frac{\mathrm{d}X_t^*}{\mathrm{d}t}(x) = -v(1-t, X_t^*(x)), \quad X_{\underline{t}}^*(x) \sim \mathrm{Law}(a_{1-\underline{t}}\mathsf{Z} + b_{1-\underline{t}}\mathsf{X}_1), \quad t \in [\underline{t}, 1], \tag{2.26}$$

where $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$. The push-forward measure satisfies $X_{t\#}^*\nu = \mathrm{Law}(a_{1-t}\mathsf{Z} + b_{1-t}\mathsf{X}_1)$. Moreover, the flow map satisfies $X_t^*(x) = X_t^{-1}(x)$.

Based on the well-posedness of the flow, we can provide an upper bound on the Lipschitz constant of the induced flow map.

Lemma 2.31. *Suppose that a flow $(X_t)_{t\in[0,1]}$ is well-posed with a velocity field $v(t,x) : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ of class C^1 in x, and that for any $(t,x) \in [0,1]\times\mathbb{R}^d$, it holds $\nabla_x v(t,x) \preceq \theta_t \mathrm{I}_d$. Let the flow map $X_{s,t} : \mathbb{R}^d \to \mathbb{R}^d$ be of class C^1 in x for any $0 \le s \le t \le 1$. Then the flow map $X_{s,t}$ is Lipschitz continuous with an upper bound of its Lipschitz constant given by*

$$\|\nabla_x X_{s,t}(x)\|_{2,2} \le \exp\left(\int_s^t \theta_u \mathrm{d}u\right). \tag{2.27}$$

Using Lemma 2.31, we show that the flow map of a GIF is Lipschitz continuous in the space variable x.

Proposition 2.32 (Lipschitz mappings)**.** *Suppose that Assumptions 2.4 and 2.5-(i) hold.*

(i) If ν is κ-semi-log-concave for some $\kappa > 0$, then the flow map $X_1(x)$ is a Lipschitz mapping, that is,

$$\|\nabla_x X_1(x)\|_{2,2} \le \frac{1}{\sqrt{\kappa a_0^2 + b_0^2}}, \quad \forall x \in \mathbb{R}^d.$$

In particular, if $a_0 = 1$ and $b_0 = 0$, then

$$\|\nabla_x X_1(x)\|_{2,2} \le \frac{1}{\sqrt{\kappa}}, \quad \forall x \in \mathbb{R}^d.$$

(ii) If ν is β-semi-log-convex for some $\beta > 0$, then the time-reversed flow map $X_1^*(x)$ is a Lipschitz mapping, that is,

$$\|\nabla_x X_1^*(x)\|_{2,2} \le \sqrt{\beta a_0^2 + b_0^2}, \quad \forall x \in \mathrm{supp}(\nu).$$

In particular, if $a_0 = 1$ *and* $b_0 = 0$, *then*

$$\|\nabla_x X_1^*(x)\|_{2,2} \le \sqrt{\beta}, \quad \forall x \in \mathrm{supp}(\nu).$$

Proposition 2.33 (Gaussian mixtures)**.** *Suppose that Assumptions 2.4 and 2.5-(iii) hold. Then the flow map* $X_1(x)$ *is a Lipschitz mapping, that is,*

$$\|\nabla_x X_1(x)\|_{2,2} \le \frac{\sigma}{\sqrt{a_0^2 + \sigma^2 b_0^2}} \exp\left(\frac{a_0^2}{a_0^2 + \sigma^2 b_0^2} \cdot \frac{R^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}^d.$$

In particular, if $a_0 = 1$ *and* $b_0 = 0$, *then*

$$\|\nabla_x X_1(x)\|_{2,2} \le \sigma \exp\left(\frac{R^2}{2\sigma^2}\right), \quad \forall x \in \mathbb{R}^d.$$

Moreover, the time-reversed flow map $X_1^*(x)$ *is a Lipschitz mapping, that is,*

$$\|\nabla_x X_1^*(x)\|_{2,2} \le \sqrt{\sigma^{-2} a_0^2 + b_0^2}, \quad \forall x \in \mathrm{supp}(\nu).$$

In particular, if $a_0 = 1$ *and* $b_0 = 0$, *then*

$$\|\nabla_x X_1^*(x)\|_{2,2} \le \frac{1}{\sigma}, \quad \forall x \in \mathrm{supp}(\nu).$$

Remark 2.34. Well-posed GIFs produce diffeomorphisms that transport the source measure onto the target measure. The diffeomorphism property of the transport maps are relevant to the auto-encoding and cycle consistency properties of their generative modeling applications. We defer a detailed discussion to Section 2.6.

Early stopping implicitly mollifies the target measure with a small Gaussian noise. For image generation tasks (with bounded pixel values), the mollified target measure is indeed a Gaussian mixture distribution considered in Theorem 2.33. The regularity of the target measure largely gets enhanced through such mollification, especially when the target measure is supported on a low-dimensional manifold in accordance with the data manifold hypothesis. Therefore, although such a diffeomorphism $X_1(x)$ may not be well-defined for general bounded target measures, an off-the-shelf solution would be to perturb the target measure with a small Gaussian noise or to employ the early stopping technique. Both approaches will smooth the landscape of the target measure.

Proposition 2.35. *Suppose the target measure ν satisfies the log-Sobolev inequality with constant $C_{\mathrm{LS}}(\nu)$. Then the marginal distribution of the GIF $(p_t)_{t\in[0,1]}$ satisfies the log-Sobolev inequality, and its log-Sobolev constant $C_{\mathrm{LS}}(p_t)$ is bounded as*

$$C_{\mathrm{LS}}(p_t) \le a_t^2 + b_t^2 C_{\mathrm{LS}}(\nu).$$

Moreover, suppose the target measure ν satisfies the Poincaré inequality with constant $C_P(\nu)$. Then the marginal distribution of the GIF $(p_t)_{t\in[0,1]}$ satisfies the Poincaré inequality, and its Poincaré constant $C_{\mathrm{P}}(p_t)$ is bounded as

$$C_{\mathrm{P}}(p_t) \le a_t^2 + b_t^2 C_{\mathrm{P}}(\nu).$$

The log-Sobolev and Poincaré inequalities (see Definitions 2.49 and 2.50) are fundamental tools for establishing convergence guarantees for Langevin Monte Carlo algorithms. From an algorithmic viewpoint, the predictor-corrector algorithm in score-based diffusion models and the corresponding probability flow ODEs essentially combine the ODE numerical solver (performing as the predictor) and the overdamped Langevin diffusion (performing as the corrector) to simulate samples from the marginal distributions [Song et al., 2021b]. Proposition 2.35 shows that the marginal distributions all satisfy the log-Sobolev and Poincaré inequalities under mild assumptions on the target distribution. This conclusion suggests that Langevin Monte Carlo algorithms are certified to have convergence guarantees for sampling from the marginal distributions of GIFs. Furthermore, the target distributions covered in Assumption 2.5 are shown to satisfy the log-Sobolev and Poincaré inequalities [Mikulincer and Shenfeld, 2024, Dai et al., 2023, Fathi et al., 2023], which suggests that the assumptions of Proposition 2.35 generally hold.

2.6 Applications to generative modeling

Auto-encoding is a primary principle in learning a latent representation with generative models [Goodfellow et al., 2016, Chapter 14]. Meanwhile, the concept of cycle consistency is important to unpaired image-to-image translation between the source and target domains [Zhu et al., 2017]. The recent work by Su et al. [2023] propose the dual diffusion implicit bridges (DDIB) for image-to-image translation, which shows a strong pattern of exact auto-encoding and image-to-image translation. DDIBs are built

upon the denoising diffusion implicit models (DDIM), which share the same probability flow ODE with VESDE (considered as VE interpolant in Table 2.1), as pointed out by [Song et al., 2021a, Proposition 1]. First, DDIBs attain latent embeddings of source images encoded with one DDIM operating in the source domain. The encoding embeddings are then decoded using another DDIM trained in the target domain to construct target images. The whole process consisting of two DDIMs seems to be cycle consistent up to numerical errors. Several phenomena of auto-encoding and cycle consistency are observed in the unpaired data generation procedure with DDIBs.

We replicate the two-dimensional experiments by Su et al. [2023] in Figures 2.3 and 2.4 to show the phenomena of approximate auto-encoding and cycle consistency of GIFs[1].

In Figure 2.3, the Concentric Rings data in the source domain (the first panel) is encoded into the latent domain (the second panel), and then decoded into the source domain (the third panel). According to the consistent color pattern and pointwise correspondences across the domains, both the learned encoder mapping and the learned decoder mapping exhibit approximate Lipschitz continuity with respect to the space variable. One justification of such auto-encoding observation is presented in Corollary 2.36 where we prove that the composition of the encoder map and the decoder map yields an identity map.

In Figure 2.4, the cycle consistency property is manifested through the consistency of color patterns across the transformations. We transform the Moons data in the source domain onto the Concentric Squares data in the target domain, and then complete the cycle by mapping the target data back to the source domain. The latent spaces play a central role in the bidirectional translation. We provide a proof in Corollary 2.37 accounting for the cycle consistency property.

To elucidate the empirical auto-encoding and cycle consistency for measure transport, we derive Corollaries 2.36 and 2.37 below and analyze the transport maps defined by GIFs (covering the probability flow ODE of VESDE used by DDIBs). We consider the continuous-time framework and the population level, which precludes learning errors including the time discretization errors and velocity field estimation errors, and show that the transport maps naturally possess the exact auto-encoding and cycle consistency

[1]The implementation is based on the GitHub repository at https://github.com/suxuann/ddib.

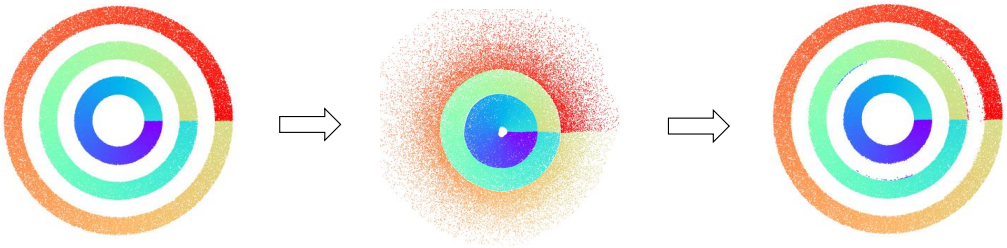

Figure 2.3. An illustration of auto-encoding using DDIBs.

properties at the population level.

Corollary 2.36 (Auto-encoding). *Suppose Assumptions 2.4 and 2.5-(i), (iii), or (iv) hold for a target measure ν. The Gaussian interpolation flow $(X_t)_{t\in[0,1]}$ and its time-reversed flow $(X_t^*)_{t\in[0,1]}$ form an auto-encoder with a Lipschitz encoder $X_1^*(x)$ and a Lipschitz decoder $X_1(x)$. The auto-encoding property holds in the sense that*

$$X_1 \circ X_1^* = \mathrm{I}_d. \tag{2.28}$$

Corollary 2.37 (Cycle consistency). *Suppose Assumptions 2.4 and 2.5-(i), (iii), or (iv) hold for the target measures ν_1 and ν_2. For the target measure ν_1, we define the Gaussian interpolation flow $(X_{1,t})_{t\in[0,1]}$ and its time-reversed flow $(X_{1,t}^*)_{t\in[0,1]}$. We also define the Gaussian interpolation flow $(X_{2,t})_{t\in[0,1]}$ and its time-reversed flow $(X_{2,t}^*)_{t\in[0,1]}$ for the target measure ν_2 using the same a_t and b_t. Then the transport maps $X_{1,1}(x)$, $X_{1,1}^*(x)$, $X_{2,1}(x)$, and $X_{2,1}^*(x)$ are Lipschitz continuous in the space variable x. Furthermore, the cycle consistency property holds in the sense that*

$$X_{1,1} \circ X_{2,1}^* \circ X_{2,1} \circ X_{1,1}^* = \mathrm{I}_d. \tag{2.29}$$

Corollaries 2.36 and 2.37 show that the auto-encoding and cycle consistency properties hold for the flows at the population level. These results provide insights to the approximate auto-encoding and cycle consistency properties at the sample level.

There are several types of errors introduced in the training of GIFs. On the one hand, the approximation in specifying source measures would exert influence on modeling the distribution. On the other hand, the approximation in the velocity field also affects the distribution learning error. We use the stability analysis method in the differential equations theory to address the potential effects of these errors.

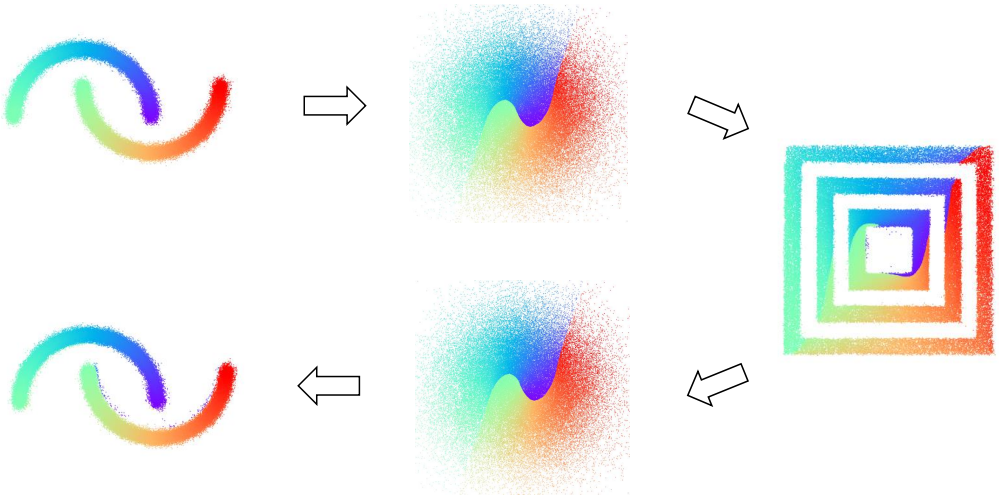

Figure 2.4. An illustration of cycle consistency using DDIBs.

Corollary 2.38. *Suppose Assumptions 2.4 and 2.5-(i), (iii), or (iv) hold. It holds that*

$$C_1 := \sup_{x\in\mathbb{R}^d} \|\nabla_x X_1(x)\|_{2,2} < \infty, \quad C_2 := \sup_{(t,x)\in[0,1]\times\mathbb{R}^d} \|\nabla_x v(t,x)\|_{2,2} < \infty.$$

Proposition 2.39 (Stability in the source distribution)**.** *Suppose Assumptions 2.4 and 2.5-(i), (iii), or (iv) hold. If the source measure $\mu = \mathrm{Law}(a_0 Z + b_0 X_1)$ is replaced with the Gaussian measure γ_{d,a_0^2}, then the stability of the transport map X_1 is guaranteed by the W_2 distance between the push-forward measure $X_{1\#}\gamma_{d,a_0^2}$ and the target measure $\nu = \mathrm{Law}(X_1)$ as follows*

$$W_2(X_{1\#}\gamma_{d,a_0^2}, \nu) \le C_1 b_0 \sqrt{\mathbb{E}_\nu[\|X_1\|^2]} \exp(C_2 d). \tag{2.30}$$

The stability analysis in Proposition 2.39 provides insights into the selection of source measures for learning probability flow ODEs and GIFs. The error bound (2.30) demonstrates that when the signal intensity is reasonably small in the source measure, that is, $b_0 \ll 1$, the distribution estimation error, induced by the approximation with a Gaussian source measure, is small as well in the sense of the quadratic Wasserstein distance. Using a Gaussian source measure to replace the true convolution source measure is a common approximation method for learning probability flow ODEs and GIFs. Our analysis shows this replacement is reasonable for the purpose of distribution estimation.

The Alekseev-Gröbner formula and its stochastic variants [Del Moral and Singh, 2022] have been shown effective in quantifying the stability of well-posed ODE and

SDE flows against perturbations of its velocity field or drift [Bortoli, 2022, Benton et al., 2023]. We state these results below for convenience.

Lemma 2.40. *[Hairer et al., 1993, Theorem 14.5] Let $(X_t)_{t\in[0,1]}$ and $(Y_t)_{t\in[0,1]}$ solve the following IVPs, respectively*

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = v(t, X_t), \quad X_0 = x_0, \quad t \in [0,1],$$
$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = \tilde{v}(t, Y_t), \quad Y_0 = x_0, \quad t \in [0,1],$$

where $v(t,x) : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ and $\tilde{v}(t,x) : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ are the velocity fields.

(i) Suppose that v is of class C^1 in x. Then the Alekseev-Gröbner formula for the difference $X_t(x_0) - Y_t(x_0)$ is given by

$$X_t(x_0) - Y_t(x_0) = \int_0^t (\nabla_x X_{s,t})(Y_s(x_0))^\top (v(s, Y_s(x_0)) - \tilde{v}(s, Y_s(x_0)))\,\mathrm{d}s \tag{2.31}$$

where $\nabla_x X_{s,t}(x)$ satisfies the variational equation

$$\partial_t(\nabla_x X_{s,t}(x)) = (\nabla_x v)(t, X_{s,t}(x))\nabla_x X_{s,t}(x), \quad \nabla_x X_{s,s}(x) = \mathrm{I}_d. \tag{2.32}$$

(ii) Suppose that $\tilde{v}$ is of class C^1 in x. Then the Alekseev-Gröbner formula for the difference $Y_t(x_0) - X_t(x_0)$ is given by

$$Y_t(x_0) - X_t(x_0) = \int_0^t (\nabla_x Y_{s,t})(X_s(x_0))^\top (\tilde{v}(s, X_s(x_0)) - v(s, X_s(x_0)))\,\mathrm{d}s \tag{2.33}$$

where $\nabla_x Y_{s,t}(x)$ satisfies the variational equation

$$\partial_t(\nabla_x Y_{s,t}(x)) = (\nabla_x \tilde{v})(t, Y_{s,t}(x))\nabla_x Y_{s,t}(x), \quad \nabla_x Y_{s,s}(x) = \mathrm{I}_d. \tag{2.34}$$

Exploiting the Alekseev-Gröbner formulas in Lemma 2.40 and uniform Lipschitz properties of the velocity field, we deduce two error bounds in terms of the quadratic Wasserstein (W_2) distance to show the stability of the ODE flow when the velocity field is not accurate.

Proposition 2.41 (Stability in the velocity field)**.** *Suppose Assumptions 2.4 and 2.5 hold. Let $\tilde{q}_t$ denote the density function of $Y_{t\#}\mu$.*

(i) Suppose that

$$\int_0^1 \int_{\mathbb{R}^d} \|v(t,x) - \tilde{v}(t,x)\|^2 \tilde{q}_t(x) \mathrm{d}x\mathrm{d}t \leq \varepsilon. \tag{2.35}$$

Then

$$W_2^2(Y_{1\#}\mu, \nu) \leq \varepsilon \int_0^1 \exp\left(2\int_s^1 \theta_u \mathrm{d}u\right)\mathrm{d}s. \tag{2.36}$$

(ii) Suppose that

$$\sup_{(t,x)\in[0,1]\times\mathbb{R}^d} \|\nabla_x \tilde{v}(t,x)\|_{2,2} \leq C_3.$$

Then

$$W_2^2(Y_{1\#}\mu, \nu) \leq \frac{\exp(2C_3) - 1}{2C_3} \int_0^1 \int_{\mathbb{R}^d} \|v(t,x) - \tilde{v}(t,x)\|^2 p_t(x) \mathrm{d}x\mathrm{d}t. \tag{2.37}$$

Proposition 2.41 provides a stability analysis against the estimation error of the velocity field using the W_2 distance. The estimation error originates from the flow matching or score matching procedures and the approximation error rising from using deep neural networks in estimating the velocity field or the score function. These two W_2 bounds imply that the distribution estimation error is controlled by the L_2 estimation error of flow matching and score matching. Indeed, this point justifies the soundness of the approximation method through flow matching and score matching. The first W_2 bound (2.36) relies on the L_2 control (2.35) of the perturbation error of the velocity field. The second W_2 bound (2.37) is slightly better than that provided in [Albergo and Vanden-Eijnden, 2023, Proposition 3] but still has exponential dependence on the Lipschitz constant of $\tilde{v}(t,x)$.

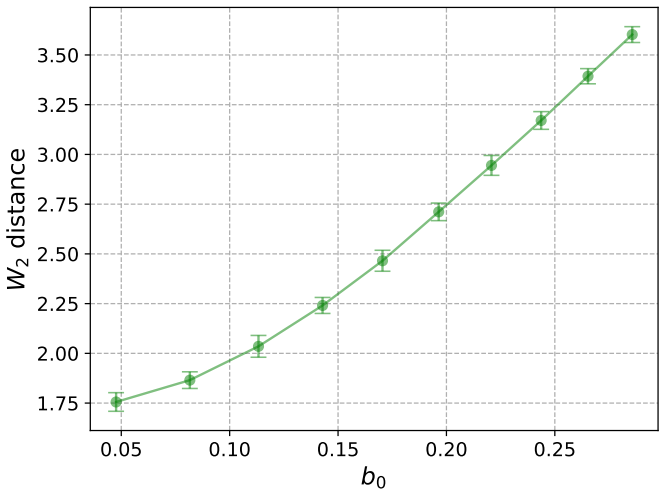

Figure 2.5. An approximately linear relation between b_0 and the Wasserstein-2 distance.

To demonstrate the bounds presented in Propositions 2.39 and 2.41, we conducted further experiments with a mixture of eight two-dimensional Gaussian distributions. These propositions provide bounds for the stability of the flow when subjected to perturbations in either the source distribution or the velocity field. Let the target distribution be the following two-dimensional Gaussian mixture

$$p(x) = \sum_{j=1}^{8} \phi(x; \mu_j, \Sigma_j),$$

where $\phi(x; \mu_j, \Sigma_j)$ is the probability density function for the Gaussian distribution with mean $\mu_j = 12(\sin(2(j-1)\pi/8), \cos(2(j-1)\pi/8))^\top$ and covariance matrix $\Sigma_j = 0.03^2 \mathrm{I}_2$ for $j = 1, \cdots, 8$. For Gaussian mixtures, the velocity field has an explicit formula, which facilitates the perturbation analysis.

To illustrate the bound in Proposition 2.39, we consider a perturbation of the source distribution for the following model:

$$\mathsf{X}_t = a_t \mathsf{Z} + b_t \mathsf{X} \quad \text{with} \quad a_t = 1 - \frac{t+\zeta}{1+\zeta}, \quad b_t = \frac{t+\zeta}{1+\zeta},$$

where $\zeta \in [0, 0.3]$ is a value controlling the perturbation level. It is easy to see $a_0 = 1/(1+\zeta), b_0 = \zeta/(1+\zeta)$. Thus, the source distribution $\mathrm{Law}(a_0\mathsf{Z} + b_0\mathsf{X})$ is a mixture of Gaussian distributions. Practically, we can use a Gaussian distribution γ_{2,a_0^2} to replace

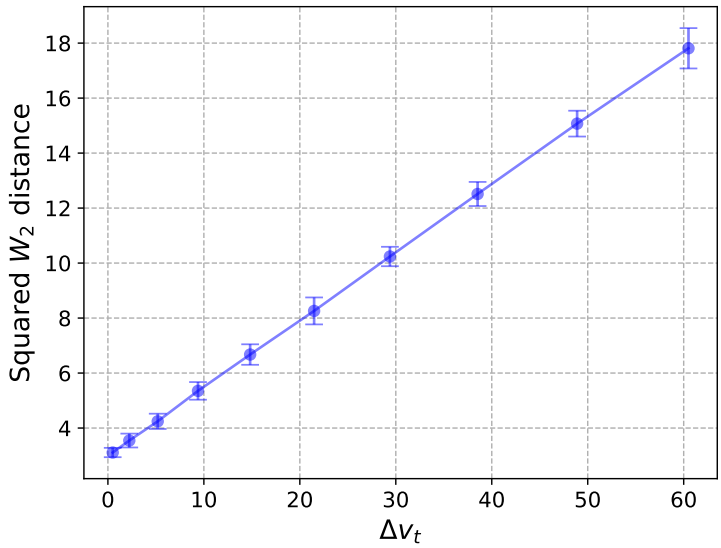

Figure 2.6. A linear relation between Δv_t and the squared Wasserstein-2 distance.

this source distribution. In Proposition 2.39, we bound the error between the distributions of generated samples due to the replacement, that is,

$$W_2(X_{1\#}\gamma_{d,a_0^2}, \nu) \leq C b_0,$$

where C is a constant. We illustrate this theoretical bound using the mixture of Gaussian distributions and the Gaussian interpolation flow given above. We consider a mesh for the variable ζ and plot the curve for b_0 and $W_2(X_{1\#}\gamma_{d,a_0^2}, \nu)$ in Figure 2.5. Through Figure 2.5, an approximate linear relation between b_0 and $W_2(X_{1\#}\gamma_{d,a_0^2}, \nu)$ is observed, which supports the results of Proposition 2.39.

We now consider perturbing the velocity field v_t by adding random noise. Let $\epsilon \in [0.5, 5.5]$. The random noise is generated using a Bernoulli random variable supported on $\{-\epsilon, \epsilon\}$. Let $\tilde{v}_t$ denote the perturbed velocity field. Then we can compute

$$\Delta v_t := \|v_t - \tilde{v}_t\|^2 = 2\epsilon^2.$$

We use the velocity field v_t and the perturbed velocity field $_t$ to generate samples and compute the squared Wasserstein-2 distance between the sample distributions. According to Proposition 2.41, the squared Wasserstein-2 distance should be linearly upper bounded as $\mathcal{O}(\Delta v_t)$, that is,

$$W_2^2(Y_{1\#}\mu, \nu) \leq \tilde{C} \int_0^1 \int_{\mathbb{R}^2} \epsilon^2 p_t(x) \mathrm{d}x \mathrm{d}t = \tilde{C}\epsilon^2,$$

where $\tilde{C}$ is a constant. This theoretical insight is illustrated in Figure 2.6, where a linear relationship between these two variables is observed.

2.7 Related work

GIFs and the induced transport maps are related to CNFs and score-based diffusion models. Mathematically, they interrelate with the literature on Lipschitz mass transport and Wasserstein gradient flows. A central question in developing the ODE flow or transport map method for generative modeling is how to construct an ODE flow or transport map that are sufficiently smooth and enable efficient computation. Various approaches have been proposed to answer the question.

CNFs construct invertible mappings between an isotropic Gaussian distribution and a complex target distribution [Chen et al., 2018, Grathwohl et al., 2019]. They fall within the broader framework of neural ODEs [Chen et al., 2018, Ruiz-Balet and Zuazua, 2023]. A major challenge for CNFs is designing a time-dependent ODE flow whose marginal distribution converges to the target distribution while allowing for efficient estimation of its velocity field. Previous work has explored several principles to construct such flows, including optimal transport, Wasserstein gradient flows, and diffusion processes. Additionally, Gaussian denoising has emerged as an effective principle for constructing simulation-free CNFs in generative modeling.

Liu et al. [2023] propose the rectified flow, which is based on a linear interpolation between a standard Gaussian distribution and the target distribution, mimicking the Gaussian denoising procedure. Albergo and Vanden-Eijnden [2023] study a similar formulation called stochastic interpolation, defining a trigonometric interpolant between a standard Gaussian distribution and the target distribution. Albergo et al. [2023b] extend this idea by proposing a stochastic bridge interpolant between two arbitrary distributions. Under a few regularity assumptions, the velocity field of the ODE flow modeling the stochastic bridge interpolant is proven to be continuous in the time variable and smooth in the space variable.

Lipman et al. [2023] introduce a nonlinear least squares method called flow matching to directly estimate the velocity field of probability flow ODEs. All of these models are encompassed within the framework of simulation-free CNFs, which have been the

focus of numerous ongoing research efforts [Neklyudov et al., 2023, Tong et al., 2023, Chen and Lipman, 2023, Albergo et al., 2023b, Shaul et al., 2023, Pooladian et al., 2023, Albergo et al., 2023a,c]. Furthermore, Marzouk et al. [2023] provide the first statistical convergence rate for the simulation-based method by placing neural ODEs within the nonparametric estimation framework.

Score-based diffusion models integrate the time reversal of stochastic differential equations with the score matching technique [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020, Song and Ermon, 2020, Song et al., 2021b,a, De Bortoli et al., 2021]. These models are capable of modeling highly complex probability distributions and have achieved state-of-the-art performance in image synthesis tasks [Dhariwal and Nichol, 2021, Rombach et al., 2022]. The probability flow ODEs of diffusion models can be considered as CNFs, whose velocity field incorporates the nonlinear score function [Song et al., 2021b, Karras et al., 2022, Lu et al., 2022b,a, Zheng et al., 2023]. In addition to the score matching method, Lu et al. [2022a] and Zheng et al. [2023] explore maximum likelihood estimation for probability flow ODEs. However, the regularity of these probability flow ODEs has not been studied and their well-posedness properties remain to be established.

A key concept in defining measure transport is Lipschitz mass transport, where the transport maps are required to be Lipschitz continuous. This ensures the smoothness and stability of the measure transport. There is a substantial body of research on the Lipschitz properties of transport maps. The celebrated Caffarelli's contraction theorem [Caffarelli, 2000, Theorem 2] establishes the Lipschitz continuity of optimal transport maps that push the standard Gaussian measure onto a log-concave measure. Colombo et al. [2017] study a Lipschitz transport map between perturbations of log-concave measures using optimal transport theory.

Mikulincer and Shenfeld [2024] demonstrate that the Brownian transport map, defined by the Föllmer process, is Lipschitz continuous when it pushes forward the Wiener measure on the Wiener space to the target measure on the Euclidean space. Additionally, Neeman [2022] and Mikulincer and Shenfeld [2023] prove that the transport map along the reverse heat flow of certain target measures is Lipschitz continuous.

Beyond studying Lipschitz transport maps, significant effort has been devoted to applying optimal transport theory in generative modeling. Zhang et al. [2018] propose the

Monge-Ampe‘re flow for generative modeling by solving the linearized Monge-Ampère equation. Optimal transport theory has been utilized as a general principle to regularize the training of continuous normalizing flows or generators for generative modeling [Finlay et al., 2020, Yang and Karniadakis, 2020, Onken et al., 2021, Makkuva et al., 2020]. Liang [2021] leverage the regularity theory of optimal transport to formalize the generator-discriminator-pair regularization of GANs under a minimax rate framework.

In our work, we study the Lipschitz transport maps defined by GIFs, which differ from the optimal transport map. GIFs naturally fit within the framework of continuous normalizing flows, and their flow mappings are examined from the perspective of Lipschitz mass transport.

Wasserstein gradient flows offer another principled approach to constructing ODE flows for generative modeling. A Wasserstein gradient flow is derived from the gradient descent minimization of a certain energy functional over probability measures endowed with the quadratic Wasserstein metric [Ambrosio et al., 2008]. The Eulerian formulation of Wasserstein gradient flows produces the continuity equations that govern the evolution of marginal distributions. After transferred into a Lagrangian formulation, Wasserstein gradient flows define ODE flows that have been widely explored for generative modeling [Johnson and Zhang, 2018, Gao et al., 2019, Liutkus et al., 2019, Johnson and Zhang, 2019, Arbel et al., 2019, Mroueh et al., 2019, Ansari et al., 2021, Mroueh and Nguyen, 2021, Fan et al., 2022, Gao et al., 2022, Duncan et al., 2023, Xu et al., 2022]. Wasserstein gradient flows are shown to be connected with the forward process of diffusion models. The variance preserving SDE of diffusion models is equivalent to the Langevin dynamics towards the standard Gaussian distribution that can be interpreted as a Wasserstein gradient flow of the Kullback–Leibler divergence for a standard Gaussian distribution [Song et al., 2021b]. In the meantime, the probability flow ODE of the variance preserving SDE conforms to the Eulerian formulation of this Wasserstein gradient flow. However, when assigning a general distribution instead of the standard Gaussian distribution, it remains unclear whether the ODE formulation of Wasserstein gradient flows possesses well-posedness.

The main contribution of this chapter lies in establishing the theoretical properties of GIFs and their associated flow maps in a unified way. Our theoretical results encompass the Lipschitz continuity of both the flow velocity field and the flow map, addressing the

existence, uniqueness, and stability of the flow. We also demonstrate that both the flow map and its inverse possess Lipschitz properties.

Our proposed framework for Gaussian interpolation flow builds upon previous research on probability flow methods in diffusion models [Song et al., 2021b,a] and stochastic interpolation methods for generative modeling [Liu et al., 2023, Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023]. Rather than adopting a methodological perspective, we focus on elucidating the theoretical aspects of these flows from a unified standpoint, thereby enhancing the understanding of various methodological approaches. Our theoretical results are derived from geometric considerations of the target distribution and from analytic calculations that exploit the Gaussian denoising property.

2.8 Conclusion

Gaussian denoising as a framework for constructing continuous normalizing flows holds great promise in generative modeling. Through a unified framework and rigorous analysis, we have established the well-posedness of these flows, shedding light on their capabilities and limitations. We have examined the Lipschitz regularity of the corresponding flow maps for several rich classes of probability measures. When applied to generative modeling based on Gaussian denoising, we have shown that GIFs possess auto-encoding and cycle consistency properties at the population level. Additionally, we have established stability error bounds for the errors accumulated during the process of learning GIFs. Although our analysis has partially established the well-posedness of the GIFs, it remains unclear whether the well-posedness holds for learning more general distributions. Moreover, it remains interesting to investigate the advantages of the denoising framework beyond Gaussian denoising.

2.9 Proofs and supplementary results

In this section, we provide proofs of the lemmas and theorems shown in the previous sections of the chapter.

2.9.1 Proofs of Theorem 2.14 and Lemma 2.20

Dynamical properties of Gaussian interpolation flow $(\mathsf{X}_t)_{t\in[0,1]}$ form the cornerstone of the measure interpolation method. Following Albergo and Vanden-Eijnden [2023], Albergo et al. [2023b], we leverage an argument of characteristic functions to quantify the dynamics of its marginal flow, and in result, to prove Theorem 2.14.

Proof of Theorem 2.14. Let $\omega \in \mathbb{R}^d$. For the Gaussian stochastic interpolation $(\mathsf{X}_t)_{t\in[0,1]}$, we define the characteristic function of X_t by

$$\Psi(t,\omega) := \mathbb{E}[\exp(i\langle\omega,\mathsf{X}_t\rangle)] = \mathbb{E}[\exp(i\langle\omega,a_t\mathsf{Z}+b_t\mathsf{X}_1\rangle)] = \mathbb{E}[\exp(ia_t\langle\omega,\mathsf{Z}\rangle)]\mathbb{E}[\exp(ib_t\langle\omega,\mathsf{X}_1\rangle)],$$

where the last equality is due to the independence of between $\mathsf{Z}\sim\gamma_d$ and $\mathsf{X}_1\sim\nu$. Taking the time derivative of $\Psi(t,\omega)$ for $t\in(0,1)$, we derive that

$$\partial_t\Psi(t,\omega) = i\langle\omega,\psi(t,\omega)\rangle$$

where

$$\psi(t,\omega) := \mathbb{E}[\exp(i\langle\omega,\mathsf{X}_t\rangle)(\dot{a}_t\mathsf{Z}+\dot{b}_t\mathsf{X}_1)].$$

We first define

$$v(t,\mathsf{X}_t) := \mathbb{E}[\dot{a}_t\mathsf{Z}+\dot{b}_t\mathsf{X}_1|\mathsf{X}_t]. \tag{2.38}$$

Using the double expectation formula, we deduce that

$$\psi(t,\omega) = \mathbb{E}[\exp(i\langle\omega,\mathsf{X}_t\rangle)\mathbb{E}[\dot{a}_t\mathsf{Z}+\dot{b}_t\mathsf{X}_1|\mathsf{X}_t]] = \mathbb{E}[\exp(i\langle\omega,\mathsf{X}_t\rangle)v(t,\mathsf{X}_t)].$$

Applying the inverse Fourier transform to $\psi(t,\omega)$, it holds that

$$j(t,x) := (2\pi)^{-d}\int_{\mathbb{R}^d}\exp(-i\langle\omega,x\rangle)\psi(t,\omega)\mathrm{d}\omega = p_t(x)v(t,x),$$

where $v(t,x) := \mathbb{E}[\dot{a}_t\mathsf{Z}+\dot{b}_t\mathsf{X}_1|\mathsf{X}_t=x]$. Then it further yields that

$$\partial_t p_t + \nabla_x\cdot j(t,x) = 0,$$

that is,

$$\partial_t p_t + \nabla_x\cdot(p_t v(t,x)) = 0.$$

Next, we study the property of $v(t,x)$ at $t=0$ and $t=1$. Notice that

$$x = a_t\mathbb{E}[\mathsf{Z}|\mathsf{X}_t=x] + b_t\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t=x]. \tag{2.39}$$

Combining Eq. (2.38) and (2.39), it implies that

$$v(t,x) = \frac{\dot{a}_t}{a_t}x + \left(\dot{b}_t - \frac{\dot{a}_t}{a_t}b_t\right)\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x], \quad t \in (0,1). \tag{2.40}$$

According to Tweedie's formula in Lemma 2.51, it holds that

$$s(t,x) = \frac{b_t}{a_t^2}\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x] - \frac{1}{a_t^2}x, \quad t \in (0,1), \tag{2.41}$$

where $s(t,x)$ is the score function of the marginal distribution of $\mathsf{X}_t \sim p_t$.

Combining Eq. (2.40), (2.41), it holds that the velocity field is a gradient field and its nonlinear term is the score function $s(t,x)$, namely, for any $t \in (0,1)$,

$$v(t,x) = \frac{\dot{b}_t}{b_t}x + \left(\frac{\dot{b}_t}{b_t}a_t^2 - \dot{a}_t a_t\right)s(t,x). \tag{2.42}$$

By the regularity properties that $a_t, b_t \in C^2([0,1)), a_t^2 \in C^1([0,1]), b_t \in C^1([0,1])$, we have that $\dot{a}_0, \dot{b}_0, \dot{a}_1 a_1$, and $\dot{b}_1$ are well-defined. Then by Eq. (2.40), we define that

$$v(0,x) := \lim_{t\downarrow 0} v(t,x) = \frac{\dot{a}_0}{a_0}x + \left(\dot{b}_0 - \frac{\dot{a}_0}{a_0}b_0\right)\mathbb{E}[\mathsf{X}_1|\mathsf{X}_0 = x]$$

Using Eq. (2.42) yields that

$$v(1,x) := \lim_{t\uparrow 1} v(t,x) = \frac{\dot{b}_1}{b_1}x - \dot{a}_1 a_1 s(1,x). \tag{2.43}$$

This completes the proof. □

Lemma 2.20 presents several standard properties of Gaussian channels in information theory [Wibisono and Jog, 2018a,b, Dytso et al., 2023b] that will facilitate our proof.

Proof of Lemma 2.20. By Bayes' rule, $\mathrm{Law}(\mathsf{Y}|\mathsf{X}_t = x) = p(y|t,x)$ can be represented as

$$\begin{aligned}
p(y|t,x) &= \varphi_{b_t y, a_t^2}(x) p_1(y)/p_t(x) \\
&= (2\pi)^{-d/2} a_t^{-d} \exp\left(-\frac{\|x - b_t y\|^2}{2a_t^2}\right) p_1(y)/p_t(x) \\
&= (2\pi)^{-d/2} a_t^{-d} \exp\left(-\frac{\|x\|^2}{2a_t^2} + \frac{b_t\langle x, y\rangle}{a_t^2} - \frac{b_t^2\|y\|^2}{2a_t^2}\right) p_1(y)/p_t(x) \\
&= \left\{\exp\left(\frac{b_t\langle x, y\rangle}{a_t^2} - \frac{b_t^2\|y\|^2}{2a_t^2}\right) p_1(y)\right\} \Big/ \left\{(2\pi)^{d/2} a_t^d \exp\left(\frac{\|x\|^2}{2a_t^2}\right) p_t(x)\right\}.
\end{aligned}$$

Let $\theta = \frac{b_t x}{a_t^2}, h(y) = p_1(y)\exp(-\frac{b_t^2\|y\|^2}{2a_t^2})$, and the logarithmic partition function

$$A(\theta) = \log \int_{\mathbb{R}^d} h(y)\exp(\langle y, \theta\rangle)\mathrm{d}y,$$

then by the definition of exponential family distributions, we conclude that

$$p(y|t,x) = h(y)\exp(\langle y, \theta\rangle - A(\theta))$$

is an exponential family distribution of y. By simple calculation, it follows that

$$\nabla_x^2 \log p(y|t,x) = -\frac{b_t^2}{a_t^4}\nabla_\theta^2 A(\theta).$$

For an exponential family distribution, a basic equality shows that

$$\nabla_\theta^2 A(\theta) = \mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x),$$

which further yields that $\nabla_x^2 \log p(y|t,x) = -\frac{b_t^2}{a_t^4}\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x)$. □

2.9.2 Auxiliary lemmas for Lipschitz flow maps

The following lemma, due to G. Peano [Hartman, 2002a, Theorem 3.1], describes several meaningful differential equations associated with well-posed flows and supports the derivation of Lipschitz continuity of their flow maps.

Lemma 2.42. *[Ambrosio et al., 2023, Lemma 3.4] Suppose that a flow $(X_t)_{t\in[0,1]}$ is well-posed and its velocity field $v(t,x) : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ is of class C^1. Then the flow map $X_{s,t} : \mathbb{R}^d \to \mathbb{R}^d$ is of class C^1 for any $0 \le s \le t \le 1$. Fix $(s,x) \in [0,1]\times\mathbb{R}^d$ and set the following functions defined with $t \in [s,1]$*

$$y(t) := \nabla_x X_{s,t}(x), \qquad J(t) := (\nabla_x v)(t, X_{s,t}(x)),$$
$$w(t) := \det(\nabla_x X_{s,t}(x)), \qquad b(t) := (\nabla_x \cdot v)(t, X_{s,t}(x)) = \mathrm{Tr}(J(t)).$$

Then $y(t)$ and $w(t)$ are the unique C^1 solutions of the following IVPs

$$\dot{y}(t) = J(t)y(t), \quad y(s) = \mathrm{I}_d, \tag{2.44}$$
$$\dot{w}(t) = b(t)w(t), \quad w(s) = 1. \tag{2.45}$$

We present an upper bound of the Lipschitz constant of its flow map $X_{s,t}(x)$ in Lemma 2.31. The upper bound has been deduced in Mikulincer and Shenfeld [2023], Ambrosio et al. [2023], Dai et al. [2023]. For completeness, we derive it as a direct implication of Eq. (2.44) in Lemma 2.42 and an upper bound of the Jacobian matrix of the velocity field.

Proof of Lemma 2.31. Let $y(u) = \nabla_x X_{s,u}(x)$, $J(u) = (\nabla_x v)(u, X_{s,u}(x))$. Owing to Lemma 2.42, $y(u)$ is of class C^1, and the function $u \mapsto \|y(u)\|_{2,2}$ is absolutely continuous over $[s,t]$. By Lemma 2.42, it follows that

$$\partial_u \|y(u)\|_{2,2}^2 = 2\langle y(u), \dot{y}(u)\rangle = 2\langle y(u), J(u)y(u)\rangle \le 2\theta_u \|y(u)\|_{2,2}^2.$$

Applying Grönwall's inequality yields that $\|y(t)\|_{2,2} \le \exp(\int_s^t \theta_u \mathrm{d}u)$ which concludes the proof. □

Another result is concerning the theorem of instantaneous change of variables that is widely deployed in studying neural ODEs [Chen et al., 2018, Theorem 1]. We also exploit the instantaneous change of variables to prove Proposition 2.39. To make the proof self-contained, we show that the instantaneous change of variables directly follows Eq. (2.45) in Lemma 2.42. Compared with the original proof in [Chen et al., 2018, Theorem 1], we illustrate that the well-posedness of a flow is sufficient to ensure the instantaneous change of variables property, without a boundedness condition on the flow.

Corollary 2.43 (Instantaneous change of variables). *Suppose that a flow $(X_t)_{t\in[0,1]}$ is well-posed with a velocity field $v(t,x) : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ of class C^1 in x. Let $X_0(x) \sim \pi_0(X_0(x))$ be a distribution of the initial value. Then the law of $X_t(x)$ satisfies the following differential equation*

$$\partial_t \log \pi_t(X_t(x)) = -\operatorname{Tr}((\nabla_x v)(t, X_t(x))).$$

Proof. Let $\delta(t) := \det(\nabla_x X_t(x))$. Thanks to Eq. (2.45) in Lemma 2.42, it holds that

$$\dot{\delta}(t) = \operatorname{Tr}((\nabla_x v)(t, X_t(x)))\delta(t), \quad \delta(0) = 1,$$

which implies $\delta(t) > 0$ for $t \in [0,1]$. Notice that $\log \pi_t(X_t(x)) = \log \pi_0(X_0(x)) - \log|\delta(t)|$ by change of variables. Then it follows that $\partial_t \log \pi_t(X_t(x)) = -\operatorname{Tr}((\nabla_x v)(t, X_t(x)))$. □

2.9.3 Proofs of spatial Lipschitz estimates for the velocity field

The main results in Section 2.4 are proved in this section. We first present some ancillary lemmas before proceeding to give the proofs.

Lemma 2.44 (Fathi et al., 2023). *Suppose that $f : \mathbb{R}^d \to \mathbb{R}_+$ is L-log-Lipschitz for some $L \geq 0$. Let $\mathcal{P}_t$ be the Ornstein–Uhlenbeck semigroup defined by $\mathcal{P}_t h(x) := \mathbb{E}_{Z\sim\gamma_d}[h(e^{-t}x + \sqrt{1-e^{-2t}}Z)]$ for any $h \in C(\mathbb{R}^d)$ and $t \geq 0$. Then it holds that*

$$\left\{-5Le^{-t}(L + t^{-\frac{1}{2}}) - L^2e^{-2t}\right\}\mathrm{I}_d \leq \nabla_x^2 \log \mathcal{P}_t f(x) \leq \left\{5Le^{-t}(L + t^{-\frac{1}{2}})\right\}\mathrm{I}_d.$$

Proof. This is a restatement of known results. See Proposition 2, Proposition 6, Theorem 6, and their proofs in Fathi et al. [2023]. □

Corollary 2.45. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}_+$ is L-log-Lipschitz for some $L \geq 0$. Let $\mathcal{Q}_t$ be an operator defined by*

$$\mathcal{Q}_t h(x) := \mathbb{E}_{Z\sim\gamma_d}[h(\beta_t x + \alpha_t Z)] \tag{2.46}$$

for any $h \in C(\mathbb{R}^d)$ and $t \in [0,1]$ where $0 \leq \alpha_t \leq 1, \beta_t \geq 0$ for any $t \in [0,1]$. Then it holds that

$$\left(-A_t - L^2\beta_t^2\right)\mathrm{I}_d \leq \nabla_x^2 \log \mathcal{Q}_t f(x) \leq A_t \mathrm{I}_d,$$

where $A_t := 5L\beta_t^2(1-\alpha_t^2)^{-\frac{1}{2}}(L + (-\frac{1}{2}\log(1-\alpha_t^2))^{-\frac{1}{2}})$.

Proof. It is easy to notice that $\mathcal{Q}_t f(x) = \mathcal{P}_s f(\beta_t e^s x)$ where $s = -\frac{1}{2}\log(1-\alpha_t^2)$. Then it follows that $\nabla_x^2 \log \mathcal{Q}_t f(x) = (\beta_t e^s)^2 (\nabla_x^2 \log \mathcal{P}_s f)(\beta_t e^s x)$ which yields

$$\left(-A_t - L^2\beta_t^2\right)\mathrm{I}_d \leq \nabla_x^2 \log \mathcal{Q}_t f(x) \leq A_t \mathrm{I}_d,$$

where $A_t := 5L\beta_t^2(1-\alpha_t^2)^{-\frac{1}{2}}(L + (-\frac{1}{2}\log(1-\alpha_t^2))^{-\frac{1}{2}})$. □

Lemma 2.46. *The Jacobian matrix of the velocity field (2.9) has an alternative expression over time $t \in (0,1)$, that is,*

$$\nabla_x v(t,x) = \left(\frac{\dot{b}_t}{b_t}a_t^2 - \dot{a}_t a_t\right)\left(\nabla_x^2 \log \widetilde{\mathcal{Q}}_t f(x) - \frac{1}{a_t^2+b_t^2}\mathrm{I}_d\right) + \frac{\dot{b}_t}{b_t}\mathrm{I}_d,$$

where $f(x) := \frac{\mathrm{d}\nu}{\mathrm{d}\gamma_d}(x)$ and $\widetilde{\mathcal{Q}}_t f(x) := \mathbb{E}_{Z\sim\gamma_d}[f(\frac{b_t}{a_t^2+b_t^2}x + \frac{a_t}{\sqrt{a_t^2+b_t^2}}Z)]$.

Proof. By direct calculations, it holds that

$$
\begin{aligned}
p_t(x) &= a_t^{-d} \int_{\mathbb{R}^d} p_1(y) \varphi\left(\frac{x - b_t y}{a_t}\right) \mathrm{d}y = a_t^{-d} \int_{\mathbb{R}^d} f(y) \varphi(y) \varphi\left(\frac{x - b_t y}{a_t}\right) \mathrm{d}y \\
&= a_t^{-d} \varphi\left((a_t^2 + b_t^2)^{-\frac{1}{2}} x\right) \int_{\mathbb{R}^d} f(y) \varphi\left(\left(\frac{a_t}{\sqrt{a_t^2 + b_t^2}}\right)^{-1} \left(y - \frac{b_t}{a_t^2 + b_t^2} x\right)\right) \mathrm{d}y \\
&= a_t^{-d} \varphi\left((a_t^2 + b_t^2)^{-\frac{1}{2}} x\right) \left(\frac{a_t}{\sqrt{a_t^2 + b_t^2}}\right)^d \int_{\mathbb{R}^d} f\left(\frac{b_t}{a_t^2 + b_t^2} x + \frac{a_t}{\sqrt{a_t^2 + b_t^2}} z\right) \mathrm{d}\gamma_d(z) \\
&= (a_t^2 + b_t^2)^{-d/2} \varphi\left((a_t^2 + b_t^2)^{-\frac{1}{2}} x\right) \widetilde{\mathcal{Q}}_t f(x).
\end{aligned}
$$

Taking the logarithm and then the second-order derivative of the equation above, it yields

$$
\nabla_x s(t, x) = \nabla_x^2 \log \widetilde{\mathcal{Q}}_t f(x) - \tfrac{1}{a_t^2 + b_t^2} \mathrm{I}_d.
$$

Recalling that $\nabla_x v(t, x) = \left(\frac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right) \nabla_x s(t, x) + \frac{\dot{b}_t}{b_t} \mathrm{I}_d$, it further yields that

$$
\nabla_x v(t, x) = \left(\frac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right) \nabla_x^2 \log \widetilde{\mathcal{Q}}_t f(x) + \tfrac{\dot{a}_t a_t + \dot{b}_t b_t}{a_t^2 + b_t^2} \mathrm{I}_d,
$$

which completes the proof. □

Corollary 2.47. *Suppose that $f(x) := \frac{\mathrm{d}\nu}{\mathrm{d}\gamma_d}(x)$ is L-log-Lipschitz for some $L \geq 0$. Then for $t \in (0, 1)$, it holds that*

$$
\begin{aligned}
&\left\{\left(\tfrac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right)\left(-B_t - L^2 \left(\tfrac{b_t}{a_t^2 + b_t^2}\right)^2\right) + \tfrac{\dot{a}_t a_t + \dot{b}_t b_t}{a_t^2 + b_t^2}\right\} \mathrm{I}_d \\
&\preceq \nabla_x v(t, x) \preceq \left\{\left(\tfrac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right) B_t + \tfrac{\dot{a}_t a_t + \dot{b}_t b_t}{a_t^2 + b_t^2}\right\} \mathrm{I}_d,
\end{aligned}
$$

where $B_t := 5 L b_t (a_t^2 + b_t^2)^{-\frac{3}{2}} (L + (\log(\sqrt{a_t^2 + b_t^2}/b_t))^{-\frac{1}{2}})$.

Proof. Let $\alpha_t = \frac{a_t}{\sqrt{a_t^2 + b_t^2}}$ and $\beta_t = \frac{b_t}{a_t^2 + b_t^2}$. Then these bounds hold according to Corollary 2.45 and Lemma 2.46. □

Then we are prepared to prove Proposition 2.22. The proof is mainly based on the techniques for bounding conditional covariance matrices that are developed in a series

of work [Wibisono and Jog, 2018a,b, Mikulincer and Shenfeld, 2024, 2023, Chewi and Pooladian, 2022, Dai et al., 2023].

Proof of Proposition 2.22. (a) By Jung's theorem [Danzer et al., 1963, Theorem 2.6], there exists a closed Euclidean ball with radius less than $D := (1/\sqrt{2})\mathrm{diam}(\mathrm{supp}(\nu))$ that contains $\mathrm{supp}(\nu)$ in $\mathbb{R}^d$. Then the desired bounds hold due to $0\mathrm{I}_d \preceq \mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x) \preceq D^2\mathrm{I}_d$ and Eq. (2.17).

(b) Let p_1 be β-semi-log-convex for some $\beta > 0$ on $\mathbb{R}^d$. Then for any $t \in [0,1)$, the conditional distribution $p(y|t,x)$ is $\left(\beta + \frac{b_t^2}{a_t^2}\right)$-semi-log-convex because

$$-\nabla_y^2 \log p(y|t,x) = -\nabla_y^2 \log p_1(y) - \nabla_y^2 \log p(t,x|y) \preceq \left(\beta + \frac{b_t^2}{a_t^2}\right)\mathrm{I}_d.$$

By the Cramér-Rao inequality (2.4), we obtain

$$\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x) \succeq \left(\beta + \frac{b_t^2}{a_t^2}\right)^{-1}\mathrm{I}_d.$$

Therefore, by Eq. (2.17), we obtain

$$\nabla_x v(t,x) \succeq \left\{\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)\frac{b_t^2}{\beta a_t^2 + b_t^2} + \frac{\dot{a}_t}{a_t}\right\}\mathrm{I}_d,$$

which implies

$$\nabla_x v(t,x) \succeq \frac{\beta a_t \dot{a}_t + b_t \dot{b}_t}{\beta a_t^2 + b_t^2}\mathrm{I}_d.$$

In addition, the bound above can be verified at time $t = 1$ by the definition (2.43).

(c) Let p_1 be κ-semi-log-concave for some $\kappa \in \mathbb{R}$. Then for any $t \in [0,1)$, the conditional distribution $p(y|t,x)$ is $\left(\kappa + \frac{b_t^2}{a_t^2}\right)$-semi-log-concave because

$$-\nabla_y^2 \log p(y|t,x) = -\nabla_y^2 \log p_1(y) - \nabla_y^2 \log p(t,x|y) \succeq \left(\kappa + \frac{b_t^2}{a_t^2}\right)\mathrm{I}_d.$$

When $t \in \left\{t : \kappa + \frac{b_t^2}{a_t^2} > 0, t \in (0,1)\right\}$, by the Brascamp-Lieb inequality (2.2), we obtain

$$\mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x) \preceq \left(\kappa + \frac{b_t^2}{a_t^2}\right)^{-1}\mathrm{I}_d.$$

Therefore, by Eq. (2.17), we obtain

$$\nabla_x v(t,x) \preceq \left\{ \left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t} \right) \frac{b_t^2}{\kappa a_t^2 + b_t^2} + \frac{\dot{a}_t}{a_t} \right\} \mathrm{I}_d,$$

which implies

$$\nabla_x v(t,x) \preceq \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2} \mathrm{I}_d.$$

Moreover, the bound above can be verified at time $t = 1$ by the definition (2.43).

(d) Notice that

$$\begin{aligned} p(y|t,x) &= \frac{p(t,x|y)}{p_t(x)} \frac{\mathrm{d}(\gamma_{d,\sigma^2} * \rho)}{\mathrm{d}y} \\ &= A_{x,t} \int_{\mathbb{R}^d} \varphi_{z,\sigma^2}(y) \varphi_{\frac{x}{b_t}, \frac{a_t^2}{b_t^2}}(y) \rho(\mathrm{d}z), \end{aligned}$$

where the prefactor $A_{x,t}$ only depends on x and t. Then it follows that

$$p(y|t,x) = \int_{\mathbb{R}^d} \varphi_{\frac{a_t^2 z + \sigma^2 b_t x}{a_t^2 + \sigma^2 b_t^2}, \frac{\sigma^2 a_t^2}{a_t^2 + \sigma^2 b_t^2}}(y) \tilde{\rho}(\mathrm{d}z)$$

where $\tilde{\rho}$ is a probability measure on $\mathbb{R}^d$ whose density function is a multiple of ρ by a positive function. It also indicates that $\tilde{\rho}$ is supported on the same Euclidean ball as ρ. To further illustrate $p(y|t,x)$, let $\mathsf{Q} \sim \tilde{\rho}$ and $\mathsf{Z} \sim \gamma_d$ be independent. Then it holds that

$$\frac{a_t^2}{a_t^2 + \sigma^2 b_t^2} \mathsf{Q} + \sqrt{\frac{\sigma^2 a_t^2}{a_t^2 + \sigma^2 b_t^2}} \mathsf{Z} + \frac{\sigma^2 b_t}{a_t^2 + \sigma^2 b_t^2} x \sim p(y|t,x).$$

Thus, it holds that

$$\begin{aligned} \mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x) &= \left(\frac{a_t^2}{a_t^2 + \sigma^2 b_t^2} \right)^2 \mathrm{Cov}(\mathsf{Q}) + \frac{\sigma^2 a_t^2}{a_t^2 + \sigma^2 b_t^2} \mathrm{I}_d \\ &\preceq \left\{ \left(\frac{a_t^2}{a_t^2 + \sigma^2 b_t^2} \right)^2 R^2 + \frac{\sigma^2 a_t^2}{a_t^2 + \sigma^2 b_t^2} \right\} \mathrm{I}_d. \end{aligned}$$

By Eq. (2.17), it holds that

$$\nabla_x v(t,x) \preceq \frac{b_t^2}{a_t^2} \left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t} \right) \left(\left(\frac{a_t^2}{a_t^2 + \sigma^2 b_t^2} \right)^2 R^2 + \frac{\sigma^2 a_t^2}{a_t^2 + \sigma^2 b_t^2} \right) \mathrm{I}_d + \frac{\dot{a}_t}{a_t} \mathrm{I}_d,$$

which implies

$$\nabla_x v(t,x) \leq \left\{ \frac{a_t b_t (a_t \dot{b}_t - \dot{a}_t b_t)}{(a_t^2 + \sigma^2 b_t^2)^2} R^2 + \frac{\dot{a}_t a_t + \sigma^2 \dot{b}_t b_t}{a_t^2 + \sigma^2 b_t^2} \right\} \mathrm{I}_d.$$

Analogously, due to $\mathrm{Cov}(\mathrm{Q}) \succeq 0\mathrm{I}_d$, a lower bound would be yielded as follows

$$\nabla_x v(t,x) \geq \frac{\dot{a}_t a_t + \sigma^2 \dot{b}_t b_t}{a_t^2 + \sigma^2 b_t^2} \mathrm{I}_d.$$

Then the results follow by combining the upper and lower bounds.

(e) The result follows from Corollary 2.47.

We complete the proof. □

Proof of Corollary 2.23. Let us consider that $\kappa > 0$ which is divided into two cases where $\kappa D^2 \geq 1$ and $\kappa D^2 < 1$. On the one hand, suppose that the first case $\kappa D^2 \geq 1$ holds. By Proposition 2.22, the κ-based upper bound is tighter, that is,

$$\lambda_{\max}(\nabla_x v(t,x)) \leq \theta_t := \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2}.$$

On the other hand, suppose that the second case $\kappa D^2 < 1$ holds. Let t_1 be defined in Eq. (2.19). Again, by Proposition 2.22, the D^2-based upper bound is tighter over $[0, t_1)$ and the κ-based upper bound is tighter over $[t_1, 1]$, which is denoted by

$$\lambda_{\max}(\nabla_x v(t,x)) \leq \theta_t := \begin{cases} \frac{b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right) D^2 + \frac{\dot{a}_t}{a_t}, & t \in [0, t_1), \\ \frac{\kappa a_t \dot{a}_t + b_t \dot{b}_t}{\kappa a_t^2 + b_t^2}, & t \in [t_1, 1]. \end{cases}$$

This completes the proof. □

Proof of Corollary 2.24. Let $\kappa < 0, D < \infty$ such that $\kappa D^2 < 1$ is fulfilled. Then an argument similar to the proof of Corollary 2.23 yields the desired bounds. □

Proof of Corollary 2.25. The result follows from Proposition 2.22-(d). □

Proof of Corollary 2.26. The L-based upper and lower bounds in Proposition 2.22-(e) would blow up at time $t = 1$ because the term $(\log(\sqrt{a_t^2 + b_t^2}/b_t))^{-\frac{1}{2}}$ in B_t goes to ∞ as $t \to 1$.

To ensure the spatial derivative of the velocity field $v(t,x)$ is upper bounded at time $t = 1$, we additionally require the target measure is κ-semi-log-concave with $\kappa \leq 0$. Hence, a κ-based upper bound is available for $t \in (t_0, 1]$ as shown in Proposition 2.22-(c). Next, these two upper bounds are combined by choosing any $t_2 \in (t_0, 1)$ first. Then we exploit the L-based bound over $[0, t_2)$ and κ-based bound over $[t_0, 1]$. This completes the proof. □

2.9.4 Proofs of well-posedness and Lipschitz flow maps

The proofs of main results in Section 2.5 are offered in the following. Before proceeding, let us introduce some definitions and notations about function spaces that are collected in [Evans, 2010, Chapter 5]. Let $L^1_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^\ell) := \{\text{locally integrable function } u : \mathbb{R}^d \to \mathbb{R}^\ell\}$. For integers $k \geq 0$ and $1 \leq p \leq \infty$, we define the Sobolev space $W^{k,p}(\mathbb{R}^d) := \{u \in L^1_{\text{loc}}(\mathbb{R}^d) | D^\alpha u \text{ exists and } D^\alpha u \in L^p(\mathbb{R}^d) \text{ for } |\alpha| \leq k\}$, where $D^\alpha u$ is the weak derivative of u. Then the local Sobolev space $W^{k,p}_{\text{loc}}(\mathbb{R}^d)$ is defined as the function space such that for any $u \in W^{k,p}_{\text{loc}}(\mathbb{R}^d)$ and any compact set $\Omega \subset \mathbb{R}^d$, $u \in W^{k,p}(\Omega)$. As a result, we denote the vector-valued local Sobolev space by $W^{k,p}_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^\ell)$. Provided that $v(t,x) : [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$, we use $v \in L^1([0,1]; W^{1,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^d))$ to indicate that v has a finite L^1 norm over $(t,x) \in [0,1]\times\mathbb{R}^d$ and $v(t,\cdot) \in W^{1,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^d)$ for any $t \in [0,1]$. Similarly, we say $v \in L^1([0,1]; L^\infty(\mathbb{R}^d;\mathbb{R}^d))$ when v has a finite L^1 norm over $(t,x) \in [0,1]\times\mathbb{R}^d$ and $v(t,\cdot) \in L^\infty(\mathbb{R}^d;\mathbb{R}^d)$ for every $t \in [0,1]$. We will use the definitions and notations in the following proof.

Proof of Theorem 2.27. Under Assumptions 1 and 2, we claim that the velocity field $v(t,x)$ satisfies that for any $A > 0$,

$$v \in L^1([0,1]; W^{1,\infty}_{\text{loc}}([-A,A]^d;\mathbb{R}^d)), \quad \frac{\|v\|_2}{1+\|x\|_2} \in L^1([0,1]; L^\infty([-A,A]^d;\mathbb{R}^d)).$$

where the first condition indicates the velocity field v is locally bounded and locally Lipschitz continuous in x, and the second condition is a growth condition on v. According to the Cauchy-Lipschitz theorem [Ambrosio and Crippa, 2014, Remark 2.4], we have the representation formulae for solutions of the continuity equation. As a result, there exists a flow $(X_t)_{t\in[0,1]}$ uniquely solves the IVP (2.14). Furthermore, the marginal flow

of $(X_t)_{t\in[0,1]}$ satisfies the continuity equation (2.8) in the weak sense. Then it remains to show the velocity field v is locally bounded and locally Lipschitz continuous in x, and satisfies the growth condition. By the lower and upper bounds given in Proposition 2.22, we know that v is globally Lipschitz continuous in x under Assumptions 1 and 2. Indeed, the global Lipschitz continuity leads to local boundedness and linear growth properties by simple arguments. More concretely, for any $t \in (0,1)$, it holds that

$$\begin{aligned} v(t,0) &= \left(\dot{b}_t - \frac{\dot{a}_t}{a_t} b_t\right) \mathbb{E}[\mathsf{X}_1 | \mathsf{X}_t = 0] = \left(\dot{b}_t - \frac{\dot{a}_t}{a_t} b_t\right) \int_{\mathbb{R}^d} y p(y|t,0) \mathrm{d}y \\ &\lesssim \left(\dot{b}_t - \frac{\dot{a}_t}{a_t} b_t\right) \int_{\mathbb{R}^d} y p_1(y) a_t^{-d} \exp\left(-\frac{b_t^2 \|y\|_2^2}{2a_t^2}\right) \mathrm{d}y, \end{aligned}$$

which implies $\|v(t,0)\|_2 < \infty$ due to fast growth of the exponential function. Besides, it holds that $v(0,0) = (\dot{b}_0 - \frac{\dot{a}_0}{a_0} b_0)\mathbb{E}[\mathsf{X}_1|\mathsf{X}_0 = x] < \infty, v(1,0) = \dot{a}_1 a_1 s(1,0) < \infty$. Then by the boundedness of $\|v(t,0)\|_2$ and the global Lipschitz continuity in x over $t \in [0,1]$, we bound $v(t,x)$ as follows

$$\begin{aligned} \|v(t,x)\|_2 &\le \|v(t,0)\|_2 + \|v(t,x) - v(t,0)\|_2 \\ &\le \|v(t,0)\|_2 + \left\{ \sup_{(t,y)\in[0,1]\times\mathbb{R}^d} \|\nabla_y v(t,y)\|_{2,2} \right\} \|x\|_2 \\ &\lesssim \max\{\|x\|_2, 1\}. \end{aligned}$$

Hence, the local boundedness and linear growth properties of v are proved. This completes the proof. □

Proof of Theorem 2.28. The proof is similar to that of Theorem 2.27. □

Proof of Corollary 2.29. A well-posed ODE flow has the time-reversal symmetry [Lamb and Roberts, 1998]. By Theorem 2.27, the desired results are proved. □

Proof of Corollary 2.30. The proof is similar to that of Corollary 2.29. □

Proof of Proposition 2.32. Combining Proposition 2.22-(b), (c), and Lemma 2.31, we complete the proof. □

Proof of Proposition 2.33. Combining Proposition 2.22-(d) and Lemma 2.31, we complete the proof. □

Proof of Corollary 2.36. By Theorem 2.27 and Corollary 2.29, it holds that

$$X_1 \circ X_1^* = X_1 \circ X_1^{-1} = \mathrm{I}_d.$$

This completes the proof. □

Proof of Corollary 2.37. By Theorem 2.27 and Corollary 2.29, it holds that

$$X_{1,1} \circ X_{2,1}^* \circ X_{2,1} \circ X_{1,1}^* = X_{1,1} \circ X_{2,1}^{-1} \circ X_{2,1} \circ X_{1,1}^{-1} = \mathrm{I}_d.$$

This completes the proof. □

Proof of Corollary 2.38. Let Assumptions 2.4 and 2.5 hold. According to Propositions 2.32 and 2.33, $\|\nabla_x X_1(x)\|_{2,2}$ is uniformly bounded for Case (i)-(iii) in Assumption 2.5. For Case (iv), the boundedness of $\|\nabla_x X_1(x)\|_{2,2}$ holds by combining Corollary 2.26 and Lemma 2.31. Using Proposition 2.22, we know that $\|\nabla_x v(t,x)\|_{2,2}$ is uniformly bounded. □

Proof of Proposition 2.35. The proof idea is similar to those of [Ball et al., 2003, Proposition 1] and [Cattiaux and Guillin, 2014, Proposition 18]. Let $f : \Omega \to \mathbb{R}$ be of class C^1 and $\mathsf{X}_t \sim p_t$. First, we consider the case of log-Sobolev inequalities. Using that $\mathsf{Z} \sim \gamma_d$ and $\mathsf{X}_1 \sim \nu$ both satisfy the log-Sobolev inequalities in Definition 2.49, we have

$$\begin{aligned}
&\mathbb{E}[(f^2 \log f^2)(\mathsf{X}_t)] = \mathbb{E}[(f^2 \log f^2)(a_t \mathsf{Z} + b_t \mathsf{X}_1)] \\
\leq &\int \left(\int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right) \log \left(\int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right) \mathrm{d}\nu(x) \\
&+ \int \left(2C_{\mathrm{LS}}(\gamma_d) \int a_t^2 (\|\nabla f\|_2^2)(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right) \mathrm{d}\nu(x) \\
\leq &\left(\int \int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x) \right) \log \left(\int \int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x) \right) \\
&+ 2C_{\mathrm{LS}}(\nu) \int \left\| \nabla_x \left(\int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right)^{\frac{1}{2}} \right\|_2^2 \mathrm{d}\nu(x) \\
&+ 2a_t^2 C_{\mathrm{LS}}(\gamma_d) \int \int (\|\nabla f\|_2^2)(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x) \\
\leq &\ \mathbb{E}[f^2(\mathsf{X}_t)] \log \left(\mathbb{E}[f^2(\mathsf{X}_t)] \right) + 2a_t^2 C_{\mathrm{LS}}(\gamma_d) \mathbb{E}[\|\nabla f(\mathsf{X}_t)\|_2^2] \\
&+ 2C_{\mathrm{LS}}(\nu) \int \left\| \nabla_x \left(\int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right)^{\frac{1}{2}} \right\|_2^2 \mathrm{d}\nu(x).
\end{aligned}$$

By Jensen's inequality and the Cauchy-Schwartz inequality, it holds that

$$\begin{aligned}
&\int \left\| \nabla_x \left(\int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right)^{\frac{1}{2}} \right\|_2^2 \mathrm{d}\nu(x) \\
&\leq b_t^2 \frac{\int \left(\int (\|f \nabla f\|_2)(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right)^2 \mathrm{d}\nu(x)}{\int \int f^2(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x)} \\
&\leq b_t^2 \int \int (\|\nabla f\|_2^2)(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x) \\
&\leq b_t^2 \mathbb{E}[\|\nabla f(\mathsf{X}_t)\|_2^2].
\end{aligned}$$

Hence, combining the equations above and the fact that $C_{\mathrm{LS}}(\gamma_d) \leq 1$ [Gross, 1975], it implies that

$$\mathbb{E}[(f^2 \log f^2)(\mathsf{X}_t)] - \mathbb{E}[f^2(\mathsf{X}_t)] \log \left(\mathbb{E}[f^2(\mathsf{X}_t)] \right) \leq 2 \left[a_t^2 + b_t^2 C_{\mathrm{LS}}(\nu) \right] \mathbb{E}[\|\nabla f(\mathsf{X}_t)\|_2^2],$$

that is, $C_{\mathrm{LS}}(p_t) \leq a_t^2 + b_t^2 C_{\mathrm{LS}}(\nu)$.

Next, we tackle the case of Poincaré inequalities by similar calculations. Using that $\mathsf{Z} \sim \gamma_d$ and $\mathsf{X}_1 \sim \nu$ both satisfy the Poincaré inequalities in Definition 2.50, we have

$$\begin{aligned}
&\mathbb{E}[f^2(\mathsf{X}_t)] = \mathbb{E}[f^2(a_t \mathsf{Z} + b_t \mathsf{X}_1)] \\
&\leq \int \left(\int f(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right)^2 \mathrm{d}\nu(x) \\
&\quad + \int \left(C_{\mathrm{P}}(\gamma_d) \int a_t^2 (\|\nabla f\|_2^2)(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right) \mathrm{d}\nu(x) \\
&\leq \left(\int \int f(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x) \right)^2 \\
&\quad + C_{\mathrm{P}}(\nu) \int \left\| \nabla_x \left(\int f(a_t z + b_t x) \mathrm{d}\gamma_d(z) \right) \right\|_2^2 \mathrm{d}\nu(x) \\
&\quad + a_t^2 C_{\mathrm{P}}(\gamma_d) \int \int (\|\nabla f\|_2^2)(a_t z + b_t x) \mathrm{d}\gamma_d(z) \mathrm{d}\nu(x) \\
&\leq (\mathbb{E}[f(\mathsf{X}_t)])^2 + \left[a_t^2 C_{\mathrm{P}}(\gamma_d) + b_t^2 C_{\mathrm{P}}(\nu) \right] \mathbb{E}[\|\nabla f(\mathsf{X}_t)\|_2^2].
\end{aligned}$$

Combining the expression above and $C_{\mathrm{P}}(\gamma_d) \leq 1$, it implies that

$$\mathbb{E}[f^2(\mathsf{X}_t)] - (\mathbb{E}[f(\mathsf{X}_t)])^2 \leq \left[a_t^2 + b_t^2 C_{\mathrm{P}}(\nu) \right] \mathbb{E}[\|\nabla f(\mathsf{X}_t)\|_2^2],$$

that is, $C_{\mathrm{P}}(p_t) \leq a_t^2 + b_t^2 C_{\mathrm{P}}(\nu)$. This completes the proof. □

2.9.5 Proofs of the stability results

We provide the proofs of the stability results in Section 2.6.

Proof of Proposition 2.39. Let $x_0 = a_0 z + b_0 x_1$ and suppose $X_0(x_0) \sim \mu, X_0(a_0 z) \sim \gamma_{d,a_0^2}$. According to Corollary 2.38, the Lipschitz property of $X_1(x)$ implies that $\|X_1(x_0) - X_1(a_0 z)\| \le C_1 \|x_0 - a_0 z\|$. We consider an integral defined by

$$I_t := \int \|x_0 - a_0 z\|^2 \mathrm{d}\pi_t(X_t(x_0), X_t(a_0 z)),$$

where π_t is a coupling made of the joint distribution of $(X_t(x_0), X_t(a_0 z))$. In particular, the initial value I_0 is computed by

$$I_0 = \int \|x_0 - a_0 z\|^2 p_0(x_0)\varphi(z)\mathrm{d}x_0\mathrm{d}z = \int \|b_0 x_1\|^2 p_1(x_1)\mathrm{d}x_1 = b_0^2 \mathbb{E}_\nu[\|\mathsf{X}_1\|^2].$$

Since $(X_t)_{t\in[0,1]}$ is well-posed with $X_0(x_0) \sim \mu$ or $X_0(a_0 z) \sim \gamma_{d,a_0^2}$, according to Corollary 2.43, the coupling π_t satisfies the following differential equation

$$\partial_t \log \pi_t(X_t(x_0), X_t(a_0 z)) = -\mathrm{Tr}((\nabla_x v)(t, X_t(x_0))) - \mathrm{Tr}((\nabla_x v)(t, X_t(a_0 z))). \tag{2.47}$$

Taking the derivative of I_t and using Eq. (2.47), it implies that

$$\frac{\mathrm{d}I_t}{\mathrm{d}t} \le 2\left(\sup_{(s,x)\in[0,1]\times\mathbb{R}^d} \|\mathrm{Tr}(\nabla_x v(s,x))\|\right) I_t.$$

Thanks to $\|\mathrm{Tr}(\nabla_x v(s,x))\| \le d\|\nabla_x v(s,x)\|_{2,2}$, it follows that

$$\frac{\mathrm{d}I_t}{\mathrm{d}t} \le 2C_2 d I_t, \quad I_0 = b_0^2 \mathbb{E}_\nu[\|\mathsf{X}_1\|^2].$$

By Grönwall's inequality, it holds that $I_t \le b_0^2 \mathbb{E}_\nu[\|\mathsf{X}_1\|^2]\exp(2C_2 dt)$. Therefore, we obtain the following W_2 bound

$$W_2(X_{1\#}\gamma_{d,a_0^2}, \nu) = W_2(X_{1\#}\gamma_{d,a_0^2}, X_{1\#}\mu) \le C_1\sqrt{I_1} \le C_1 b_0 \sqrt{\mathbb{E}_\nu[\|\mathsf{X}_1\|^2]}\exp(C_2 d),$$

which completes the proof. □

Proof of Proposition 2.41. (i) On the one hand, by Corollary 2.38, $v(t,x)$ is Lipschitz continuous in x uniformly over $(t,x) \in [0,1]\times\mathbb{R}^d$ with Lipschitz constant C_2. By the variational equation (2.32) and Lemma 2.31, it follows that

$$\|\nabla_x X_{s,t}(x)\|_{2,2}^2 \le \exp\left(2\int_s^t \theta_u \mathrm{d}u\right).$$

Due to the equality (2.31), we deduce that

$$\begin{aligned}
&\|X_1(x_0) - Y_1(x_0)\|^2 \\
\leq &\left(\int_0^1 \|(\nabla_x X_{s,1})(Y_s(x_0))\|_{2,2} \|v(s, Y_s(x_0)) - \tilde{v}(s, Y_s(x_0))\| \mathrm{d}s \right)^2 \\
\leq &\left(\int_0^1 \|(\nabla_x X_{s,1})(Y_s(x_0))\|_{2,2}^2 \mathrm{d}s \right) \left(\int_0^1 \|v(s, Y_s(x_0)) - \tilde{v}(s, Y_s(x_0))\|^2 \mathrm{d}s \right) \\
\leq &\int_0^1 \exp\left(2 \int_s^1 \theta_u \mathrm{d}u \right) \mathrm{d}s \int_0^1 \|v(s, Y_s(x_0)) - \tilde{v}(s, Y_s(x_0))\|^2 \mathrm{d}s.
\end{aligned}$$

Take expectation and it follows that

$$\begin{aligned}
W_2^2(Y_{1\#}\mu, \nu) &\leq \mathbb{E}_{x_0 \sim \mu} \left[\|Y_1(x_0) - X_1(x_0)\|^2 \right] \\
&\leq \int_0^1 \exp\left(2 \int_s^1 \theta_u \mathrm{d}u \right) \mathrm{d}s \int_0^1 \int_{\mathbb{R}^d} \|v(t,x) - \tilde{v}(t,x)\|^2 \tilde{q}_t(x) \mathrm{d}x \mathrm{d}t \\
&\leq \varepsilon \int_0^1 \exp\left(2 \int_s^1 \theta_u \mathrm{d}u \right) \mathrm{d}s
\end{aligned}$$

where $\tilde{q}_t$ denotes the density function of $Y_{t\#}\mu$, and we use the assumption that

$$\int_0^1 \int_{\mathbb{R}^d} \|v(t,x) - \tilde{v}(t,x)\|^2 \tilde{q}_t(x) \mathrm{d}x \mathrm{d}t \leq \varepsilon$$

in the last inequality.

(ii) On the other hand, suppose that $\tilde{v}(t,x)$ is Lipschitz continuous in x uniformly over $(t,x) \in [0,1] \times \mathbb{R}^d$ with Lipschitz constant C_3. Applying Grönwall's inequality to the variational equation (2.34), it follows that

$$\|\nabla_x Y_{s,t}(x)\|_{2,2}^2 \leq \exp(2C_3(t-s)).$$

By the equality (2.33), it holds that

$$\begin{aligned}
&\|Y_1(x_0) - X_1(x_0)\|^2 \\
\leq &\left(\int_0^1 \|(\nabla_x Y_{s,1})(X_s(x_0))\|_{2,2} \|v(s, X_s(x_0)) - \tilde{v}(s, X_s(x_0))\| \mathrm{d}s \right)^2 \\
\leq &\left(\int_0^1 \|(\nabla_x Y_{s,1})(X_s(x_0))\|_{2,2}^2 \mathrm{d}s \right) \left(\int_0^1 \|v(s, X_s(x_0)) - \tilde{v}(s, X_s(x_0))\|^2 \mathrm{d}s \right) \\
\leq &\frac{\exp(2C_3) - 1}{2C_3} \int_0^1 \|v(s, X_s(x_0)) - \tilde{v}(s, X_s(x_0))\|^2 \mathrm{d}s.
\end{aligned}$$

Taking expectations, it further yields that

$$W_2^2(Y_{1\#}\mu,\nu) \le \mathbb{E}_{x_0\sim\mu}\left[\|Y_1(x_0)-X_1(x_0)\|^2\right]$$
$$\le \frac{\exp(2C_3)-1}{2C_3}\int_0^1\int_{\mathbb{R}^d}\|v(t,x)-\tilde{v}(t,x)\|^2 p_t(x)\mathrm{d}x\mathrm{d}t$$

where $X_t(x_0)\sim p_t$.

□

2.9.6 Time derivative of the velocity field

In this section, we are interested in representing the time derivative of the velocity field via moments of $\mathsf{Y}|\mathsf{X}_t = x$. The result is efficacious for controlling the time derivative with moment estimates, though the computation is somehow tedious.

Proposition 2.48. *The time derivative of the velocity field $v(t,x)$ has an expression with moments of $\mathsf{X}_1|\mathsf{X}_t$ for any $t\in(0,1)$ as follows*

$$\partial_t v(t,x) = \left(\frac{\ddot{a}_t}{a_t}-\frac{\dot{a}_t^2}{a_t^2}\right)x + \left(a_t^2\frac{\ddot{b}_t}{b_t}-\dot{a}_t a_t\frac{\dot{b}_t}{b_t}-\ddot{a}_t a_t+\dot{a}_t^2\right)\frac{b_t}{a_t^2}M_1$$
$$+\frac{b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t}-\frac{\dot{a}_t}{a_t}\right)\left(\frac{\dot{b}_t}{b_t}-2\frac{\dot{a}_t}{a_t}\right)M_2^c x - \frac{b_t^3}{a_t^2}\left(\frac{\dot{b}_t}{b_t}-\frac{\dot{a}_t}{a_t}\right)^2(M_3-M_2M_1),$$

where $M_1 := \mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x], M_2 := \mathbb{E}[\mathsf{X}_1^\top\mathsf{X}_1|\mathsf{X}_t = x], M_2^c := \mathrm{Cov}(\mathsf{X}_1|\mathsf{X}_t = x), M_3 := \mathbb{E}[\mathsf{X}_1\mathsf{X}_1^\top\mathsf{X}_1|\mathsf{X}_t = x]$.

Proof. By direct differentiation, it implies that

$$\partial_t v(t,x) = \partial_t\left(\frac{\dot{b}_t}{b_t}\right)x + \partial_t\left(\frac{\dot{b}_t}{b_t}a_t^2-\dot{a}_t a_t\right)s(t,x) + \left(\frac{\dot{b}_t}{b_t}a_t^2-\dot{a}_t a_t\right)\partial_t s(t,x)$$
$$= \frac{\ddot{b}_t b_t-\dot{b}_t^2}{b_t^2}x + \left(\frac{\ddot{b}_t b_t-\dot{b}_t^2}{b_t^2}a_t^2+\frac{\dot{b}_t}{b_t}2\dot{a}_t a_t-\ddot{a}_t a_t-\dot{a}_t^2\right)s(t,x) + \left(\frac{\dot{b}_t}{b_t}a_t^2-\dot{a}_t a_t\right)\partial_t s(t,x).$$

We first focus on $\partial_t s(t,x)$. Since p_t satisfies the continuity equation (2.8), it holds that

$$\begin{aligned}
\partial_t s(t,x) &= \nabla_x(\partial_t \log p_t(x)) \\
&= -\nabla_x\left(\frac{\nabla_x \cdot (p_t(x)v(t,x))}{p_t(x)}\right) \\
&= -\nabla_x\left(\frac{(\nabla_x p_t(x))^\top v(t,x) + p_t(x)(\nabla_x \cdot v(t,x))}{p_t(x)}\right) \\
&= -\nabla_x\Big(s(t,x)^\top v(t,x) + \nabla_x \cdot v(t,x)\Big) \\
&= -\Big((\nabla_x s(t,x))^\top v(t,x) + (\nabla_x v(t,x))^\top s(t,x) + \nabla_x(\nabla_x \cdot v(t,x))\Big) \\
&= -\left(\nabla_x s(t,x)v(t,x) + \nabla_x v(t,x)s(t,x) + \nabla_x \operatorname{Tr}(\nabla_x v(t,x))\right).
\end{aligned}$$

By direct computation, it holds that

$$\begin{aligned}
&\nabla_x s(t,x)v(t,x) + \nabla_x v(t,x)s(t,x) \\
&= \nabla_x s(t,x)\left(\frac{\dot b_t}{b_t}x + \left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)s(t,x)\right) + \nabla_x\left(\frac{\dot b_t}{b_t}x + \left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)s(t,x)\right)s(t,x) \\
&= \frac{\dot b_t}{b_t}\nabla_x s(t,x)x + \left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)\nabla_x s(t,x)s(t,x) + \frac{\dot b_t}{b_t}s(t,x) + \left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)\nabla_x s(t,x)s(t,x) \\
&= \frac{\dot b_t}{b_t}s(t,x) + \frac{\dot b_t}{b_t}\nabla_x s(t,x)x + 2\left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)\nabla_x s(t,x)s(t,x).
\end{aligned}$$

Then we focus on the trace term

$$\begin{aligned}
&\nabla_x \operatorname{Tr}(\nabla_x v(t,x)) \\
&= \nabla_x \operatorname{Tr}\left(\left(\frac{\dot b_t}{b_t} - \frac{\dot a_t}{a_t}\right)\frac{b_t^2}{a_t^2}\operatorname{Cov}(\mathsf{Y}|\mathsf{X}_t = x) + \frac{\dot a_t}{a_t}\mathrm{I}_d\right) \\
&= \left(\frac{\dot b_t}{b_t} - \frac{\dot a_t}{a_t}\right)\frac{b_t^2}{a_t^2}\nabla_x \operatorname{Tr}(\operatorname{Cov}(\mathsf{Y}|\mathsf{X}_t = x)) \\
&= \left(\frac{\dot b_t}{b_t} - \frac{\dot a_t}{a_t}\right)\frac{b_t^2}{a_t^2}\nabla_x\left(\int \|y\|^2 p(y|t,x)\mathrm{d}y - \left\|\int y p(y|t,x)\mathrm{d}y\right\|^2\right) \\
&= \left(\frac{\dot b_t}{b_t} - \frac{\dot a_t}{a_t}\right)\frac{b_t^2}{a_t^2}\left(\int \|y\|^2 \nabla_x p(y|t,x)\mathrm{d}y - 2\left(\int \nabla_x p(y|t,x)\otimes y \mathrm{d}y\right)\left(\int y p(y|t,x)\mathrm{d}y\right)\right),
\end{aligned}$$

where we notice that

$$\begin{aligned}\nabla_x p(y|t,x) &= \nabla_x\left(\frac{p(t,x|y)p_1(y)}{p_t(x)}\right)\\ &= \frac{\nabla_x p(t,x|y)p_1(y)}{p_t(x)} - \frac{p(t,x|y)p_1(y)}{p_t(x)}s(t,x)\\ &= p(y|t,x)\left(\frac{b_t y - x}{a_t^2} - s(t,x)\right).\end{aligned}$$

For ease of presentation, we introduce the following notations to denote several moments of $\mathsf{Y}|\mathsf{X}_t = x$

$$\begin{aligned} M_1 &:= \mathbb{E}[\mathsf{Y}|\mathsf{X}_t = x], & M_2 &:= \mathbb{E}[\mathsf{Y}^\top\mathsf{Y}|\mathsf{X}_t = x],\\ M_2^c &:= \mathrm{Cov}(\mathsf{Y}|\mathsf{X}_t = x), & M_3 &:= \mathbb{E}[\mathsf{Y}\mathsf{Y}^\top\mathsf{Y}|\mathsf{X}_t = x].\end{aligned}$$

By Tweedie's formula in Lemma 2.51, it yields $s(t,x) = \frac{b_t}{a_t^2}M_1 - \frac{1}{a_t^2}x$. By this expression of $s(t,x)$, it yields

$$\begin{aligned}&\nabla_x s(t,x)v(t,x) + \nabla_x v(t,x)s(t,x)\\ &= \frac{\dot b_t}{b_t}s(t,x) + \frac{\dot b_t}{b_t}\nabla_x s(t,x)x + 2\left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)\nabla_x s(t,x)s(t,x)\\ &= \frac{\dot b_t}{b_t}\left(\frac{b_t}{a_t^2}M_1 - \frac{1}{a_t^2}x\right) + \frac{\dot b_t}{b_t}\left(\frac{b_t^2}{a_t^4}M_2^c - \frac{1}{a_t^2}\mathrm{I}_d\right)x\\ &\quad + 2\left(\frac{\dot b_t}{b_t}a_t^2 - \dot a_t a_t\right)\left(\frac{b_t^2}{a_t^4}M_2^c - \frac{1}{a_t^2}\mathrm{I}_d\right)\left(\frac{b_t}{a_t^2}M_1 - \frac{1}{a_t^2}x\right)\\ &= -2\frac{\dot a_t}{a_t^3}x + \frac{b_t}{a_t^2}\left(2\frac{\dot a_t}{a_t} - \frac{\dot b_t}{b_t}\right)M_1 + \frac{b_t^2}{a_t^4}\left(2\frac{\dot a_t}{a_t} - \frac{\dot b_t}{b_t}\right)M_2^c x + 2\frac{b_t^3}{a_t^4}\left(\frac{\dot b_t}{b_t} - \frac{\dot a_t}{a_t}\right)M_2^c M_1\end{aligned}$$

and $\nabla_x p(y|t,x) = \frac{b_t}{a_t^2}(y - M_1)\, p(y|t,x)$. Therefore, we obtain

$$
\begin{aligned}
&\int \|y\|^2 \nabla_x p(y|t,x)\mathrm{d}y - 2\left(\int \nabla_x p(y|t,x) \otimes y \mathrm{d}y\right)\left(\int y p(y|t,x)\mathrm{d}y\right) \\
&= \int \|y\|^2 \frac{b_t}{a_t^2}(y - M_1)\, p(y|t,x)\mathrm{d}y - 2\left(\int \frac{b_t}{a_t^2}(y - M_1) \otimes y p(y|t,x)\mathrm{d}y\right)\left(\int y p(y|t,x)\mathrm{d}y\right) \\
&= \frac{b_t}{a_t^2}\left[\int \|y\|^2 y p(y|t,x)\mathrm{d}y - \left(\int \|y\|^2 p(y|t,x)\mathrm{d}y\right) M_1 \right. \\
&\qquad \left. -2\left(\int y \otimes y p(y|t,x)\mathrm{d}y - M_1 \otimes \int y p(y|t,x)\mathrm{d}y\right)\left(\int y p(y|t,x)\mathrm{d}y\right)\right] \\
&= \frac{b_t}{a_t^2}\left(M_3 - M_2 M_1 - 2M_2^c M_1\right).
\end{aligned}
$$

Combining the equations above, we obtain

$$
\begin{aligned}
\partial_t v(t,x) = &\frac{\ddot{b}_t b_t - \dot{b}_t^2}{b_t^2} x + \left(\frac{\ddot{b}_t b_t - \dot{b}_t^2}{b_t^2} a_t^2 + \frac{\dot{b}_t}{b_t} 2\dot{a}_t a_t - \ddot{a}_t a_t - \dot{a}_t^2\right)\left(\frac{b_t}{a_t^2} M_1 - \frac{1}{a_t^2} x\right) \\
&- \left(\frac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right)\left[-2\frac{\dot{a}_t}{a_t^3} x + \frac{b_t}{a_t^2}\left(2\frac{\dot{a}_t}{a_t} - \frac{\dot{b}_t}{b_t}\right) M_1 \right. \\
&\qquad \left. + \frac{b_t^2}{a_t^4}\left(2\frac{\dot{a}_t}{a_t} - \frac{\dot{b}_t}{b_t}\right) M_2^c x + 2\frac{b_t^3}{a_t^4}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right) M_2^c M_1\right] \\
&- \left(\frac{\dot{b}_t}{b_t} a_t^2 - \dot{a}_t a_t\right)\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)\frac{b_t^3}{a_t^4}(M_3 - M_2 M_1 - 2M_2^c M_1) \\
= &\left(\frac{\ddot{a}_t}{a_t} - \frac{\dot{a}_t^2}{a_t^2}\right) x + \left(a_t^2 \frac{\ddot{b}_t}{b_t} - \dot{a}_t a_t \frac{\dot{b}_t}{b_t} - \ddot{a}_t a_t + \dot{a}_t^2\right)\frac{b_t}{a_t^2} M_1 \\
&+ \frac{b_t^2}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)\left(\frac{\dot{b}_t}{b_t} - 2\frac{\dot{a}_t}{a_t}\right) M_2^c x - \frac{b_t^3}{a_t^2}\left(\frac{\dot{b}_t}{b_t} - \frac{\dot{a}_t}{a_t}\right)^2 (M_3 - M_2 M_1).
\end{aligned}
$$

Then we complete the proof. □

2.9.7 Functional inequalities and Tweedie's formula

This section is devoted to an exposition of functional inequalities and Tweedie's formula that would assist in our proof.

For a probability measure μ on a compact set $\Omega \subset \mathbb{R}^d$, we define the variance of a

function $f \in L^2(\Omega, \mu)$ as

$$\mathrm{Var}_\mu(f) := \int_\Omega f^2 \mathrm{d}\mu - \left(\int_\Omega f \mathrm{d}\mu \right)^2.$$

Moreover, for a probability measure μ on a compact set $\Omega \subset \mathbb{R}^d$ and any positive integrable function $f : \Omega \to \mathbb{R}$ such that $\int_\Omega f |\log f| \mathrm{d}\nu < \infty$, we define the entropy of f as

$$\mathrm{Ent}_\mu(f) := \int_\Omega f \log f \, \mathrm{d}\mu - \int_\Omega f \mathrm{d}\mu \log \left(\int_\Omega f \mathrm{d}\mu \right).$$

Definition 2.49 (Log-Sobolev inequality)**.** A probability measure $\mu \in \mathcal{P}(\Omega)$ is said to satisfy a log-Sobolev inequality with constant $C > 0$, if for all functions $f : \Omega \to \mathbb{R}$, it holds that

$$\mathrm{Ent}_\mu(f^2) \leq 2C \int_\Omega \|\nabla f\|_2^2 \mathrm{d}\mu.$$

The best constant $C > 0$ for which such an inequality holds is referred to as the log-Sobolev constant $C_{\mathrm{LS}}(\mu)$.

Definition 2.50 (Poincaré inequality)**.** A probability measure $\mu \in \mathcal{P}(\Omega)$ is said to satisfy a Poincaré inequality with constant $C > 0$, if for all functions $f : \Omega \to \mathbb{R}$, it holds that

$$\mathrm{Var}_\mu(f) \leq C \int_\Omega \|\nabla f\|_2^2 \mathrm{d}\mu.$$

The best constant $C > 0$ for which such an inequality holds is referred to as the Poincaré constant $C_{\mathrm{P}}(\mu)$.

Finally, for ease of reference, we present Tweedie's formula that was first reported in Robbins [1956], and then was used as a simple empirical Bayes approach for correcting selection bias [Efron, 2011]. Here, we use Tweedie's formula to link the score function with the expectation conditioned on an observation with Gaussian noise.

Lemma 2.51 (Tweedie's formula)**.** *Suppose that* $\mathsf{X} \sim \mu$ *and* $\epsilon \sim \gamma_{d,\sigma^2}$*. Let* $\mathsf{Y} = \mathsf{X} + \epsilon$ *and* $p(y)$ *be the marginal density of* Y*. Then* $\mathbb{E}[\mathsf{X}|\mathsf{Y} = y] = y + \sigma^2 \nabla_y \log p(y)$.

Chapter 3

Convergence of Continuous Normalizing Flows

Continuous normalizing flows are a generative method for learning probability distributions, which is based on ODEs. This method has shown remarkable empirical success across various applications, including large-scale image synthesis, protein structure prediction, and molecule generation. In this chapter, we study the theoretical properties of CNFs with linear interpolation in learning probability distributions from a finite random sample, using a flow matching objective function. We establish non-asymptotic error bounds for the distribution estimator based on CNFs, in terms of the Wasserstein-2 distance. We present a convergence analysis framework that encompasses the error due to velocity estimation, the discretization error, and the early stopping error. A key step in our analysis involves establishing the regularity properties of the velocity field and its estimator for CNFs constructed with linear interpolation. This necessitates the development of uniform error bounds with Lipschitz regularity control of deep ReLU networks that approximate the Lipschitz function class, which could be of independent interest. Our nonparametric convergence analysis offers theoretical guarantees for using CNFs to learn probability distributions from a finite random sample.

3.1 Introduction

In this chapter, we study the theoretical properties of simulation-free CNFs. We develop a general framework for error analyses of CNFs with flow matching for learning probability distributions based on a random sample. Central to simulation-free CNFs, deep ReLU networks are employed for function approximation and nonparametric estimation of the velocity field. We establish the approximation properties of deep ReLU networks with Lipschitz regularity control, which is essential for analyzing the impact of the estimated velocity field on the distribution of the data generated through the flow. In particular, it is crucial to control the Lipschitz regularity of the estimated velocity field to ensure that the associated IVP is well-posed.

3.1.1 Preview of main results

The following informal descriptions provide a preview of our main results.

Our first main result concerns the regularity of the velocity fields of the CNFs constructed with linear interpolation. For a detailed definition of such CNFs, see Lemma 3.12 and equation (3.3).

Theorem 3.1 (Informal). *Assume that the target distribution either has a bounded support, is strongly log-concave, or is a mixture of Gaussians. Let $0 < \underline{t} \ll 1$. The velocity fields of the CNFs with linear interpolation have the following regularity properties:*

(i) The velocity field v^* is Lipschitz continuous in the space variable x for $(t, x) \in [0, 1] \times \mathbb{R}^d$, where the Lipschitz constant is uniformly bounded;

(ii) The velocity field v^* is Lipschitz continuous in the time variable t for $(t, x) \in [0, 1 - \underline{t}] \times \mathbb{R}^d$, where the Lipschitz constant grows at the order of $\mathcal{O}(\underline{t}^{-2})$ as $\underline{t} \downarrow 0$;

(iii) The velocity field v^* spatially has a linear growth on $\mathbb{R}^d$ for each $t \in [0, 1]$.

Remark 3.2. The regularity properties of the velocity fields stated in Theorem 3.1 are derived from the assumptions made about the underlying target distribution. These properties are essential for studying the distributions generated by the corresponding CNFs.

Theorem 3.3 (Informal). *Suppose that the target distribution is strongly log-concave or is a mixture of Gaussians. Let n be the sample size and $0 < \underline{t} \ll 1$ satisfying $\underline{t} \asymp n^{-1/(d+5)}$. By properly setting the deep ReLU network structure and the forward Euler discretization step sizes, the distribution estimation error of the CNFs learned with linear interpolation and flow matching is evaluated by*

$$\mathbb{E}\mathcal{W}_2(\hat{\nu}_{1-\underline{t}}, \nu) = \widetilde{\mathcal{O}}(n^{-\frac{1}{d+5}}), \tag{3.1}$$

where the expectation is taken with respect to all random samples, $\hat{\nu}_{1-\underline{t}}$ is the law of generated data, ν is the law of target data, $\mathcal{W}_2(\cdot, \cdot)$ is the Wasserstein-2 distance, and a poly-logarithmic prefactor in n is omitted.

Remark 3.4. As can be seen from Theorem 3.1 or Theorem 3.26 below, the velocity fields associated with the CNFs based on linear interpolation may be singular in the time variable at $t = 1$ due to the exploding Lipschitz constant bound. This singularity affects the

convergence rate in (3.1). Without the time singularity of the velocity field, the distribution estimation error would be bounded by $\widetilde{\mathcal{O}}(n^{-1/(d+3)})$. The time singularity leads to a necessary trade-off regarding $\underline{t}$ between the error due to velocity estimation and the early stopping error. The trade-off reduces the nonparametric convergence rate of the distribution estimator to $\widetilde{\mathcal{O}}(n^{-1/(d+5)})$. However, this rate $\widetilde{\mathcal{O}}(n^{-1/(d+5)})$ is slower than the minimax rate $\mathcal{O}(n^{-2/(d+4)})$ for nonparametric density estimation. In our analysis, we first consider the convergence rate of the velocity estimator, when the smoothness index of the velocity function is 1 and an additional time variable is included. Due to the relation between the velocity function and the score function, we know that the smoothness index of the density function differs from that of the velocity function. Therefore, the gap between our derived rate and the optimal rate may be due to the additional time dimension, the loss of smoothness, and the time singularity. See also Remark 3.24 for additional comments and explanation.

3.1.2 Our contributions

We present a comprehensive error analysis of simulation-free CNFs with linear interpolation, trained using flow matching. To the best of our knowledge, this is the first analysis of its kind in the context of simulation-free CNFs. Our results are based solely on assumptions about the target distribution, and all regularity conditions are rigorously derived from these assumptions. Our analysis accurately reflects the practical computational implementation of flow matching for learning simulation-free CNFs. Although our focus is on CNFs based on linear interpolation due to their widespread use and for the sake of simplicity, our analytical framework can be applied to CNFs based on other types of interpolation as well.

We summarize our main contributions into four points:

(1) We establish non-asymptotic error bounds for distribution estimators based on simulation-free CNFs with linear interpolation and flow matching. We present a convergence analysis framework that encompasses the error due to velocity estimation, the discretization error, and the early stopping error. We show that the nonparametric convergence rate of the distribution estimator is $\widetilde{\mathcal{O}}(n^{-1/(d+5)})$ up to a polylogarithmic prefactor in the sample size n.

(2) We derive regularity properties for the velocity field of the CNF with linear interpolation. We demonstrate the Lipschitz regularity properties of the velocity field in both the space and time variables and establish bounds for the Lipschitz constants. We also show that the velocity field grows at most linearly with respect to the space variable.

(3) We establish error bounds for deep ReLU network approximation within the Lipschitz function class, demonstrating that the constructed approximation function maintains Lipschitz regularity. We also derive time-space approximation bounds for approximating the velocity field in both the time and space variables. These time-space approximation bounds are novel in three respects. Firstly, the approximation bounds are derived in terms of the L^∞ norm. Secondly, we demonstrate that the constructed time-space approximation function is Lipschitz in both the time and space variables, with Lipschitz constants of the same order as those of the target function. Lastly, the time-space approximation function can exhibit different Lipschitz regularity in the time and space variables. These neural network approximation results, which maintain Lipschitz regularity, could be of independent interest.

(4) We establish the statistical consistency of the flow matching estimator for the velocity field. By rigorously bounding the stochastic and approximation errors, we show that the convergence rate of the flow matching estimator coincides with the minimax optimal rate of nonparametric estimation of regression functions belonging to the Sobolev space $W^{1,\infty}([0,1]^d)$.

The remainder of this chapter is organized as follows. In Section 3.2, we present the preliminary materials required for subsequent sections. In Section 3.3, we describe simulation-free CNFs and outline the steps for using these CNFs for generative learning. In Section 3.4, we derive our main result concerning the error bounds for the distribution estimator based on CNFs with linear interpolation. In Section 3.5, we first present some useful regularity properties of the velocity field and establish error bounds for the estimated velocity fields through flow matching. Section 3.6 contains discussions on related works in the existing literature.

3.2 Preliminaries

In this section, we summarize several useful definitions, assumptions, the background of CNFs and deep ReLU networks.

3.2.1 Definitions

We list a few useful definitions in this subsection.

The rectified linear unit (ReLU) activation function is defined as $\rho(x) := \max(0, x)$ for $x \in \mathbb{R}$, which also operates coordinate-wise on elements of $x \in \mathbb{R}^d$. For $k \geq 2$, the ReLU^k activation function is given by $\rho_k(x) := (\max(0, x))^k$.

Definition 3.5 (Deep ReLU networks)**.** Deep ReLU networks stand for a class of feed-forward artificial neural networks defined with ReLU activation functions. The function $f_\theta(x) : \mathbb{R}^{\mathsf{k}} \to \mathbb{R}^{\mathsf{d}}$ implemented by a deep ReLU network with parameter θ is expressed as composition of a sequence of functions

$$f_\theta(x) := l_{\mathsf{D}} \circ \rho \circ l_{\mathsf{D}-1} \circ \rho \circ \cdots \circ l_1 \circ \rho \circ l_0(x)$$

for any $x \in \mathbb{R}^{\mathsf{k}}$, where $\rho(x)$ is the ReLU activation function and the depth D is the number of hidden layers. For $i = 0, 1, \cdots, \mathsf{D}$, the i-th layer is represented by $l_i(x) := A_i x + b_i$, where $A_i \in \mathbb{R}^{d_{i+1} \times d_i}$ is the weight matrix, $b_i \in \mathbb{R}^{d_{i+1}}$ is the bias vector, and d_i is the width of the i-th layer. The network f_θ contains $\mathsf{D}+1$ layers in all. We use a $(\mathsf{D}+1)$-dimension vector $(d_0, d_1, \cdots, d_{\mathsf{D}})^\top$ to describe the width of each layer. In particular, $d_0 = \mathsf{k}$ is the dimension of the domain and $d_{\mathsf{D}} = \mathsf{d}$ is the dimension of the codomain. The width W is defined as the maximum width of hidden layers, that is, $\mathsf{W} = \max\{d_1, d_2, \cdots, d_{\mathsf{D}}\}$. The size S denotes the total number of nonzero parameters in the network f_θ. The bound B denotes the L^∞ bound of f_θ, that is, $\sup_{x \in \mathbb{R}^{\mathsf{k}}} \|f_\theta(x)\|_\infty \leq \mathsf{B}$. We denote the function class $\{f_\theta : \mathbb{R}^{\mathsf{k}} \to \mathbb{R}^{\mathsf{d}}\}$ implemented by deep ReLU networks with size S, width W, depth D, and bound B as $\mathcal{NN}(\mathsf{S}, \mathsf{W}, \mathsf{D}, \mathsf{B}, \mathsf{k}, \mathsf{d})$.

Definition 3.6 (Wasserstein-2 distance)**.** The Wasserstein-2 distance between two probability distributions on $\mathbb{R}^d$ is the L^2 optimal transportation cost defined by

$$\mathcal{W}_2(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left(\mathbb{E}_{(\mathsf{X}, \mathsf{Y}) \sim \pi} \|\mathsf{X} - \mathsf{Y}\|_2^2\right)^{1/2},$$

where $\Pi(\mu, \nu)$ denotes the set of all joint probability distributions π whose marginals are respectively μ and ν. A distribution $\pi \in \Pi(\mu, \nu)$ is called a coupling of μ and ν.

3.2.2 Assumptions

We state the assumptions on the target distribution ν on $\mathbb{R}^d$ to support the main results. Below we use $\mathrm{supp}(\nu)$ to denote the support of a probability distribution ν. We also use $\mathrm{diam}(\Omega)$ to represent the diameter of a set $\Omega \subset \mathbb{R}^d$.

Assumption 3.7. The probability distribution ν is absolutely continuous with respect to the Lebesgue measure and has a zero mean.

Assumption 3.8. The probability distribution ν satisfies any one of the following conditions:

(i) ν is β-semi-log-convex for some $\beta > 0$ and κ-semi-log-concave for some $\kappa > 0$ with $\mathrm{supp}(\nu) = \mathbb{R}^d$;

(ii) $\nu = \gamma_{d,\sigma^2} * \rho$ where ρ is a probability distribution supported on a Euclidean ball of radius R on $\mathbb{R}^d$.

Remark 3.9 (Distribution classes)**.** Assumption 3.8 covers two classes of distributions that are of great interest in the literature on generative learning and sampling. Let us briefly remark on the properties of the distributions:

(1) Strongly log-concave distributions are considered in Assumption 3.8-(i). For distributions in this class, the Hessian matrix of the potential function is both positively lower bounded and positively upper bounded.

(2) Mixtures of Gaussians are considered in Assumption 3.8-(ii). A mixture of Gaussians is notably a multimodal probability distribution, which is neither strongly log-concave nor bounded.

Remark 3.10 (Score Lipschitzness)**.** For any $\beta > 0, \kappa \in \mathbb{R}$, and $\kappa \leq \beta$, the β-semi-log-convexity and κ-semi-log-concavity measure the Lipschitzness of the smooth score function $S(x) := \nabla_x \log \frac{\mathrm{d}\nu}{\mathrm{d}x}(x)$ in the sense that $\kappa \mathrm{I}_d \preceq \nabla_x S(x) \preceq \beta \mathrm{I}_d$. The Lipschitzness of the score function for a probability distribution is a common assumption in the literature studying convergence properties of Langevin Monte Carlo algorithms and score-based diffusion models (cf. Dalalyan [2017], Durmus and Moulines [2017], Chen et al. [2023d,a]).

Remark 3.11 (Sub-Gaussianity)**.** The probability distribution ν considered in Assumption 3.8 satisfies the log-Sobolev inequality with a finite constant C_{LSI} depending on $\kappa > 0$ or (σ, R) [Mikulincer and Shenfeld, 2024, Dai et al., 2023]. Let $\mathsf{V} \sim \nu$ and $\mathsf{V} = [\mathsf{V}_1, \mathsf{V}_2, \cdots, \mathsf{V}_d]^\top$. Then a standard Herbst's argument shows that V_i is sub-Gaussian, and its sub-Gaussian norm $\|\mathsf{V}_i\|_{\psi_2} \asymp \sqrt{C_{\mathrm{LSI}}}$ for $1 \le i \le d$ owing to Ledoux [2001, Theorem 5.3]. In addition, a sub-Gaussian random variable has a finite fourth moment.

3.3 Simulation-free continuous normalizing flows

The basic idea of simulation-free CNFs is to construct an ODE-based IVP with a tractable velocity field. The flow map of the IVP pushes forward a simple source distribution onto the underlying target distribution. It is essential to be able to efficiently estimate the velocity field with a random sample from the target distribution. Since CNFs use ODEs to model a target distribution, the corresponding velocity fields depend on the target distribution and may have a complex, unknown structure. Therefore, it is natural to employ nonparametric methods with deep neural networks to estimate velocity fields.

In Table 3.1, we summarize the four steps of using simulation-free CNFs for generative learning, and we discuss each step in detail below.

3.3.1 Construction of simulation-free CNFs

Based on the concept of stochastic interpolation, Liu et al. [2023] and Albergo and Vanden-Eijnden [2023] proposed a new class of simulation-free CNFs. Gao et al. [2024a] analyzed probability flows of diffusion models and denoising diffusion implicit models in the framework of stochastic interpolation. Let $a_t : [0,1] \to \mathbb{R}_+, b_t : [0,1] \to \mathbb{R}_+$ satisfy the following conditions:

$$\begin{aligned}
&\dot{a}_t \le 0, \quad \dot{b}_t \ge 0, \quad a_0 > 0, \quad b_0 \ge 0, \quad a_1 = 0, \quad b_1 = 1, \\
&a_t > 0 \ \text{ for any } t \in (0,1), \quad b_t > 0 \ \text{ for any } t \in (0,1), \\
&a_t, b_t \in C^2([0,1)), \quad a_t^2 \in C^1([0,1]), \quad b_t \in C^1([0,1]).
\end{aligned}$$

Then a general class of simulation-free CNFs is constructed in Lemma 3.12 and satisfies the requirements of Step 1 in Table 3.1. For simplicity, we use $\mathrm{Law}(\mathsf{X})$ to denote the distribution of a random variable X.

Task: Generating samples from a distribution approximating the target distribution ν.

Step 1. Construct a simulation-free CNF defined by the IVP such that $\nu = X_{T\#}\mu$:

$$\frac{\mathrm{d}X_t}{\mathrm{d}t}(x) = v(t, X_t(x)), \quad X_0(x) = x \sim \mu, \quad (t,x) \in [\tau, T] \times \mathbb{R}^d,$$

where $T > 0, v : [\tau, T] \times \mathbb{R}^d \to \mathbb{R}^d$ is the velocity field, and μ is a simple source distribution.

Step 2. Estimate the velocity field $v(t,x)$ with a deep neural network $\hat{v}(t,x)$.

Step 3. Use a proper numerical solver to solve the IVP associated with $\hat{v}(t,x)$, and return the numerical solution $\hat{Y}_T(x)$ at time $t = T$:

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t}(x) = \hat{v}(t, Y_t(x)), \quad Y_0(x) = x \sim \mu, \quad (t,x) \in [\tau, T] \times \mathbb{R}^d.$$

Step 4. Generate samples from $\hat{Y}_{T\#}\mu$, which approximates $\nu = X_{T\#}\mu$.

Table 3.1. Four steps to conduct generative learning via simulation-free CNFs.

Lemma 3.12 (Theorem 5.1 in Gao et al. [2024a]). *Suppose that a probability distribution ν satisfies Assumptions 3.7 and 3.8. Let $\mu = \mathrm{Law}(a_0\mathsf{Z} + b_0\mathsf{X}_1)$ with $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$, $\mathsf{X}_t := a_t\mathsf{Z} + b_t\mathsf{X}_1$ for any $t \in (0,1)$. Let the velocity field $v(t,x)$ be defined by*

$$v(t,x) := \mathbb{E}[\dot{a}_t\mathsf{Z} + \dot{b}_t\mathsf{X}_1 | \mathsf{X}_t = x], \quad (t,x) \in (0,1) \times \mathbb{R}^d,$$

$$v(0,x) := \lim_{t \downarrow 0} v(t,x), \quad v(1,x) := \lim_{t \uparrow 1} v(t,x), \quad x \in \mathbb{R}^d.$$

Then there exists a unique solution $(X_t)_{t\in[0,1]}$ to the IVP

$$\frac{\mathrm{d}X_t}{\mathrm{d}t}(x) = v(t, X_t(x)), \quad X_0(x) = x \sim \mu, \quad (t,x) \in [0,1] \times \mathbb{R}^d. \tag{3.2}$$

Moreover, the push-forward distribution satisfies $X_{t\#}\mu = \mathrm{Law}(a_t\mathsf{Z} + b_t\mathsf{X}_1)$ with $\mathsf{Z} \sim \gamma_d, \mathsf{X}_1 \sim \nu$.

Lemma 3.12 shows that the velocity field $v(t,x)$ of the simulation-free CNF (3.2) takes the form of conditional expectations. As a result, the least squares method, also

known as the flow matching method for simulation-free CNFs [Lipman et al., 2023], is an effective approach to estimating the velocity field $v(t,x)$.

Among various choices of the coefficients a_t and b_t, the linear interpolation scenario, where

$$a_t = 1 - t,\ b_t = t, \tag{3.3}$$

has been shown to have excellent properties for generative learning tasks [Liu et al., 2023, Albergo et al., 2023b]. A CNF model with linear interpolation has been used in the implementation of a large image generative model [Esser et al., 2024]. The coefficients of the linear interpolation are the same as those of the displacement interpolation in optimal transport [McCann, 1997, Villani, 2009]. In this chapter, we focus on the regularity and convergence properties of the CNF with linear interpolation (3.3).

Corollary 3.13. *Suppose that a probability distribution ν satisfies Assumptions 3.7 and 3.8. Let $\mathsf{X}_0 = \mathsf{Z} \sim \gamma_d$, $\mathsf{X}_1 \sim \nu$. Consider the linear interpolation $\mathsf{X}_t := (1-t)\mathsf{Z} + t\mathsf{X}_1$ for any $t \in (0,1)$ and the velocity field $v^*(t,x)$ defined by*

$$v^*(t,x) := \mathbb{E}[\mathsf{X}_1 - \mathsf{Z} | \mathsf{X}_t = x], \quad (t,x) \in [0,1] \times \mathbb{R}^d. \tag{3.4}$$

Then there exists a unique solution $(X_t)_{t\in[0,1]}$ to the IVP

$$\frac{\mathrm{d}X_t}{\mathrm{d}t}(x) = v^*(t, X_t(x)), \quad X_0(x) = x \sim \gamma_d, \quad (t,x) \in [0,1] \times \mathbb{R}^d, \tag{3.5}$$

and the push-forward distribution satisfies $X_{t\#}\gamma_d = \mathrm{Law}(\mathsf{X}_t)$.

Corollary 3.13 implies that the push-forward distributions $(X_{t\#}\gamma_d)_{t\in[0,1]}$ coincide with the marginal distributions of the Gaussian channel $\mathsf{X}_t = (1-t)\mathsf{Z} + t\mathsf{X}_1$. The connections between the velocity fields and Gaussian channels have inspired moment expressions for the derivatives of the velocity field. We show these expressions in Lemmas 3.15 and 3.17 for the purpose of examining the regularity properties of velocity fields.

Remark 3.14. Since $x = \mathbb{E}[(1-t)\mathsf{Z} + t\mathsf{X}_1 | \mathsf{X}_t = x]$ for $(t,x) \in [0,1] \times \mathbb{R}^d$, an alternative expression of the velocity field is given by

$$v^*(t,x) = -\frac{1}{1-t}x + \frac{1}{1-t}\mathbb{E}[\mathsf{X}_1 | \mathsf{X}_t = x], \quad (t,x) \in [0,1) \times \mathbb{R}^d. \tag{3.6}$$

Expression (3.6) shows that the velocity field $v^*(t,x)$ only depends on the conditional expectation $\mathbb{E}[\mathsf{X}_1 | \mathsf{X}_t = x]$.

Lemma 3.15 (Lemma 4.1 in Gao et al. [2024a]). *The Jacobian matrix $\nabla_x v^*(t,x)$ has a covariance expression as follows*

$$\nabla_x v^*(t,x) = \frac{t}{(1-t)^3}\mathrm{Cov}(\mathsf{X}_1|\mathsf{X}_t = x) - \frac{1}{1-t}\mathrm{I}_d, \quad (t,x) \in [0,1)\times\mathbb{R}^d. \tag{3.7}$$

Remark 3.16. The covariance expression (3.7) has been used to derive regularity properties of the velocity field $v^*(t,x)$ in the space variable x. For example, see Proposition 4.1 in Gao et al. [2024a].

Lemma 3.17 (Proposition F.1 in Gao et al. [2024a]). *The time derivative of the velocity field $v^*(t,x)$ has a moment expression for any $t \in [0,1)$ as follows*

$$\partial_t v^*(t,x) = -\frac{1}{(1-t)^2}x + \frac{1}{(1-t)^2}M_1 + \frac{t+1}{(1-t)^4}M_2^c x - \frac{t}{(1-t)^4}(M_3 - M_2 M_1), \tag{3.8}$$

where $M_1 := \mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = x], M_2 := \mathbb{E}[\mathsf{X}_1^\top \mathsf{X}_1|\mathsf{X}_t = x], M_2^c := \mathrm{Cov}(\mathsf{X}_1|\mathsf{X}_t = x)$, and $M_3 := \mathbb{E}[\mathsf{X}_1\mathsf{X}_1^\top \mathsf{X}_1|\mathsf{X}_t = x]$ with omitted dependence on (t,x).

Remark 3.18. To quantify the regularity of the velocity field $v^*(t,x)$ in the time variable t, one can try to bound the moments in Eq. (3.8) defined in Lemma 3.17. Following this idea, we conduct regularity analyses on the velocity field in Section 3.8.1 and summarize the results in Theorem 3.26.

3.3.2 Flow matching

This subsection concerns Step 2 in Table 3.1, which estimates the velocity field of a CNF with linear interpolation. As shown in (3.4), the velocity field $v^*(t,x) = \mathbb{E}[\mathsf{X}_1 - \mathsf{Z}|\mathsf{X}_t = x]$ for each $t \in [0,1]$. For notational simplicity, let $\mathsf{Y} := \mathsf{X}_1 - \mathsf{Z}$, and note that $\mathsf{X}_t = (1-t)\mathsf{Z} + t\mathsf{X}_1$. Given $\tau \in (0,1]$, we consider the time interval $[0,\tau]$. For $t \in [0,\tau]$, we denote $\mathsf{X}_t \sim p_t$. When the time is a random variable distributed as a uniform distribution on $[0,\tau]$, that is, $\mathsf{t} \sim \mathrm{U}(0,\tau)$, we denote $\mathsf{X}_\mathsf{t}|\mathsf{t} = t \sim p_t$. Then the flow matching method [Lipman et al., 2023, Liu et al., 2023] solves a nonlinear least squares problem for estimating $v^*(t,x) = \mathbb{E}[\mathsf{Y}|\mathsf{X}_t = x]$ on the domain $[0,\tau]\times\mathbb{R}^d$:

$$v^* \in \arg\min_v \left\{\mathcal{L}(v) := \mathbb{E}_{\mathsf{t}\sim\mathrm{U}(0,\tau)}\mathbb{E}_{\mathsf{X}_1\sim\nu,\mathsf{Z}\sim\gamma_d}\|v(\mathsf{t},\mathsf{X}_\mathsf{t}) - \mathsf{Y}\|_2^2\right\}. \tag{3.9}$$

In practice, τ is often taken as 1. Here, we leave τ as a quantity adaptive to the time regularity of the velocity field v^*. We will analyze the time regularity of v^* in Subsection 3.5.1.

Let $\{\mathsf{Z}_i\}_{i=1}^n$, $\{\mathsf{X}_{1,i}\}_{i=1}^n$, and $\{\mathsf{t}_i\}_{i=1}^n$ be i.i.d. random samples from γ_d, ν, and $\mathrm{U}(0,\tau)$, respectively. For $i = 1, 2, \cdots, n$, we denote $\mathsf{X}_{\mathsf{t}_i} := (1-\mathsf{t}_i)\mathsf{Z}_i + \mathsf{t}_i\mathsf{X}_{1,i}$ and $\mathsf{Y}_i := \mathsf{X}_{1,i} - \mathsf{Z}_i$. The population risk $\mathcal{L}(v)$ defined in (3.9) leads to the empirical risk

$$\mathcal{L}_n(v) := \frac{1}{n}\sum_{i=1}^{n} \|v(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathsf{Y}_i\|_2^2. \tag{3.10}$$

We approximate the velocity field v^* using deep neural networks. We consider a deep ReLU network class with the input dimension $\mathtt{k} = d+1$ and the output dimension $\mathtt{d} = d$. Let $\mathcal{F}_n := \mathcal{NN}(\mathsf{S}, \mathtt{W}, \mathtt{D}, \mathsf{B}, \mathsf{L}_x, \mathsf{L}_t, d+1, d)$ denote a function class $\{f_\theta(t,x) : \mathbb{R}^{d+1} \to \mathbb{R}^d\}$ implemented by deep ReLU networks with size S, width $\mathtt{W}$, depth $\mathtt{D}$, bound B, and Lipschitz constants at most L_x in x and L_t in t over $(t,x) \in [0,\tau]\times\mathbb{R}^d$. The network parameters can depend on the sample size n, and we show the fact by making $\mathcal{F}_n$ depend on n. For any $f \in \mathcal{F}_n$, the Lipschitz continuity of f implies that $\|f(t,x) - f(s,x)\|_\infty \le \mathsf{L}_t|t-s|$ and $\|f(t,x) - f(t,y)\|_\infty \le \mathsf{L}_x\|x-y\|_\infty$ for any $s,t \in [0,\tau]$ and $x,y \in \mathbb{R}^d$. It is easy to see that $\mathcal{F}_n \subseteq \mathcal{NN}(\mathsf{S}, \mathtt{W}, \mathtt{D}, \mathsf{B}, d+1, d)$, that is a deep ReLU network class without Lipschitz regularity control. To estimate the velocity field within the hypothesis class $\mathcal{F}_n$, we consider the empirical risk minimization problem

$$\hat{v}_n \in \operatorname*{arg\,min}_{v\in\mathcal{F}_n} \mathcal{L}_n(v), \tag{3.11}$$

where $\hat{v}_n$ is a deep ReLU network estimator for the velocity field v^*. We call $\hat{v}_n$ the flow matching estimator because it is a minimizer of the empirical flow matching objective.

3.3.3 Forward Euler discretization

In this subsection, we proceed to Step 3 in Table 3.1, where a numerical solver is used to solve the IVP:

$$\frac{\mathrm{d}\tilde{X}_t}{\mathrm{d}t}(x) = \hat{v}_n(t, \tilde{X}_t(x)), \quad \tilde{X}_0(x) \sim \gamma_d, \quad (t,x) \in [0,\tau]\times\mathbb{R}^d. \tag{3.12}$$

The forward Euler method is a first-order numerical procedure for solving ODE-based IVPs, which is commonly used in algorithms for sampling and generative learning. First,

we set a time grid of $[0, \tau]$ as $t_0 = 0 < t_1 < \cdots < t_K = \tau \leq 1$. The forward Euler method for solving the IVP (3.12) yields the numerical iterations:

$$\frac{\mathrm{d}\hat{X}_t}{\mathrm{d}t}(x) = \hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x)), \quad \hat{X}_0(x) \sim \gamma_d, \quad t \in [t_{k-1}, t_k], \; k = 1, 2, \cdots, K. \tag{3.13}$$

Finally, Step 4 in Table 3.1 is accomplished by drawing samples from γ_d and running the numerical iterations given in (3.13).

3.4 Main result: Error bounds for distribution estimation

In this section, we derive our main result, Theorem 3.23, on the error bounds for the distribution estimator based on CNFs with linear interpolation.

3.4.1 Error decomposition

We begin with three IVPs (3.5), (3.12), and (3.13) in Section 3.3. The IVP (3.5) defines the true process without approximation of the velocity field v^* and discretization in time. The IVP (3.12) defines the neural process resulting from replacing the velocity field v^* with the flow matching estimator $\hat{v}_n$. The IVP (3.13) is the forward Euler discretization counterpart of the IVP (3.12).

Let $0 < \underline{t} \ll 1$. As shown in Theorem 3.1 or Theorem 3.26 below, the Lipschitz constant bound of the velocity field in the time variable is of the order $\mathcal{O}(\underline{t}^{-2})$ on the time interval $[0, 1 - \underline{t}]$. This order shows that the bound explodes at the time $t = 1$. To maintain the Lipschitz regularity in the time variable for flow matching, we need to consider an early stopping time by letting the end time $\tau = 1 - \underline{t}$.

Solving the IVPs (3.5), (3.12), and (3.13), we obtain the flow maps $(X_t)_{t\in[0,1]}$, $(\tilde{X}_t)_{t\in[0,1-\underline{t}]}$, and $(\hat{X}_t)_{t\in[0,1-\underline{t}]}$. To simplify the notations, we denote the push-forward distributions $X_{t\#}\gamma_d, \tilde{X}_{t\#}\gamma_d, \hat{X}_{t\#}\gamma_d$ by $p_t, \tilde{p}_t, \hat{p}_t$, respectively. We summarize the three processes (3.5), (3.12), and (3.13), their corresponding flow maps and push-forward distributions defined by the three IVPs, their corresponding velocity fields and density functions, and sources of errors in Table 3.2. In the first column, we present three processes defined by the three IVPs referred in the second column. The corresponding notations are given in the

following columns. Particularly, we show the error source for each process in the last column.

Process	IVP	Flow map	Velocity field	Push-forward distribution	Density	Error source
True process	(3.5)	$X_t(x)$	$v^*(t, X_t(x))$	$X_{t\#}\gamma_d$	p_t	Early stopping
Neural process	(3.12)	$\tilde{X}_t(x)$	$\hat{v}_n(t, \tilde{X}_t(x))$	$\tilde{X}_{t\#}\gamma_d$	$\tilde{p}_t$	Velocity estimation
Discrete process	(3.13)	$\hat{X}_t(x)$	$\hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x))$	$\hat{X}_{t\#}\gamma_d$	$\hat{p}_t$	Discretization

Table 3.2. A list of three IVPs and related notations defining the generative learning process.

There are three sources of errors introduced in the generative learning process (3.5), (3.12), and (3.13). The discretization error comes from the forward Euler discretization. The error due to velocity estimation results from flow matching with deep ReLU networks. The early stopping error is due to the time singularity of the velocity field at time $t = 1$. We use the Wasserstein-2 distance $\mathcal{W}_2$ to measure the difference between the estimated generative distribution $\hat{p}_{1-\underline{t}}$ and the target distribution p_1. We derive an upper bound for $\mathcal{W}_2(\hat{p}_{1-\underline{t}}, p_1)$, which takes into account all the three sources of error.

It is important to consider the trade-off between the different sources of errors. The early stopping error is reduced when the parameter $\underline{t}$ gets smaller. However, a smaller value of $\underline{t}$ increases the time singularity of the velocity field on the time interval $[0, 1-\underline{t}]$, thus leads to a larger error due to velocity estimation.

Keeping the error trade-off in mind, we consider a basic decomposition of the total error in terms of the Wasserstein-2 distance as follows:

$$\mathcal{W}_2(\hat{p}_{1-\underline{t}}, p_1) \le \underbrace{\mathcal{W}_2(\hat{p}_{1-\underline{t}}, \tilde{p}_{1-\underline{t}})}_{\text{discretization}} + \underbrace{\mathcal{W}_2(\tilde{p}_{1-\underline{t}}, p_{1-\underline{t}})}_{\text{velocity estimation}} + \underbrace{\mathcal{W}_2(p_{1-\underline{t}}, p_1)}_{\text{early stopping}}. \tag{3.14}$$

In (3.14), the first term $\mathcal{W}_2(\hat{p}_{1-\underline{t}}, \tilde{p}_{1-\underline{t}})$ measures the discretization error, the second term $\mathcal{W}_2(\tilde{p}_{1-\underline{t}}, p_{1-\underline{t}})$ measures the error due to velocity estimation, and the third term $\mathcal{W}_2(p_{1-\underline{t}}, p_1)$ measures the early stopping error. We evaluate each error term in Lemmas 3.19, 3.21, and 3.22 below, respectively.

Lemma 3.19 (Discretization error). *Suppose that Assumptions 3.7 and 3.8 hold. Let* $\Upsilon \equiv t_k - t_{k-1}$ *for* $k = 1, 2, \cdots, K$. *Then the discretization error is evaluated by*

$$\mathcal{W}_2(\hat{p}_{1-\underline{t}}, \tilde{p}_{1-\underline{t}}) = \mathcal{O}\Big(\sqrt{d} e^{\mathsf{L}_x}(\mathsf{L}_x \mathsf{B} + \mathsf{L}_t)\Upsilon\Big).$$

Lemma 3.19 shows that the error due to the forward Euler discretization is well controlled when the discretization step size Υ is sufficiently small. We use the perturbation analysis of ODE flows to derive the error bound in Lemma 3.19, and the proof can be found in Section 3.8.4.

Remark 3.20. Lemma 3.19 considers a uniform step size for the forward Euler discretization. For more general choices of the step size, we need the general condition

$$\Upsilon^2 = \sum\nolimits_{k=1}^{K} (t_k - t_{k-1})^3 \tag{3.15}$$

to ensure that the error bound in Lemma 3.19 holds. This can be shown following the proof of Lemma 3.19.

In the sequel, we frequently take expectations over $(\mathsf{t}, \mathsf{X}_\mathsf{t})$ whose joint distribution is specified by $\mathsf{t} \sim \mathrm{U}(0, 1 - \underline{t})$ and $\mathsf{X}_\mathsf{t} | \mathsf{t} = t \sim p_t$. For ease of notation, we may omit the notation of the joint distribution when taking expectations over $(\mathsf{t}, \mathsf{X}_\mathsf{t})$.

Lemma 3.21 (Error due to velocity estimation). *Suppose that Assumptions 3.7 and 3.8 hold, and let* $\hat{v}_n \in \mathcal{F}_n$ *satisfy* $\|\hat{v}_n(t, x) - \hat{v}_n(t, y)\|_\infty \le \mathsf{L}_x \|x - y\|_\infty$ *for any* $t \in [0, 1 - \underline{t}]$ *and* $x, y \in \mathbb{R}^d$. *Then the error due to velocity estimation is bounded by*

$$\mathcal{W}_2^2(\tilde{p}_{1-\underline{t}}, p_{1-\underline{t}}) \le \exp(2\mathsf{L}_x + 1) \mathbb{E}_{(\mathsf{t}, \mathsf{X}_\mathsf{t})} \|\hat{v}_n(\mathsf{t}, \mathsf{X}_\mathsf{t}) - v^*(\mathsf{t}, \mathsf{X}_\mathsf{t})\|_2^2. \tag{3.16}$$

Lemma 3.21 states that the error due to velocity estimation is controlled by the excess risk of the flow matching estimator $\hat{v}_n$ when the Lipschitz constant L_x is bounded. We will analyze the excess risk of the flow matching estimator in Section 3.5.

By combining the excess risk bound for $\hat{v}_n$ and the $\mathcal{W}_2$ distance bound (3.16), we can deduce the error bound attributable to the estimated velocity field. Generally, bounding the error due to velocity estimation involves establishing a perturbation error bound for the ODE flow associated with the velocity field $\hat{v}_n$ or v^*, as well as an estimation error bound for the velocity field v^*. We will use the Grönwall's inequality to establish the perturbation error bound (3.16) based on the $\mathcal{W}_2$ distance. Similar perturbation

error bounds have been obtained in Albergo and Vanden-Eijnden [2023, Proposition 3], Benton et al. [2024b, Theorem 1], and Gao et al. [2024a, Proposition 54]. The proof of Lemma 3.21 is given in Section 3.8.4.

Lemma 3.22 (Early stopping error)**.** *Suppose that Assumptions 3.7 and 3.8 hold. The early stopping error is evaluated by*

$$\mathcal{W}_2(p_{1-\underline{t}}, p_1) \lesssim \underline{t},$$

where we omit a polynomial prefactor in d and $\mathbb{E}[\|X_1\|_2^2]$.

The $\mathcal{W}_2$ distance bound in Lemma 3.22 formalizes the intuition that the early stopping error scales with the early stopping parameter $\underline{t}$. The proof of Lemma 3.22 uses a coupling argument, and is given in Section 3.8.4.

3.4.2 Error bounds for the estimated distribution

We now apply the error bounds in the preceding subsection to derive error bounds for the distribution estimator $\hat{\nu}_{1-\underline{t}}(\mathrm{d}x) = \hat{p}_{1-\underline{t}}(x)\mathrm{d}x$.

By Lemma 3.19, it is clear that the discretization error can be controlled by choosing the step size Υ properly. Lemma 3.21 shows that the error due to velocity estimation is upper bounded by the excess risk of the flow matching estimator $\hat{v}_n$. Furthermore, we will provide a detailed nonparametric analysis of the flow matching estimator $\hat{v}_n$ in Section 3.5.

Before presenting our main result, we first describe the trade-off between the different sources of errors. By Theorem 3.40, the excess risk of the flow matching estimator $\hat{v}_n$ satisfies

$$\mathbb{E}_{\mathbb{D}_n}\mathbb{E}_{(t,X_t)}\|\hat{v}_n(t, X_t) - v^*(t, X_t)\|_2^2 \lesssim (n\underline{t}^2)^{-2/(d+3)}\mathrm{polylog}(n)\log(1/\underline{t}),$$

where $\mathrm{polylog}(n)$ stands for a polylogarithmic prefactor in n. Consequently, the error due to velocity estimation satisfies

$$\mathbb{E}_{\mathbb{D}_n}\mathcal{W}_2(\tilde{p}_{1-\underline{t}}, p_{1-\underline{t}}) \lesssim (n\underline{t}^2)^{-1/(d+3)}\mathrm{polylog}(n)\log(1/\underline{t}),$$

where we use Lemma 3.21 and the bound $L_x \asymp \log(\log n)$. According to Lemma 3.22, the early stopping error is upper bounded by $\mathcal{W}_2(p_{1-\underline{t}}, p_1) \lesssim \underline{t}$. By substituting the error

bounds into the error decomposition (3.14), it follows that

$$\begin{aligned}\mathbb{E}_{\mathbb{D}_n}\mathcal{W}_2(\hat{p}_{1-\underline{t}}, p_1) &\le \mathcal{W}_2(\hat{p}_{1-\underline{t}}, \tilde{p}_{1-\underline{t}}) + \mathbb{E}_{\mathbb{D}_n}\mathcal{W}_2(\tilde{p}_{1-\underline{t}}, p_{1-\underline{t}}) + \mathcal{W}_2(p_{1-\underline{t}}, p_1) \\ &\lesssim \Big\{ \underbrace{e^{\mathsf{L}_x}(\mathsf{L}_x\mathsf{B} + \mathsf{L}_t)\Upsilon}_{\text{controlled by } \Upsilon} + \underbrace{(n\underline{t}^2)^{-1/(d+3)} + \underline{t}}_{\text{trade-off on } \underline{t}} \Big\} \mathrm{polylog}(n)\log(1/\underline{t}) \\ &\lesssim \Big\{ \underline{t}^{-2}\Upsilon + (n\underline{t}^2)^{-1/(d+3)} + \underline{t} \Big\} \mathrm{polylog}(n)\log(1/\underline{t}), \end{aligned} \tag{3.17}$$

where we use the following bounds in deriving (3.17)

$$\mathsf{L}_x \asymp \log(\log n), \quad e^{\mathsf{L}_x} \asymp \log n, \quad \mathsf{B} \asymp \log(\log n), \quad \mathsf{L}_t \asymp \log(\log n)\underline{t}^{-2}.$$

Our main result stated below is obtained by balancing the error terms on the right-hand side of (3.17).

Theorem 3.23 (Distribution estimation error)**.** *Suppose that Assumptions 3.7 and 3.8 hold. Let us set* $A \asymp \log(\log n)$, $NL \asymp n^{d/(2d+10)}$, $\underline{t} \asymp n^{-1/(d+5)}$, *and* $\Upsilon \lesssim n^{-3/(d+5)}$. *We consider the deep ReLU network class* $\mathcal{F}_n = \mathcal{NN}(\mathsf{S}, \mathsf{W}, \mathsf{D}, \mathsf{B}, \mathsf{L}_x, \mathsf{L}_t, d+1, d)$ *whose parameters satisfy the following bounds*

$$\mathsf{S} \asymp \underline{t}^{-2}(NL)^{2/d}(N\log N)^2 L\log L, \quad \mathsf{W} \asymp \underline{t}^{-2}(NL)^{2/d}N\log N,$$
$$\mathsf{D} \asymp L\log L, \quad \mathsf{B} \asymp A, \quad \mathsf{L}_x \asymp A, \quad \mathsf{L}_t \asymp A\underline{t}^{-2}.$$

For any random sample $\mathbb{D}_n := \{(\mathsf{Z}_i, \mathsf{X}_{1,i}, \mathsf{t}_i)\}_{i=1}^n$ *satisfying* $n \ge \mathrm{Pdim}(\mathcal{F}_n)$, *the distribution estimation error of the CNF learned with linear interpolation and flow matching is upper bounded as*

$$\mathbb{E}_{\mathbb{D}_n}\mathcal{W}_2(\hat{p}_{1-\underline{t}}, p_1) = \widetilde{\mathcal{O}}(n^{-\frac{1}{d+5}}),$$

where we omit a polylogarithmic prefactor in n.

The proof of Theorem 3.23 is given in Section 3.8.4.

Remark 3.24. As shown in (3.17), we need to consider a trade-off on the early stopping parameter $\underline{t}$ as well as an appropriate order of the step size Υ. By setting $\underline{t} \asymp n^{-1/(d+5)}$ and $\Upsilon \lesssim n^{-3/(d+5)}$, we attain a concrete convergence rate $\widetilde{\mathcal{O}}(n^{-1/(d+5)})$ of the distribution estimator $\hat{p}_{1-\underline{t}}$.

Remark 3.25. We consider the uniform step size Υ in deriving the distribution estimation error bound in Theorem 3.23. The uniform step size is common in implementing numerical solvers for ODE flows. Additionally, the condition (3.15) in Remark 3.20 provides a guideline on general settings of the discretization step size.

3.5 Error analysis of flow matching

In this section, we first present some useful regularity properties of the velocity field v^*, which are essential to the convergence analysis of the flow matching estimator $\hat{v}_n$ defined in (3.11). The error from the flow matching estimation constitutes a main source of the total error for the distribution estimation given in Theorem 3.23.

3.5.1 Regularity of velocity fields

The regularity properties of the velocity field are needed in studying the nonparametric estimation error of flow matching. We summarize the regularity results of the velocity field in Theorem 3.26.

Theorem 3.26. *Suppose that Assumptions 3.7 and 3.8 are satisfied, and let* $0 < \underline{t} \ll 1$. *Then the velocity field* $v^*(t,x) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ *has the following regularity properties:*

(1) For any $s,t \in [0, 1-\underline{t}]$ *and* $x \in \Omega_A$, $\|v^*(t,x) - v^*(s,x)\|_\infty \le L_t|t-s|$ *with* $L_t \lesssim A\underline{t}^{-2}$;

(2) For any $x, y \in \mathbb{R}^d$ *and* $t \in [0,1]$, $\|v^*(t,x) - v^*(t,y)\|_\infty \le L_x\|x-y\|_\infty$ *with* $L_x \lesssim 1$;

(3) $\sup_{(t,x)\in[0,1]\times\Omega_A} \|v^*(t,x)\|_\infty \le B$ *with* $B \lesssim A$,

where we omit constants in $d, \kappa, \beta, \sigma, R$ *and denote* $\Omega_A := [-A, A]^d$.

Theorem 3.26 states that the Lipschitz regularity of the velocity field v^* holds in the time variable t and the space variable x. Moreover, the Lipschiz constant in x is uniformly bounded for any $(t,x) \in [0,1]\times\mathbb{R}^d$, and the Lipschitz constant in t is bounded for any $(t,x) \in [0, 1-\underline{t}] \times \Omega_A$ but depends on $\underline{t}$. Due to the uniform Lipschitzness in x, the velocity field v^* further satisfies the linear growth property (3).

Remark 3.27. We note that the velocity field may be singular at time $t = 1$, since the Lipschitz constant bound of L_t explodes at time $t = 1$. We quantify the time singularity through the upper bound $L_t \lesssim \underline{t}^{-2}$ where $0 < \underline{t} \ll 1$. Taking the time singularity into account, we set the end time $\tau = 1 - \underline{t}$ in Subsection 3.4.1.

Remark 3.28. The global Lipschitz continuity of the velocity field in x ensures that the associated IVP has a unique solution, according to the Cauchy-Lipschitz theorem [Hartman, 2002b].

Proof idea of Theorem 3.26. The proof of Theorem 3.26 can be found in Section 3.8.1. As shown in Lemmas 3.15 and 3.17, the derivatives of the velocity field v^* can be expressed in terms of the moments of $X_1|X_t$. The key idea of the proof is to bound these moments. Under Assumptions 3.7 and 3.8, Gao et al. [2024a] have shown that the covariance matrix $\mathrm{Cov}(X_1|X_t)$ is both lower and upper bounded uniformly in $t \in [0,1]$ (cf. Lemma 3.42 in Section 3.8.1). As a result, we can prove that the Lipschitz property (1) holds. The linear growth property (3) follows from the Lipschitz property (1). To prove the Lipschitz property (2), we derive bounds for the moments $M_1 = \mathbb{E}[X_1|X_t], M_2 = \mathbb{E}[X_1^\top X_1|X_t]$, and $M_3 = \mathbb{E}[X_1 X_1^\top X_1|X_t]$ by transforming the bounds of the covariance matrix $M_2^c = \mathrm{Cov}(X_1|X_t)$. In particular, the bound transformations can be derived based on the Hatsell-Nolte identity [Dytso et al., 2023b, Proposition 1], the Brascamp-Lieb inequality [Brascamp and Lieb, 1976], and a basic inequality on $M_3 - M_2 M_1$. Moreover, we validate the sharpness of moment bounds using a Gaussian example.

3.5.2 Error decomposition of flow matching

The starting point of our analysis is the decomposition of the excess risk of $\hat{v}_n$ below.

Lemma 3.29. *Let* $t \sim \mathrm{U}(0, 1-\underline{t})$. *For any random sample* $\mathbb{D}_n := \{(Z_i, X_{1,i}, t_i)\}_{i=1}^n$, *the excess risk of the flow matching estimator* $\hat{v}_n$ *satisfies*

$$\begin{aligned}\mathbb{E}_{\mathbb{D}_n}\mathbb{E}_{(t,X_t)}\|\hat{v}_n(t,X_t) - v^*(t,X_t)\|_2^2 &= \mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(\hat{v}_n) - \mathcal{L}(v^*)] \\ &\le \mathcal{E}_{\mathrm{stoc}} + 2\mathcal{E}_{\mathrm{appr}}, \end{aligned} \tag{3.18}$$

where the stochastic error $\mathcal{E}_{\mathrm{stoc}} := \mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(v^*) - 2\mathcal{L}_n(\hat{v}_n) + \mathcal{L}(\hat{v}_n)]$ *and the approximation error* $\mathcal{E}_{\mathrm{appr}} := \inf_{v\in\mathcal{F}_n} \mathbb{E}_{(t,X_t)}\|v(t,X_t) - v^*(t,X_t)\|_2^2$.

The proof of Lemma 3.29 is given in Section 3.8.3. The decomposition (3.18) of the excess risk can be considered a bias-variance decomposition. The stochastic error $\mathcal{E}_{\mathrm{stoc}}$ bounds the variance term of the flow matching estimator, and the approximation error $\mathcal{E}_{\mathrm{appr}}$ represents the bias term of the flow matching estimator. We derive bounds for $\mathcal{E}_{\mathrm{stoc}}$ and $\mathcal{E}_{\mathrm{appr}}$. Then the best bound for the excess risk under the decomposition (3.18) is obtained by balancing these two error bounds.

3.5.3 Approximation error

We derive the error bounds for approximating Lipschitz functions using deep ReLU networks with Lipschitz regularity. The results are presented in Theorem 3.31, which is crucial for bounding the approximation error of the flow matching estimator $\hat{v}_n$. To address the challenge posed by the unbounded support of the velocity field v^* in the space variable x, we use the standard technique of *truncated approximation*. This allows us to divide the approximation error into the truncated approximation error and the truncation error.

Lemma 3.30. *For $\bar{v} \in \mathcal{F}_n$ and any $A > 0$, the approximation error satisfies a basic inequality as follows:*

$$\mathcal{E}_{\text{appr}} = \inf_{v \in \mathcal{F}_n} \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})} \|v(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2 \lesssim \mathcal{E}_{\text{appr}}^{\text{trunc}} + \mathcal{E}_{\text{trunc}}, \tag{3.19}$$

where the truncated approximation error

$$\mathcal{E}_{\text{appr}}^{\text{trunc}} := \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})} \|[\bar{v}(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})] \operatorname{Id}_{\Omega_A}(\mathsf{X}_\mathsf{t})\|_2^2,$$

and the truncation error

$$\mathcal{E}_{\text{trunc}} := \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})} \|[\bar{v}(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})] \operatorname{Id}_{\Omega_A^c}(\mathsf{X}_\mathsf{t})\|_2^2.$$

Lemma 3.30 follows from the triangle inequality and the inequality $2ab \le a^2 + b^2$ for any $a, b \in \mathbb{R}$. We bound the truncated approximation error $\mathcal{E}_{\text{appr}}^{\text{trunc}}$ by considering deep ReLU network approximation of the velocity field on the $(d+1)$-dimensional hypercube. The truncation error $\mathcal{E}_{\text{trunc}}$ measures how fast the approximation error decays according to the tail property of the probability distribution p_t with $t \in [0, 1 - \underline{t}]$.

Approximation with Lipschitz regularity control. We study the capacity of an approximation function $\bar{v}(t,x)$ implemented by a deep ReLU network with Lipschitz regularity for approximating the velocity field $v^*(t,x)$. For balancing the approximation error with the stochastic and discretization errors to obtain an overall error bound for the distribution estimation error, we construct the approximation function $\bar{v}(t,x)$ so that it satisfies the following three requirements:

(a) Good approximation power under the sup norm over the hypercube $\Omega_{\underline{t},A} := [0, 1 - \underline{t}] \times [-A, A]^d$,

(b) Lipschitz continuity with respect to both the time variable t and the space variable x,

(c) Independent regularity in the time variable t and the space variable x.

Let us briefly comment on each of these requirements. Requirement (a) is needed for bounding the approximation error of the flow matching estimator $\hat{v}_n$. The time-space Lipschitz regularity of the approximation function $\bar{v}(t,x)$ required in (b) is essential to bounding the discretization error and the error due to velocity estimation. Requirement (c) stems from the time singularity of the velocity field at $t = 1$ and the different roles of the time regularity and the space regularity in the error analysis.

Theorem 3.31. *For any $N, L \in \mathbb{N}$, there exists a function $\bar{v}(t,x)$ implemented by a deep ReLU network with width $\mathcal{O}(\underline{t}^{-2}(NL)^{2/d}N\log N)$, depth $\mathcal{O}(L\log L)$, and size $\mathcal{O}(\underline{t}^{-2}(NL)^{2/d}(N\log N)^2 L\log L)$ such that the following properties hold simultaneously:*

(i) Boundedness and Lipschitz regularity: for any $s,t \in [0,1-\underline{t}]$ and any $x,y \in \mathbb{R}^d$,

$$\sup_{(t,x)\in[0,1-\underline{t}]\times\mathbb{R}^d} \|\bar{v}(t,x)\|_\infty \lesssim A,$$

$$\sup_{x\in\mathbb{R}^d} \|\bar{v}(t,x)-\bar{v}(s,x)\|_\infty \lesssim A\underline{t}^{-2}|t-s|,$$

$$\sup_{t\in[0,1-\underline{t}]} \|\bar{v}(t,x)-\bar{v}(t,y)\|_\infty \lesssim A\|x-y\|_\infty.$$

(ii) Approximation error bound:

$$\sup_{(t,x)\in\Omega_{\underline{t},A}} \|\bar{v}(t,x)-v^*(t,x)\|_\infty \lesssim A^2(NL)^{-2/d}.$$

Note that we omit some prefactors in d,κ,β,σ,R and denote $\Omega_{\underline{t},A} := [0,1-\underline{t}]\times[-A,A]^d$.

The proof of Theorem 3.31 is given in Section 3.8.2. Let $l \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^l$ denote a subset of $\mathbb{R}^d$. We denote by $L^p(\Omega)$ the standard Lebesgue space on Ω with the Lebesgue norm $\|\cdot\|_{L^p(\Omega)}$ for $p \in [1,\infty]$. Let $k \in \mathbb{N}$. We show the definitions of the Sobolev space $W^{k,\infty}(\Omega)$, the Sobolev norm $\|\cdot\|_{W^{k,\infty}(\Omega)}$, and the Sobolev semi-norm $|\cdot|_{W^{k,\infty}(\Omega)}$ in Section 3.8.5.

Proof idea of Theorem 3.31. To fulfill Requirement (c), we use different approaches to approximation in the time variable and the space variable. Our approximation approaches can ensure the constructed approximation function to be global Lipschitz for fulfilling Requirement (b). We take four steps to construct the time-space approximation function $\bar{v}$ with Lipschitz regularity control.

The first step is to derive an $L^\infty([0,1]^d)$ error bound of using deep ReLU networks for approximating a Lipschitz function in the space variable. We show that the constructed deep neural approximation function is globally Lipschitz in the space variable. The approximation results are presented in Section 3.8.2, and we summarize them here. Lemma 3.49 and Corollary 3.50 show that for any $N, L \in \mathbb{N}$ and any $f \in W^{1,\infty}((0,1)^d)$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(2^d dN \log N)$ and depth $\mathcal{O}(d^2 L \log L)$ such that $|\phi|_{W^{1,\infty}((0,1)^d)} \lesssim |f|_{W^{1,\infty}((0,1)^d)}$ and that $\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim (NL)^{-2/d}$, omitting the prefactors depending only on d.

The second step is to derive an $L^\infty([0,1])$ approximation error bound of deep ReLU networks for approximating a Lipschitz function in the time variable. We establish an $L^\infty([0,1])$ approximation bound with Lipschitz regularity control in Section 3.8.2. Lemma 3.57 states the main results for approximation in time in such a way: for any $M \in \mathbb{N}$ and any $f \in W^{1,\infty}((0,1))$, there exists a function ξ implemented by a deep ReLU network with width $\mathcal{O}(M)$, depth $\mathcal{O}(1)$, and size $\mathcal{O}(M)$ such that $|\xi|_{W^{1,\infty}((0,1))} \lesssim |f|_{W^{1,\infty}((0,1))}$ and that $\|\xi - f\|_{L^\infty([0,1])} \lesssim |f|_{W^{1,\infty}((0,1))}/M$.

In the third step, we combine the constructed approximation in Lemmas 3.49 and 3.57 to establish an $L^\infty(\Omega_{t,A})$ approximation bound for the time-space approximation of the velocity field v^*. This guarantees that Requirement (a) is fulfilled.

The last step is to show that the constructed time-space approximation satisfies the remaining Requirements (b) and (c). We summarize these discussions in Theorem 3.31 and present detailed construction and derivations in the proof of Theorem 3.31.

Remark 3.32 (Optimality)**.** Under the assumption of continuous parameter selection, DeVore et al. [1989, Theorem 4.2] and Yarotsky [2017, Theorem 3] provided a lower bound $\Omega(\epsilon^{-d/k})$ on the number of parameters for parametric approximations in the Sobolev space $W^{k,\infty}([0,1]^d)$, using the approach of continuous nonlinear widths, when the L^∞ approximation error is no more than ϵ. Our approximation rate for the time

variable in the Sobolev space $W^{1,\infty}([0,1])$ matches this lower bound in the sense that a deep ReLU network with size $\mathcal{O}(S)$ can yield an L^∞ approximation error no more than $\mathcal{O}(1/S)$. Suppose that the deep ReLU network has width $\mathcal{O}(N)$, depth $\mathcal{O}(L)$, and size $\mathcal{O}(S)$ with $S \asymp N^2 L$. The approximation rate $\mathcal{O}(S^{-k/d})$ in $W^{k,\infty}([0,1]^d)$ can be improved to the nearly optimal rate $\mathcal{O}((NL)^{-2k/d}\text{polylog}(NL))$ with the bit-extraction technique [Bartlett et al., 1998, 2019, Lu et al., 2021]. Our approximation rate for the space variable in the Sobolev space $W^{1,\infty}([0,1]^d)$ is nearly optimal in the sense that a deep ReLU network with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ can yield an L^∞ approximation error no more than $\mathcal{O}((NL)^{-2/d}(\log(NL))^{2/d})$.

The $L^\infty(\Omega_{\underline{t},A})$ approximation error bound of Theorem 3.31 implies the following L^2 bound of the truncated approximation error for analyzing the flow matching estimator $\hat{v}_n$.

Corollary 3.33. *The truncated approximation error satisfies*

$$\mathcal{E}_{\text{appr}}^{\text{trunc}} = \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|[\bar{v}(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})]\,\mathrm{Id}_{\Omega_A}(\mathsf{X}_\mathsf{t})\|_2^2 \lesssim A^4(NL)^{-4/d},$$

where we omit a constant in $d, \kappa, \beta, \sigma, R$.

As elaborated in the proof of Theorem 3.31 given in Section 3.8.2, the deep ReLU network implementing $\bar{v}$ consists of $\mathcal{O}(\underline{t}^{-2}(NL)^{2/d})$ parallel subnetworks which have width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$. We take advantage of the parallel structure and the construction of each subnetwork to estimate the complexity of the deep ReLU network class $\mathcal{F}_n$ implementing $\bar{v}$. We derive the complexity of the deep ReLU network class $\mathcal{F}_n$ in Lemma 3.34.

Lemma 3.34. *Suppose that Assumptions 3.7 and 3.8 hold. The complexity of the deep ReLU network class $\mathcal{F}_n$ implementing $\bar{v}$ is quantified by*

$$\mathsf{S} \asymp \underline{t}^{-2}(NL)^{2/d}(N\log N)^2 L\log L, \quad \mathsf{W} \asymp \underline{t}^{-2}(NL)^{2/d} N\log N,$$
$$\mathsf{D} \asymp L\log L, \quad \mathsf{B} \asymp A, \quad \mathsf{L}_x \asymp A, \quad \mathsf{L}_t \asymp A\underline{t}^{-2},$$

where we omit some prefactors in $d, \kappa, \beta, \sigma, R$.

Lemma 3.34 follows from the bounds for the number of parameters in the deep ReLU network implementing $\bar{v}$ in Theorem 3.31.

Through Theorem 3.31 and Corollary 3.33, we have established bounds for the truncated approximation error $\mathcal{E}_{\text{appr}}^{\text{trunc}}$. In what follows, we focus on the truncation error $\mathcal{E}_{\text{trunc}}$. We show the sub-Gaussian property of $\mathsf{X}_t \sim p_t$ in Lemma 3.35 under Assumptions 3.7 and 3.8. In Lemma 3.36, we prove that the truncation error $\mathcal{E}_{\text{trunc}}$ decays very fast in the parameter A, as a result of the sub-Gaussian property of p_t.

Lemma 3.35 (Tail probability). *Let* $\mathsf{X}_t = (1-t)\mathsf{Z} + t\mathsf{X}_1$ *with* $\mathsf{Z} \sim \gamma_d$, $\mathsf{X}_1 \sim \nu$, *and* $t \in [0,1]$. *Suppose that Assumptions 3.7 and 3.8 are satisfied. For any* $A > 0$, *it holds that*

$$\sup_{t\in[0,1]} \mathbb{P}(\mathsf{X}_t \in \Omega_A^c) \le 2d \exp\left(-\frac{C_2 A^2}{C_{\text{LSI}}}\right), \tag{3.20}$$

where C_2 *is a universal constant, and* $C_{\text{LSI}} > 0$ *depends on* κ, β, σ, D, *and* R.

Lemma 3.36 (Truncation error). *Suppose that Assumptions 3.7 and 3.8 are satisfied. For any* $A > 0$, *the truncation error satisfies*

$$\mathcal{E}_{\text{trunc}} = \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})} \| [\bar{v}(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})] \, \mathrm{Id}_{\Omega_A^c}(\mathsf{X}_\mathsf{t}) \|_2^2 \lesssim A^2 \exp(-C_3 A^2 / C_{\text{LSI}}),$$

where C_3 *is a universal constant, and we omit a constant in* $d, \kappa, \beta, \sigma, R$, *and the fourth moment of the target* X_1.

The proofs of Lemmas 3.35 and 3.36 are given in Section 3.8.3. We are now ready to provide an upper bound for the approximation error $\mathcal{E}_{\text{appr}}$.

Corollary 3.37. *Suppose that Assumptions 3.7 and 3.8 hold. For any* $N, L \in \mathbb{N}$ *and* $A > 0$, *the approximation error is evaluated by*

$$\mathcal{E}_{\text{appr}} \lesssim A^4 (NL)^{-4/d} + A^2 \exp(-C_3 A^2 / C_{\text{LSI}}).$$

Corollary 3.37 holds by combining (3.19) in Lemma 3.30, Lemma 3.36, and Corollary 3.33.

3.5.4 Stochastic error

We now establish upper bounds for the stochastic error of the estimated velocity field based on a class of deep ReLU networks.

Lemma 3.38. *Consider the flow matching model and the hypothesis class* $\mathcal{F}_n \subseteq \mathcal{NN}(\mathsf{S}, \mathtt{W}, \mathsf{D}, \mathsf{B}, d+1, d)$. *For any* $n \in \mathbb{N}$ *satisfying* $n \ge \mathrm{Pdim}(\mathcal{F}_n)$, *the stochastic error satisfies*

$$\mathcal{E}_{\text{stoc}} = \mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(v^*) - 2\mathcal{L}_n(\hat{v}_n) + \mathcal{L}(\hat{v}_n)] \lesssim \frac{1}{n} (\log n)^4 d A^4 \mathsf{S}\mathsf{D} \log(\mathsf{S}) \log(An^2).$$

The proof of Lemma 3.38 is given in Section 3.8.3.

Corollary 3.39. *Suppose that Assumptions 3.7 and 3.8 hold. The stochastic error satisfies*

$$\mathcal{E}_{\text{stoc}} \lesssim \frac{1}{n}\underline{t}^{-2}(NL)^{2+2/d}(\log N \log L)^2 A^4 \log(A)\log(\underline{t}^{-2}(NL)^{2/d}(N\log N)^2 L\log L),$$

where we omit a polylogarithmic prefactor in n and a prefactor in $d, \kappa, \beta, \sigma, R$.

Corollary 3.39 follows from Lemmas 3.34 and 3.38.

3.5.5 Overall error bound for the estimated velocity field

By Lemma 3.29, the overall error for the estimated velocity field is bounded by the sum of the approximation error $\mathcal{E}_{\text{appr}}$ and the stochastic error $\mathcal{E}_{\text{stoc}}$. The approximation error is analyzed in Subsection 3.5.3 and an upper bound is given in Corollary 3.37. It decreases at a fast rate as the depth and width of the deep ReLU networks grow. The stochastic error $\mathcal{E}_{\text{stoc}}$ is analyzed in Subsection 3.5.4 and its upper bound is provided in Corollary 3.39. The stochastic error increases when the size and depth of the deep ReLU networks grow, as a result of the increasing complexity of the hypothesis class $\mathcal{F}_n$. By balancing the bounds for $\mathcal{E}_{\text{appr}}$ and $\mathcal{E}_{\text{stoc}}$, we obtain the best error bound for the flow matching estimator $\hat{v}_n$ under the error decomposition in Lemma 3.29.

Theorem 3.40 (Flow matching error)**.** *Suppose that Assumptions 3.7 and 3.8 are satisfied. Let $NL \asymp (n\underline{t}^2)^{d/(2d+6)}$ and $A \asymp \log(\log n)$. Then the excess risk of flow matching satisfies*

$$\mathbb{E}_{\mathbb{D}_n}\mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|\hat{v}_n(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2 \lesssim (n\underline{t}^2)^{-2/(d+3)},$$

where we omit a polylogarithmic prefactor in n, a prefactor in $\log(1/\underline{t})$, and a prefactor in $d, \kappa, \beta, \sigma, D$, and R.

The proof of Theorem 3.40 is given in Section 3.8.3.

Remark 3.41. In Theorem 3.40, the polynomial prefactor in $1/\underline{t}$ is due to the singularity of v^* in the time variable t. Without the singularity at time $t = 1$, the convergence rate of the flow matching error in Theorem 3.40 becomes $n^{-2/(d+3)}\text{polylog}(n)$, which is nearly minimax optimal for nonparametric least squares regression in the Sobolev space $W^{1,\infty}([0,1]^{d+1})$ according to Stone [1982].

3.6 Related work

Process-based generative models aim to construct a stochastic process that transports an easy-to-sample source probability distribution to the target distribution. This goal is achieved by estimating a nonlinear transport map implemented through deep neural networks based on a random sample from the target distribution. CNFs and diffusion models are two prominent approaches that have been developed for deep generative learning. Many researchers have considered the theoretical properties of various generative learning methods. In this section, we discuss the connections and differences between this chapter and the existing studies. We focus on the studies concerning CNFs and diffusion models that are most relevant to this chapter. We also discuss the differences between the neural network approximation theory developed In this chapter and those in the existing literature, focusing on the regularity properties of the neural network functions. In particular, we highlight the fact that our approximation results concern velocity field functions that have different regularities in the space and time variables, while the existing results are only applicable to functions with the same regularity in all the variables.

3.6.1 Continuous normalizing flows

CNFs are an ODE-based generative learning approach which estimates a stochastic process for sampling from the target distribution. Marzouk et al. [2023] conducted a nonparametric statistical convergence analysis for simulation-based CNF distribution estimators trained through likelihood maximization. However, this analysis does not extend to simulation-free CNFs. Probability flow ODEs [Song et al., 2021b], denoising diffusion implicit models (DDIMs) [Song et al., 2021a], and flow matching methods [Liu et al., 2023, Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023] all fall under the category of simulation-free CNFs. In these models, either the score function or the velocity field is estimated. The overall error analysis needs to address both the estimation error of the velocity field (or score function) and the discretization error.

In existing literature, it is typical to assume strong regularity conditions directly on the velocity field (or score function) and its estimator. Furthermore, current studies often only consider certain sources of errors, neglecting either the discretization error or

the estimation error of the velocity field (or score function). In contrast, our results are derived based on assumptions about the target distribution. Additionally, our analysis encompasses the error due to velocity estimation, the discretization error of the forward Euler solver, and the early stopping error. These errors are included in the overall error bound. We provide a summary of the comparison between this chapter and relevant existing studies in Table 3.3. We use W_2, KL, TV to represent the Wasserstein-2 distance, the Kullback-Leibler divergence, and the total variation distance. We say a numerical sampler is "mixed" if it is a combination of deterministic and stochastic samplers. For assumptions on velocity fields or velocity field estimators, we mark "Yes" if the assumptions are required and "No" if not. Since the unknown nonlinear part of the velocity field is a score function, assumptions and estimation error bounds on score functions or assumptions on score estimators can be regarded as those on velocity fields or velocity field estimators.

	Metric	Sampler	Estimation error bound of velocity fields	Perturbation error bound	Discretization error bound	Assumptions on velocity fields	Assumptions on estimated velocity fields
Albergo and Vanden-Eijnden [2023]	W_2	Deterministic	✗	✓	✗	No	Yes
Chen et al. [2023e]	KL	Deterministic	✗	✗	✓	Yes	Yes
Albergo et al. [2023b]	KL	Deterministic	✗	✓	✗	No	No
Chen et al. [2023c]	TV	Mixed	✗	✓	✓	Yes	Yes
Benton et al. [2024b]	W_2	Deterministic	✗	✓	✗	Yes	Yes
Li et al. [2024]	TV	Deterministic	✗	✓	✓	No	Yes
Gao et al. [2024b]	W_2	Deterministic	✗	✓	✓	Yes	Yes
This chapter	W_2	Deterministic	✓	✓	✓	No	No

Table 3.3. Comparison of convergence analyses of simulation-free CNFs.

Albergo and Vanden-Eijnden [2023] derived a perturbation error bound similar to that in Lemma 3.16 for the CNF distribution estimator, under a Lipschitz assumption for the estimated velocity field. Chen et al. [2023e] conducted convergence analyses of DDIM-type samplers with the Kullback-Leibler (KL) divergence, assuming second-order smoothness in the space variable and Hölder-type regularity in the time variable for the score function, while ignoring the score estimation error. Albergo et al. [2023b] also derived a new perturbation error bound on the CNF distribution estimator using the KL divergence.

Chen et al. [2023c] provided polynomial-time convergence guarantees for distribution estimation using the probability flow ODE trained with denoising score matching and simulated with additional randomness. To derive these convergence rates, Chen et al. [2023c] assumed that the score function and the score estimator both have Lips-

chitz regularity in the space variable, and that the score estimation error is sufficiently small in the L^2 distance.

Benton et al. [2024b] studied the distribution estimation error of the flow matching method, but their results rely on the small L^2 estimation error assumption, the existence and uniqueness of smooth flows assumption, and the spatial Lipschitzness of estimated velocity field assumption. Li et al. [2024] derived convergence rates of probability flow ODEs in the total variation distance, and their results depend on a small L^2 score estimation error assumption and a small L^2 Jacobian estimation error assumption.

Cui et al. [2023] studied the problem of learning a high-dimensional mixture of two Gaussians with the flow matching method, in which the velocity field is parametrized by a two-layer auto-encoder. Furthermore, Cui et al. [2023] conducted convergence analyses of the Gaussian mixture distribution estimator in the asymptotic limit $d \to \infty$.

Cheng et al. [2023b] presented a theoretical analysis of the distribution estimator defined by Jordan-Kinderleherer-Otto (JKO) flow models, which implements the JKO scheme in a normalizing flow network. Gao et al. [2024b] assumed a small L^2 score estimation error, Lipschitz-type time regularity of the score function, and a smooth log-concave data distribution, and then studied the distribution estimation error for a general class of probability flow distribution estimators in the Wasserstein-2 distance.

Finally, Chang et al. [2024] considered a conditional generative learning model, in which the predictor X and the response Y are both random variables with bounded support. They provided an error analysis for learning the conditional distribution of $\mathsf{Y}|\mathsf{X}$ via the Föllmer flow.

In this study, we derive non-asymptotic error bounds for the estimated velocity fields and discretization error bounds for the forward Euler sampler. These error bounds are incorporated into the end-to-end convergence analysis of the CNF distribution estimator with flow matching. Furthermore, we only stipulate general assumptions on the target distribution, rather than making assumptions on the velocity field (or score function) and its estimator. We believe that these theoretical contributions set this chapter apart from previous studies.

3.6.2 Diffusion models

Diffusion models [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020, Song et al., 2021b,a] have emerged as a powerful SDE-based framework for deep generative learning. The diffusion model estimators share a deep connection with the CNF distribution estimators due to the correspondence between an Itô SDE and its probability flow ODE. There has been a growing interest in the statistical analyses of diffusion model estimators, as evidenced by the works of Lee et al. [2022], Bortoli [2022], Chen et al. [2023d], Lee et al. [2023], Chen et al. [2023a], Oko et al. [2023], Li et al. [2024], among others.

Unlike the deterministic sampler of CNF distribution estimators, diffusion model estimators employ a stochastic sampler (such as the Euler-Maruyama method) to simulate the time-reversed Itô SDEs. This stochasticity plays a crucial role in the discretization error analysis of diffusion model estimators and leads to the development of useful techniques such as Girsanov's theorem [Chen et al., 2023d], a chain rule-based variant [Chen et al., 2023a] of the interpolation technique [Vempala and Wibisono, 2019], and the stochastic interpolation formula [Bortoli, 2022]. However, it remains uncertain whether these techniques can be generalized for analyzing the CNF distribution estimators.

Compared to the CNF distribution estimators, the diffusion model estimators have been extensively investigated from a statistical perspective. For instance, the estimation error bounds of the score function have been established by Oko et al. [2023], Chen et al. [2023b], Huang et al. [2023], Cole and Lu [2024]. There is also a vast body of literature on analyzing the discretization error of diffusion model estimators, including works by Wibisono and Yang [2022], Benton et al. [2024a], Pedrotti et al. [2023], Gao et al. [2023], Bruno et al. [2023], Shah et al. [2023] and others. However, the absence of stochasticity presents significant challenges when attempting to analyze the ODE-based CNF distribution estimators using techniques developed for diffusion model estimators.

3.6.3 Neural network approximation with Lipschitz regularity control

The approximation theory of deep ReLU networks has developed rapidly since the seminal work of Yarotsky [2017]. Previous studies have shown that deep ReLU networks can efficiently approximate functions in a smooth function class, such as the Hölder class, the Sobolev class, and the Besov class, under the L^∞ norm [Yarotsky, 2017, Petersen and Voigtlaender, 2018, Suzuki, 2019, Yarotsky, 2018, Gühring et al., 2020, DeVore et al., 2021, Daubechies et al., 2022, Lu et al., 2021, Jiao et al., 2023a, Siegel, 2023]. Recent works have also considered nonparametric or semiparametric estimation using deep ReLU networks, including least squares regression [Bauer and Kohler, 2019, Schmidt-Hieber, 2020, Nakada and Imaizumi, 2020, Kohler and Langer, 2021, Suzuki and Nitanda, 2021, Chen et al., 2022, Jiao et al., 2023a], quantile regression [Shen et al., 2022a, Padilla et al., 2022] semiparametric inference [Farrell et al., 2021], factor augmented sparse throughput models [Fan and Gu, 2024], among others. In the convergence analysis of these models, it is sufficient to know the error bounds of using deep neural networks for approximating smooth functions.

Analyzing deep generative distribution estimators becomes more challenging as it requires not only approximation error bounds but also additional regularity properties of the constructed neural network approximation functions. For instance, the error analysis of Wasserstein GANs necessitates an upper bound of the Lipschitz constant of the discriminator network [Chen et al., 2020, Huang et al., 2022]. Chen et al. [2020] demonstrated that the wide and shallow ReLU network constructed by Yarotsky [2017], for which the depth grows logarithmically but the width grows polynomially, can approximate 1-Lipschitz functions with a uniformly bounded Lipschitz constant. Huang et al. [2022] provided a Lipschitz constant bound for the deep ReLU network approximation function proposed by Lu et al. [2021]. However, this bound increases with the width and depth of the network. Furthermore, Jiao et al. [2023b] succeeded in controlling the Lipschitz constant of deep ReLU networks by enforcing a norm constraint on the neural network weights, and applied the approximation bound to analyze the distribution estimation error of GANs. In addition to the error analyses of GANs, the convergence analysis of simulation-based CNFs by Marzouk et al. [2023], also requires a Lipschitz regularity control of the constructed approximation function to ensure the CNFs are

well-posed.

In the current context, the Lipschitz regularity of the neural network approximation functions is crucial for analyzing the behavior of the estimated velocity field. Indeed, a key step in our error analysis involves constructing deep ReLU networks to approximate the Lipschitz velocity field $v^*(t,x)$ for $(t,x) \in [0, 1-\underline{t}] \times \mathbb{R}^d$. To achieve this target, we need to derive an L^∞ bound of the approximation error and demonstrate that the Lipschitz constant of the constructed deep ReLU network is uniformly bounded, regardless of the varying width and depth of the neural network. Establishing the Lipschitz regularity of the neural network approximation functions, in addition to the approximation error bounds, is a more challenging task that requires different techniques. Specifically, our uniform bounds of the Lipschitz constants are sharper than those obtained by Huang et al. [2022] for varying width and depth of the deep ReLU network. Compared to the approximation bound of Chen et al. [2020], our approximation bound is valid for any network width and depth specified by the parameters N and L. Marzouk et al. [2023] considered the Lipschitz regularity of deep neural networks activated by the smooth function ReLU^k with $k \geq 2$, which is based on spline approximation and technically differs from this chapter.

3.7 Conclusion

We have established non-asymptotic error bounds for the CNF distribution estimator trained via flow matching, using the Wasserstein-2 distance. Assuming that the target distribution belongs to several rich classes of probability distributions, we have established Lipschitz regularity properties of the velocity field for simulation-free CNFs defined with linear interpolation. To meet the regularity requirements of flow matching estimators, we have developed L^∞ approximation bounds of deep ReLU networks for Lipschitz functions, along with Lipschitz regularity control of the constructed deep ReLU networks. By integrating the regularity results, the deep approximation bounds, and perturbation analyses of ODE flows, we have shown that the convergence rate of the CNF distribution estimator is $\widetilde{\mathcal{O}}(n^{-1/(d+5)})$, up to a polylogarithmic prefactor of n. Our error analysis framework can be extended to study more general CNFs based on interpolation, beyond the CNFs constructed with linear interpolation.

3.8 Proofs and supplementary results

In this section, we present proofs of the main results given in this chapter.

3.8.1 Regularity of the velocity field

In this section, we study the regularity properties of the velocity field and present necessary lemmas, theorems, propositions, and their proofs.

We first introduce several auxiliary conditions to assist studying the regularity properties of the velocity field. These conditions are covered in the two cases of Assumption 3.8.

Condition 1 (Semi-log-concavity)**.** Let $\nu(\mathrm{d}x) = \exp(-U(x))\mathrm{d}x$. The potential function $U(x)$ is of class C^2 and satisfies $\nabla^2 U(x) \succeq \kappa \mathrm{I}_d$ for some $\kappa \in \mathbb{R}$.

Condition 2 (Gaussian smoothing)**.** The target distribution $\nu = \gamma_{d,\sigma^2} * \rho$ where ρ is a probability distribution supported on a Euclidean ball of radius R on $\mathbb{R}^d$.

Lemma 3.42 (Proposition 4.1 in Gao et al. [2024a])**.** *Let $\nu(\mathrm{d}y) = p(y)\mathrm{d}y$ be a probability distribution on $\mathbb{R}^d$ with $D := (1/\sqrt{2})\mathrm{diam}(\mathrm{supp}(\nu))$.*

(1) For any $t \in (0,1)$,

$$-\frac{1}{1-t}\mathrm{I}_d \preceq \nabla_x v^*(t,x) \preceq \left\{\frac{t}{(1-t)^3}D^2 - \frac{1}{1-t}\right\}\mathrm{I}_d, \quad \mathrm{Cov}(\mathrm{X}_1|\mathrm{X}_t = x) \preceq D^2\mathrm{I}_d.$$

(2) Suppose that p is β-semi-log-convex with $\beta > 0$ and $\mathrm{supp}(p) = \mathbb{R}^d$. Then for any $t \in (0,1)$,

$$\nabla_x v^*(t,x) \succeq \frac{(\beta+1)t-\beta}{\beta(1-t)^2+t^2}\mathrm{I}_d, \quad \mathrm{Cov}(\mathrm{X}_1|\mathrm{X}_t = x) \succeq \frac{(1-t)^2}{\beta(1-t)^2+t^2}\mathrm{I}_d.$$

(3) Suppose that p is κ-semi-log-concave with $\kappa \in \mathbb{R}$. Then for any $t \in (t_0,1)$,

$$\nabla_x v^*(t,x) \preceq \frac{(\kappa+1)t-\kappa}{\kappa(1-t)^2+t^2}\mathrm{I}_d, \quad \mathrm{Cov}(\mathrm{X}_1|\mathrm{X}_t = x) \preceq \frac{(1-t)^2}{\kappa(1-t)^2+t^2}\mathrm{I}_d,$$

where t_0 is the root of the equation $\kappa + t^2/(1-t)^2 = 0$ over $t \in (0,1)$ if $\kappa < 0$ and $t_0 = 0$ if $\kappa \geq 0$.

(4) *Fix a probability distribution ρ on $\mathbb{R}^d$ supported on a Euclidean ball of radius R, and let $\nu := \gamma_{d,\sigma^2} * \rho$ with $\sigma > 0$. Then for any $t \in (0,1)$,*

$$\frac{(\sigma^2+1)t-1}{(1-t)^2+\sigma^2t^2}\mathrm{I}_d \preceq \nabla_x v^*(t,x) \preceq \left\{\frac{t(1-t)}{((1-t)^2+\sigma^2t^2)^2}R^2 + \frac{(\sigma^2+1)t-1}{(1-t)^2+\sigma^2t^2}\right\}\mathrm{I}_d,$$

$$\mathrm{Cov}(\mathsf{X}_1|\mathsf{X}_t = x) \preceq \left\{\left(\frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}\right)^2 R^2 + \frac{\sigma^2(1-t)^2}{(1-t)^2+\sigma^2t^2}\right\}\mathrm{I}_d,$$

$$(\mathsf{X}_1|\mathsf{X}_t = x) \overset{d}{=} \frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}\mathsf{Q} + \sqrt{\frac{\sigma^2(1-t)^2}{(1-t)^2+\sigma^2t^2}}\mathsf{Z} + \frac{\sigma^2t^2}{(1-t)^2+\sigma^2t^2}x$$

where $\mathsf{Q} \sim \tilde{\rho}$ is supported on the same ball as ρ, $\mathsf{Z} \sim \gamma_d$, and Q, Z are independent.

In Lemma 3.43 below, we show that the velocity field and its spatial derivative is (locally) bounded under mild regularity conditions. The boundedness of the spatial derivative directly follows from Lemma 3.42. Since a Lipschitz property results in a linear growth property, we obtain the the velocity field is locally bounded. For ease of presentation, let us define two parameter sets by

$$\mathcal{S}_1 := \begin{cases} \{\kappa, \beta\} & \text{if Assumption 3.8-(i) holds,} \\ \{R, \sigma\} & \text{if Assumption 3.8-(ii) holds,} \end{cases}$$

$$\mathcal{S}_2 := \begin{cases} \{d, \kappa, \beta\} & \text{if Assumption 3.8-(i) holds,} \\ \{d, R, \sigma\} & \text{if Assumption 3.8-(ii) holds.} \end{cases}$$

We say a prefactor scales polynomially with $\mathcal{S}_1$ if it scales polynomially with parameters in $\mathcal{S}_1$.

Lemma 3.43. *Suppose that Assumptions 3.7 and 3.8 hold. Then it holds that*

$$\sup_{(t,x)\in[0,1]\times\Omega_A} \|v^*(t,x)\|_2 \lesssim A, \qquad \sup_{(t,x)\in[0,1]\times\mathbb{R}^d} \|\nabla_x v^*(t,x)\|_{2,2} \lesssim 1,$$

where we omit prefactors scaling polynomially with $\mathcal{S}_2$.

Proof. Under Assumptions 3.7 and 3.8, Lemma 3.42 shows that

$$C_1(\mathcal{S}_1)\mathrm{I}_d \preceq \nabla_x v^*(t,x) \preceq C_2(\mathcal{S}_1)\mathrm{I}_d,$$

where $C_1(\mathcal{S}_1)$ and $C_2(\mathcal{S}_1)$ are constants scaling polynomially with $\mathcal{S}_1$. It further yields that

$$\sup_{(t,x)\in[0,1]\times\mathbb{R}^d} \|\nabla_x v^*(t,x)\|_{2,2} \lesssim 1, \tag{3.21}$$

when we omit a prefactor scaling polynomially with $\mathcal{S}_1$. Notice that for any $t \in (0,1)$, it holds that

$$\begin{aligned} v^*(t,0) &= \frac{1}{1-t}\mathbb{E}[\mathsf{X}_1|\mathsf{X}_t = 0] = \frac{1}{1-t}\int_{\mathbb{R}^d} yq(y|t,0)\mathrm{d}y \\ &\lesssim \frac{1}{1-t}\int_{\mathbb{R}^d} yp(y)(1-t)^{-d}\exp\left(-\frac{t^2\|y\|_2^2}{2(1-t)^2}\right)\mathrm{d}y, \end{aligned}$$

which implies $\|v^*(t,0)\|_2 < \infty$ due to fast growth of the exponential function. Besides, $v^*(0,0) = \mathbb{E}[\mathsf{X}_1], v^*(1,0) = 0$. Then by the boundedness of $\|v^*(t,0)\|_2$ over $[0,1]$ and (3.21), we bound $v^*(t,x)$ as follows

$$\begin{aligned} \|v^*(t,x)\|_2 &\le \|v^*(t,0)\|_2 + \|v^*(t,x) - v^*(t,0)\|_2 \\ &\le \|v^*(t,0)\|_2 + \left\{\sup_{(t,y)\in[0,1]\times\mathbb{R}^d} \|\nabla_y v^*(t,y)\|_{2,2}\right\}\|x\|_2 \\ &\lesssim \|x\|_2 \vee 1, \end{aligned}$$

where we omit a prefactor scaling polynomially with $\mathcal{S}_1$. It further yields that

$$\sup_{(t,x)\in[0,1]\times\Omega_A} \|v^*(t,x)\|_2 \lesssim A$$

by omitting a prefactor scaling polynomially with $\mathcal{S}_2$. This completes the proof. □

Control with semi-log-concavity

We derive moment bounds under Condition 1. The moment bounds are useful to estimate the time regularity of the velocity field.

Lemma 3.44 (Moment bounds)**.** *Suppose that Condition 1 holds. Let $\eta \in (0,1)$ be a constant. Let t_1 be the root of the equation $\kappa(1-t)^2 + t^2 = \eta$ over $t \in (0,1)$ if $\kappa \le 0$ or $t_1 = 0$ if $\kappa > 0$. Then for any $t \in [t_1, 1-\underline{t}]$, it holds that*

$$\sup_{x\in\Omega_A}\|M_1\|_2 \lesssim A, \quad \sup_{x\in\mathbb{R}^d}\|M_2^c\|_{2,2} \lesssim (1-t)^2, \quad \sup_{x\in\Omega_A}\|M_3 - M_2M_1\|_2 \lesssim A(1-t)^2,$$

where we omit polynomial prefactors in d, κ, η.

Proof. First, we bound M_2^c. According to Lemma 3.42, the following covariance bound holds for any $t \in [t_1, 1)$

$$0\mathrm{I}_d \preceq \mathrm{Cov}(\mathsf{X}_1|\mathsf{X}_t = x) \preceq \frac{(1-t)^2}{\kappa(1-t)^2 + t^2}\mathrm{I}_d \quad \text{with } \kappa(1-t)^2 + t^2 \geq \begin{cases} C_\kappa, & \text{if } \kappa > 0, \\ \eta, & \text{if } \kappa \leq 0, \end{cases}$$

where $C_\kappa := \kappa/(\kappa+1)$. Then it implies that for any $t \in [t_1, 1-\underline{t}]$, $\sup_{x\in\mathbb{R}^d} \|M_2^c\|_{2,2} \lesssim (1-t)^2$ with omitting a polynomial prefactor in κ, η.

Then, we bound M_1 and M_2. By the Hatsell-Nolte identity [Dytso et al., 2023b, Proposition 1], we obtain $\nabla_x M_1(t,x) = (t/(1-t)^2)M_2^c$ which implies that

$$\sup_{(t,x)\in[t_1,1-\underline{t}]\times\mathbb{R}^d} \|\nabla_x M_1(t,x)\|_{2,2} = \sup_{(t,x)\in[t_1,1-\underline{t}]\times\mathbb{R}^d} \frac{t}{(1-t)^2}\|M_2^c\|_{2,2} \lesssim 1 \tag{3.22}$$

with a polynomial prefactor in κ, η hidden. Notice that for any $t \in (0,1)$, it holds that

$$M_1(t,0) = \int_{\mathbb{R}^d} yq(y|t,0)\mathrm{d}y \lesssim \int_{\mathbb{R}^d} yp(y)(1-t)^{-d}\exp\left(-\frac{t^2\|y\|_2^2}{2(1-t)^2}\right)\mathrm{d}y,$$

which implies $\|M_1(t,0)\|_2 < \infty$ due to fast growth of the exponential function. Besides, $M_1(0,0) = \mathbb{E}_p[\mathsf{X}_1]$ and $M_1(1,0) = 0$. By the boundedness of $\|M_1(t,0)\|_2$ for any $t \in [0,1]$ and (3.22), we further bound $M_1(t,x)$ for any $(t,x) \in [t_1, 1-\underline{t}] \times \mathbb{R}^d$ as follows

$$\begin{aligned} \|M_1(t,x)\|_2 &\leq \|M_1(t,0)\|_2 + \|M_1(t,x) - M_1(t,0)\|_2 \\ &\leq \|M_1(t,0)\|_2 + \left\{\sup_{(t,y)\in[t_1,1-\underline{t}]\times\mathbb{R}^d} \|\nabla_y M_1(t,y)\|_{2,2}\right\}\|x\|_2 \\ &\lesssim \|x\|_2 \vee 1, \end{aligned}$$

where a polynomial prefactor in κ, η is hidden. It further yields that

$$\sup_{(t,x)\in[t_1,1-\underline{t}]\times\Omega_A} \|M_1\|_2 \lesssim A$$

when omitting a polynomial prefactor in d, κ, η. Moreover, notice that $M_2 = \mathrm{Tr}(M_2^c) + \|M_1\|_2^2$, which further yields that

$$\sup_{(t,x)\in[t_1,1-\underline{t}]\times\Omega_A} |M_2| \lesssim A^2$$

with an omitted polynomial prefactor in d, κ, η.

Lastly, we bound $M_3 - M_2 M_1$. For any $i \in \{1, 2, \cdots, d\}$, let $\mathsf{X}_{1,i}$ denote the i-th element of X_1. Then it holds that

$$
\begin{aligned}
&\|M_3 - M_2 M_1\|_2^2 \\
&= \sum_{i=1}^{d} \Big(\mathbb{E}[\mathsf{X}_{1,i} \mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x] - \mathbb{E}[\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x] \mathbb{E}[\mathsf{X}_{1,i} | \mathsf{X}_t = x] \Big)^2 \\
&= \sum_{i=1}^{d} \Big(\mathrm{Cov}(\mathsf{X}_1^\top \mathsf{X}_1, \mathsf{X}_{1,i} | \mathsf{X}_t = x) \Big)^2 \\
&\le \sum_{i=1}^{d} \mathrm{Var}(\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x) \mathrm{Var}(\mathsf{X}_{1,i} | \mathsf{X}_t = x) \\
&\quad \text{(By the Cauchy-Schwarz inequality)} \\
&= \mathrm{Var}(\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x) \sum_{i=1}^{d} \mathrm{Var}(\mathsf{X}_{1,i} | \mathsf{X}_t = x) \\
&= \mathrm{Var}(\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x) \, \mathrm{Tr}(M_2^c) \\
&\le d \, \mathrm{Var}(\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x) \|M_2^c\|_{2,2}.
\end{aligned}
$$

Let $\mathsf{X}_1 \sim p(y)$ be κ-semi-log-concave for some $\kappa \in \mathbb{R}$. Then for any $t \in [0, 1)$, $\mathsf{X}_1 | \mathsf{X}_t \sim q(y|t, x)$ is $(\kappa + t^2/(1-t)^2)$-semi-log-concave because

$$
-\nabla_y^2 \log q(y|t,x) = -\nabla_y^2 \log p(y) - \nabla_y^2 \log q(t, x|y) \succeq \left(\kappa + \frac{t^2}{(1-t)^2} \right) \mathrm{I}_d.
$$

When $t \in \Big\{ t : \kappa + t^2/(1-t)^2 > 0, t \in (0,1) \Big\}$, by the Brascamp-Lieb inequality [Brascamp and Lieb, 1976], it yields that

$$
\mathrm{Var}(\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x) \le 4 M_2 \left(\kappa + \frac{t^2}{(1-t)^2} \right)^{-1} = 4 M_2 \frac{(1-t)^2}{\kappa (1-t)^2 + t^2}.
$$

Analogous to the control of $\|M_2^c\|_{2,2}$, we further obtain that for any $(t, x) \in [t_1, 1 - \underline{t}] \times \Omega_A$,

$$
\mathrm{Var}(\mathsf{X}_1^\top \mathsf{X}_1 | \mathsf{X}_t = x) \lesssim \begin{cases} A^2 (1-t)^2 / C_\kappa, & \text{if } \kappa > 0, \\ A^2 (1-t)^2 / \eta, & \text{if } \kappa \le 0. \end{cases}
$$

Hence, we deduce that for any $t \in [t_1, 1 - \underline{t}]$,

$$
\sup_{x \in \Omega_A} \|M_3 - M_2 M_1\|_2 \lesssim A (1-t)^2,
$$

where we omit a polynomial prefactor in d, κ, η. This completes the proof. □

Lemma 3.45. *Suppose that Condition 1 holds. Then it holds that*

$$\sup_{(t,x)\in[t_1,1-\underline{t}]\times\Omega_A} \|\partial_t v^*(t,x)\|_2 \lesssim A/\underline{t}^2,$$

where we omit a polynomial prefactor in d, κ, η.

Proof. By Lemma 3.17, it holds that

$$\begin{aligned}\|\partial_t v^*(t,x)\|_2 \leq & \frac{1}{(1-t)^2}\|x\|_2 + \frac{1}{(1-t)^2}\|M_1\|_2 + \frac{1}{(1-t)^4}\|M_2^c\|_{2,2}\cdot\|x\|_2 \\ & + \frac{1}{(1-t)^4}\|M_3 - M_2M_1\|_2.\end{aligned}$$

Applying Lemma 3.44, we obtain

$$\sup_{(t,x)\in[t_1,1-\underline{t}]\times\Omega_A} \|\partial_t v^*(t,x)\|_2 \lesssim \frac{A}{\underline{t}^2},$$

where we omit a polynomial prefactor in d, κ, η. □

Control with Gaussian smoothing

We derive moment bounds under Condition 2. The moment bounds are useful to estimate the time regularity of the velocity field.

Lemma 3.46 (Moment bounds)**.** *Suppose that Condition 2 holds. Then for any $t \in [0, 1-\underline{t}]$, it holds that*

$$\sup_{x\in\Omega_A}\|M_1\|_2 \lesssim A, \quad \sup_{x\in\mathbb{R}^d}\|M_2^c\|_{2,2} \lesssim (1-t)^2, \quad \sup_{x\in\Omega_A}\|M_3 - M_2M_1\|_2 \lesssim A(1-t)^2,$$

where we omit polynomial prefactors in d, R, σ.

Proof. The proof idea is partially similar to that of Lemma 3.44.

First, we bound M_2^c. According to Lemma 3.42, the following covariance bound holds for any $t \in [0, 1)$,

$$0\mathrm{I}_d \preceq \mathrm{Cov}(\mathsf{X}_1|\mathsf{X}_t = x) \preceq (1-t)^2\left\{\frac{R^2(1-t)^2}{((1-t)^2+\sigma^2t^2)^2} + \frac{\sigma^2}{(1-t)^2+\sigma^2t^2}\right\}\mathrm{I}_d.$$

Notice that

$$\frac{R^2(1-t)^2}{((1-t)^2+\sigma^2t^2)^2}+\frac{\sigma^2}{(1-t)^2+\sigma^2t^2}\le\left(1+\frac{1}{\sigma^2}\right)^2R^2+\sigma^2+1.$$

It implies that for any $t\in[0,1-\underline{t}]$,

$$\sup_{x\in\mathbb{R}^d}\|M_2^c\|_{2,2}\lesssim(1-t)^2 \tag{3.23}$$

with omitting a polynomial prefactor in R,σ.

Then, we bound M_1 and M_2. Again, by the Hatsell-Nolte identity [Dytso et al., 2023b, Proposition 1], we obtain that

$$\sup_{(t,x)\in[0,1-\underline{t}]\times\mathbb{R}^d}\|\nabla_xM_1(t,x)\|_{2,2}=\sup_{(t,x)\in[0,1-\underline{t}]\times\mathbb{R}^d}\frac{t}{(1-t)^2}\|M_2^c\|_{2,2}\lesssim1 \tag{3.24}$$

with a polynomial prefactor in R,σ hidden. Identical to how we proceed in the proof of Lemma 3.44, we have

$$\sup_{(t,x)\in\Omega_{\underline{t},A}}\|M_1\|_2\lesssim A$$

when omitting a polynomial prefactor in d,R,σ.

Finally, we bound $M_3-M_2M_1$. Recall that we have deduced the following inequality in the proof of Lemma 3.44

$$\|M_3-M_2M_1\|_2^2\le d\ \mathrm{Var}(\mathsf{X}_1^\top\mathsf{X}_1|\mathsf{X}_t=x)\|M_2^c\|_{2,2}. \tag{3.25}$$

We next focus on bounding $\mathrm{Var}(\mathsf{X}_1^\top\mathsf{X}_1|\mathsf{X}_t=x)$. By Lemma 3.42-(4), it is shown that

$$(\mathsf{X}_1|\mathsf{X}_t=x)\overset{d}{=}\mathsf{P}_x:=\frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}\mathsf{Q}+\sqrt{\frac{\sigma^2(1-t)^2}{(1-t)^2+\sigma^2t^2}}\mathsf{Z}+\frac{\sigma^2t^2}{(1-t)^2+\sigma^2t^2}x$$

where $\mathsf{Q}\sim\tilde{\rho}$ is supported on the same ball as ρ, $\mathsf{Z}\sim\gamma_d$, and Q,Z are independent. In the expression above, we note that the denominator $(1-t)^2+\sigma^2t^2$ is lower bounded by $\sigma^2/(\sigma^2+1)$ over $t\in[0,1]$. Let $\mathsf{R}_{x,y}:=\mathsf{P}_x^\top\mathsf{P}_x|\mathsf{Q}=y$. Then by the law of total variance, it yields that

$$\mathrm{Var}(\mathsf{X}_1^\top\mathsf{X}_1|\mathsf{X}_t=x)=\mathrm{Var}(\mathsf{P}_x^\top\mathsf{P}_x)=\mathbb{E}[\mathrm{Var}(\mathsf{R}_{x,y})]+\mathrm{Var}(\mathbb{E}[\mathsf{R}_{x,y}]). \tag{3.26}$$

We claim that $\mathsf{R}_{x,y}/\eta$ is distributed as a noncentral chi-squared distribution with degrees of freedom d and the noncentrality parameter $\xi_{x,y}$ where

$$\eta = \frac{\sigma^2(1-t)^2}{(1-t)^2+\sigma^2t^2}, \quad \xi_{x,y} = \frac{1}{\eta}\left\|\frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}y + \frac{\sigma^2t^2}{(1-t)^2+\sigma^2t^2}x\right\|_2^2.$$

By properties of the noncentral chi-squared distribution, it holds that

$$\mathbb{E}(\mathsf{R}_{x,y}) = \eta(d+\xi_{x,y}), \quad \mathrm{Var}(\mathsf{R}_{x,y}) = 2\eta^2(d+2\xi_{x,y}).$$

Then we bound the first term in the variance decomposition (3.26) as follows

$$\mathbb{E}[\mathrm{Var}(\mathsf{R}_{x,y})] = 2\eta\mathbb{E}\left[\eta d + 2\eta\xi_{x,y})\right] \lesssim (1-t)^2(\|x\|_2^2 \vee 1)$$

where we omit a polynomial prefactor in d, R, σ. To bound the second term, we do the following calculations

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[\mathsf{R}_{x,y}]) = \mathrm{Var}(\eta d + \eta\xi_{x,y}) &= \mathrm{Var}\left(\eta d + \left\|\frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}\mathsf{Q} + \frac{\sigma^2t^2}{(1-t)^2+\sigma^2t^2}x\right\|_2^2\right)\\ &= \mathrm{Var}\left(\left\|\frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}\mathsf{Q}\right\|_2^2 + 2\left\langle\frac{(1-t)^2}{(1-t)^2+\sigma^2t^2}\mathsf{Q}, \frac{\sigma^2t^2}{(1-t)^2+\sigma^2t^2}x\right\rangle\right)\\ &\lesssim (1-t)^4(\|x\|_2^2 \vee 1),\end{aligned}$$

where we omit a polynomial prefactor in d, R, σ. Combining the control of the two terms, we obtain that

$$\mathrm{Var}(\mathsf{X}_1^\top\mathsf{X}_1|\mathsf{X}_t = x) \lesssim (1-t)^2(\|x\|_2^2 \vee 1) \tag{3.27}$$

by omitting a polynomial prefactor in d, R, σ. Therefore, using (3.23), (3.25), and (3.27), the bound of $M_3 - M_2M_1$ is deduced for any $t \in [0, 1-\underline{t}]$ by

$$\sup_{x\in\Omega_A}\|M_3 - M_2M_1\|_2 \lesssim A(1-t)^2$$

with a polynomial prefactor in d, R, σ hidden. This completes the proof. □

Lemma 3.47. *Suppose that Condition 2 holds. Then it holds that*

$$\sup_{(t,x)\in[0,1-\underline{t}]\times\Omega_A}\|\partial_t v^*(t,x)\|_2 \lesssim A/\underline{t}^2,$$

where we omit a polynomial prefactor in d, κ, η.

Proof. Based on Lemma 3.46, the proof is almost identical to that of Lemma 3.45. □

Sharpness of moment bounds

The moment bounds in Lemmas 3.44 and 3.46 are sharp in (t,x) because of a Gaussian example.

Proposition 3.48. *Let $\mathsf{X}_1 \sim \gamma_d$. The conditional distribution of $\mathsf{X}_1|\mathsf{X}_t$ has the following explicit expression*

$$\mathsf{X}_1|\mathsf{X}_t = x \sim N\left(\frac{t}{(1-t)^2+t^2}x, \frac{(1-t)^2}{(1-t)^2+t^2}\mathrm{I}_d\right).$$

Moreover, for any $t \in (0,1]$, the moment bounds are given by

$$\sup_{x\in\Omega_A} \|M_1\|_2 \asymp A, \ \sup_{x\in\mathbb{R}^d} \|M_2^c\|_{2,2} \asymp (1-t)^2, \ \sup_{x\in\Omega_A} \|M_3 - M_2M_1\|_2 \asymp A(1-t)^2,$$

where we omit polynomial prefactors in d.

Proof. By Bayes' rule, for $\mathsf{X}_1 \sim \gamma_d$, it implies that

$$\mathsf{X}_1|\mathsf{X}_t = x \sim N\left(\frac{t}{(1-t)^2+t^2}x, \frac{(1-t)^2}{(1-t)^2+t^2}\mathrm{I}_d\right).$$

By properties of the Gaussian distribution, the desired moment bounds hold. □

Proof of Theorem 3.26

Proof of Theorem 3.26. By Lemma 3.43, it holds that for any $x,y \in \mathbb{R}^d$ and $t \in [0,1]$, $\|v^*(t,x) - v^*(t,y)\|_\infty \lesssim \|x-y\|_\infty$, and that $\sup_{(t,x)\in[0,1]\times\Omega_A} \|v^*(t,x)\|_\infty \lesssim A$, where we omit constants in $d, \kappa, \beta, \sigma, R$.

Then we show that the Lipschitz continuity of $v^*(t,x)$ in t. Concretely, for any $s,t \in [0, 1-\underline{t}]$ and $x \in \Omega_A$, $\|v^*(t,x) - v^*(s,x)\|_\infty \le L_t|t-s|$ with $L_t \lesssim A\underline{t}^{-2}$ by omitting a constant in $d, \kappa, \beta, \sigma, R$. We analyze the cases in Assumption 3.8 one by one as follows:

- Suppose that Assumptions 3.7 and 3.8-(i) holds. Condition 1 holds as well. We use controls with semi-log-concavity in Lemma 3.45 and derive the desired Lipschitz continuity in t.
- Suppose that Assumptions 3.7 and 3.8-(ii) holds. Condition 2 holds as well. We use controls with with Gaussian smoothing in Lemma 3.47, and the Lipschitz continuity in t follows.

We complete the proof. □

3.8.2 Approximation error of the velocity field

In this section, we analyze the approximation error of the velocity field by a constructive approach.

Before proceeding, we present a few useful notations for the Sobolev function class. A d-dimensional multi-index is a d-tuple $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)^\top \in \mathbb{N}_0^d$. We define $\|\alpha\|_1 = \sum_{i=1}^d \alpha_i$ and $\partial^\alpha := \partial_1^{\alpha_1}\partial_2^{\alpha_2}\cdots\partial_d^{\alpha_d}$ to represent the partial derivative of a d-dimensional function. We also use D to denote the weak derivative of a single variable function and D^α to denote the partial derivative $D_1^{\alpha_1}D_2^{\alpha_2}\cdots D_d^{\alpha_d}$ of a d-dimensional function with α_i as the order of derivative D_i in the i-th variable.

Approximation in space with Lipschitz regularity

In this subsection, we study the approximation capacity of deep ReLU networks joint with an estimate of the Lipschitz regularity. The strong expressive power of deep ReLU networks has been studied with the localized or averaged Taylor polynomials. We follow the localized approximation approach, and establish the global Lipschitz continuity and non-asymptotic approximation estimate of deep ReLU networks.

Lemma 3.49. *Given any $f \in W^{1,\infty}((0,1)^d)$ with $\|f\|_{W^{1,\infty}((0,1)^d)} \le 1$, for any $N, L \in \mathbb{N}$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(2^d dN\log N)$ and depth $\mathcal{O}(d^2 L\log L)$ such that $\|\phi\|_{W^{1,\infty}((0,1)^d)} \lesssim 1$ and*

$$\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim (NL)^{-2/d},$$

where we omit some prefactors depending only on d.

Corollary 3.50. *Given any $f \in W^{1,\infty}((0,1)^d)$ with $\|f\|_{W^{1,\infty}((0,1)^d)} < \infty$, for any $N, L \in \mathbb{N}$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(2^d dN\log N)$ and depth $\mathcal{O}(d^2 L\log L)$ such that $\|\phi\|_{W^{1,\infty}((0,1)^d)} \lesssim \|f\|_{W^{1,\infty}((0,1)^d)}$ and*

$$\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim \|f\|_{W^{1,\infty}((0,1)^d)}(NL)^{-2/d},$$

where we omit some prefactors depending only on d.

Remark 3.51. The approximation rate is nearly optimal for the unit ball of functions in $W^{1,\infty}((0,1)^d)$ according to Shen et al. [2020, 2022b] and Lu et al. [2021].

Proof sketch of Lemma 3.49. The proof idea is similar to that of Yang et al. [2023, Theorem 3], and we divide the proof into three steps.

Step 1. Discretization. We use a partition of unity to discretize the set $(0,1)^d$. As in Definitions 3.52 and 3.53, we construct a partition of unity $\{g_m\}_{m\in\{1,2\}^d}$ on $(0,1)^d$ with $\operatorname{supp}(g_m)\cap(0,1)^d \subset \Omega_m$ for any $m \in \{1,2\}^d$. Then we approximate the partition of unity $\{g_m\}_{m\in\{1,2\}^d}$ by a collection of deep ReLU networks $\{\phi_m\}_{m\in\{1,2\}^d}$ as in Lemma 3.54.

Step 2. Approximation on Ω_m. Given any $m \in \{1,2\}^d$, for each subset $\Omega_m \subset [0,1]^d$, we find a piecewise constant function $f_{K,m}$ satisfying

$$\|f_{K,m} - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|f_{K,m} - f\|_{L^\infty(\Omega_m)} \lesssim 1/K,$$

where we omit constants in d. Piecewise constant functions can be approximated by deep ReLU networks. Then, following Lu et al. [2021] and Yang et al. [2023], we construct a deep ReLU network ψ_m with width $\mathcal{O}(2^d dN\log N)$ and depth $\mathcal{O}(d^2 L\log L)$ such that

$$\|\psi_m - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim (NL)^{-2/d},$$

where we omit constants in d.

Step 3. Approximation on $[0,1]^d$. Combining the approximations on each subset Ω_m properly, we construct an approximation of the target function f on the domain $[0,1]^d$. That is, for any $N, L \in \mathbb{N}$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$ such that

$$\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim (NL)^{-2/d} \quad \text{with} \quad \|\phi\|_{W^{1,\infty}((0,1)^d)} \lesssim 1,$$

where we omit constants in d.

Definition 3.52. Given $K, d \in \mathbb{N}$, and for any $m = [m_1, m_2, \cdots, m_d]^\top \in \{1,2\}^d$, we define $\Omega_m := \prod_{i=1}^d \Omega_{m_j}$ where $\Omega_1 := \bigcup_{i=1}^{K-1}\left[\frac{i}{K}, \frac{i}{K}+\frac{3}{4K}\right]$ and $\Omega_2 := \bigcup_{i=0}^{K}\left[\frac{i}{K}-\frac{1}{2K}, \frac{i}{K}+\frac{1}{4K}\right]\cap [0,1]$.

Definition 3.53. Given $K, d \in \mathbb{N}$, for any integer $i \in \mathbb{Z}$, we define

$$g_1(x) := \begin{cases} 1, & x \in \left[\frac{i}{K}+\frac{1}{4K}, \frac{i}{K}+\frac{1}{2K}\right], \\ 0, & x \in \left[\frac{i}{K}+\frac{3}{4K}, \frac{i}{K}+\frac{1}{K}\right], \\ 4K\left(x-\frac{i}{K}\right), & x \in \left[\frac{i}{K}, \frac{i}{K}+\frac{1}{4K}\right], \\ -4K\left(x-\frac{i}{K}-\frac{3}{4K}\right), & x \in \left[\frac{i}{K}+\frac{1}{2K}, \frac{i}{K}+\frac{3}{4K}\right], \end{cases} \qquad g_2(x) := g_1\left(x+\frac{1}{2K}\right).$$

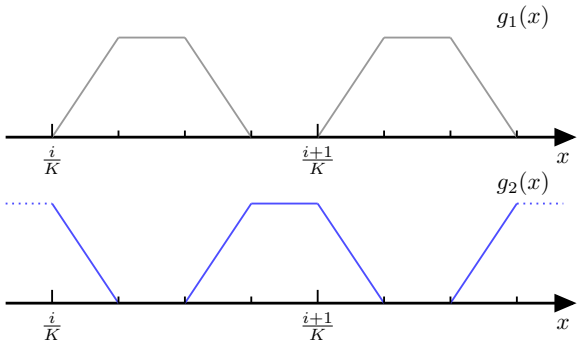

Figure 3.1. Functions g_1 and g_2 for defining a partition of unity.

For any $m = [m_1, m_2, \cdots, m_d]^\top \in \{1,2\}^d$, we further define $g_m(x) := \prod_{j=1}^d g_{m_j}(x_j)$ where $x = [x_1, x_2, \cdots, x_d]^\top$.

Lemma 3.54 (Proposition 1 in Yang et al. [2023]). *Given any $N, L \in \mathbb{N}$ and any $m \in \{1,2\}^d$, for $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, there exists a function ϕ_m implemented by a deep ReLU network with width $\mathcal{O}(dN)$ and depth $\mathcal{O}(d^2 L)$ such that*

$$\|\phi_m - g_m\|_{W^{1,\infty}((0,1)^d)} \leq 50 d^{5/2} (N+1)^{-4dL}.$$

Lemma 3.55. *Let $K \in \mathbb{N}$. For any $f \in W^{1,\infty}((0,1)^d)$ with $\|f\|_{W^{1,\infty}((0,1)^d)} \leq 1$ and $m \in \{1,2\}^d$, there exists a piecewise constant function $f_{K,m}$ on Ω_m satisfying*

$$\|f_{K,m} - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|f_{K,m} - f\|_{L^\infty(\Omega_m)} \lesssim 1/K$$

with prefactors in d omitted.

Proof. We leverage approximation properties of averaged Taylor polynomials [Brenner and Scott, 2008, Definition 4.1.3] and the Bramble-Hilbert Lemma [Brenner and Scott, 2008, Lemma 4.3.8] to deduce local estimates and then combine them through a partition of unity to obtain a global estimate. The key observation is that the L^∞ approximation bound can be established while uniformly controlling the Lipschitz constant of the piecewise constant function with a mild regularity assumption on the target function such as $f \in W^{1,\infty}((0,1)^d)$.

Without loss of generality, let us assume $m = m_* := [1, 1, \cdots, 1]^\top$. Following the proofs of Gühring et al. [2020, Lemma C.4] and Yang et al. [2023, Theorem 6], we first define an extension operator $E : W^{1,\infty}((0,1)^d) \to W^{1,\infty}(\mathbb{R}^d)$ to handle the boundary. Accordingly, let $\tilde{f} := Ef$ and C_E be the norm of the extension operator. Then for any

$\Omega \subset \mathbb{R}^d$, it holds that

$$|\tilde{f}|_{W^{1,\infty}(\Omega)} \le \|\tilde{f}\|_{W^{1,\infty}(\Omega)} \le C_E \|f\|_{W^{1,\infty}((0,1)^d)} \le C_E.$$

Next, we define an average Taylor polynomial of order 1 over $B_{i,K} := \mathbb{B}^d(\frac{8i+3}{8K}, \frac{1}{4K}, \|\cdot\|_2)$ by

$$p_{f,i}(x) := \int_{B_{i,K}} T_y^1 \tilde{f}(x) \phi_K(y) \mathrm{d}y$$

where ϕ_K is a cut-off function supported on $\bar{B}_{i,K}$ as given in Example 3.67. By Definition 3.66, $T_y^1 \tilde{f}(x) = \tilde{f}(y)$. Then, it implies that $p_{f,i}(x)$ is a constant function as

$$p_{f,i}(x) = \int_{B_{i,K}} \tilde{f}(y) \phi_K(y) \mathrm{d}y.$$

Step 1. Get local estimates. For any $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K\}^d$, we would like to employ the Bramble-Hilbert lemma 3.72 on the subset

$$\Omega_{m_*,i} = \bar{\mathbb{B}}^d\left(\frac{8i+3}{8K}, \frac{3}{8K}, \|\cdot\|_\infty\right) = \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right].$$

It is easy to check the conditions of the Bramble-Hilbert lemma are fulfilled as

$$\frac{1}{4K} \ge \frac{1}{2} \times \frac{3}{8K} = \frac{1}{2} r_{\max}(\Omega_{m_*,i}), \quad \gamma(\Omega_{m_*,i}) = \frac{d_{\Omega_{m_*,i}}}{r_{\max}(\Omega_{m_*,i})} = 2\sqrt{d}.$$

Hence, by the Bramble-Hilbert Lemma 3.72, it yields that

$$\|\tilde{f} - p_{f,i}\|_{L^\infty(\Omega_{m_*,i})} \le C_1(d) |\tilde{f}|_{W^{1,\infty}(\Omega_{m_*,i})}/K,$$
$$|\tilde{f} - p_{f,i}|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(d) |\tilde{f}|_{W^{1,\infty}(\Omega_{m_*,i})}.$$

Combining $|\tilde{f}|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_E$ and the inequalities above, it implies that

$$\|\tilde{f} - p_{f,i}\|_{L^\infty(\Omega_{m_*,i})} \le C_1(d) C_E/K, \tag{3.28}$$
$$\|\tilde{f} - p_{f,i}\|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(d) C_E. \tag{3.29}$$

Step 2. Define a partition of unity. We construct a partition of unity in order to combine the local estimates. Let $K \in \mathbb{N}$. For any $0 \le i \le K$, we define $h_i : \mathbb{R} \to \mathbb{R}$ by

$$h_i(x) = h\left(4K\left(x - \frac{8i+3}{8K}\right)\right) \text{ where } h(x) = \begin{cases} 1, & |x| < 3/2, \\ 0, & |x| > 2, \\ 4 - 2|x|, & 3/2 \le |x| \le 2. \end{cases}$$

One can verify that $\{h_i\}_{i=1}^K$ is a partition of unity of $[0,1]$ and $h_i(x) = 1$ for any $x \in \left[\frac{i}{K}, \frac{3+4i}{4K}\right]$. Considering the multidimensional case, for any $x = [x_1, x_2, \cdots, x_d]^\top \in \mathbb{R}^d$ and any $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K\}^d$, let us define

$$h_i(x) := \prod_{j=1}^d h_{i_j}(x_j).$$

Then a partition of unity of $[0,1]^d$ is defined by $\{h_i : i \in \{0, 1, \cdots, K\}^d\}$. Moreover, $h_i(x) = 1$ for any $x \in \Omega_{m_*,i} = \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right]$ and $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K\}^d$. By the definition of $h_i(x)$ on $\Omega_{m_*,i}$ and equation 3.28, equation 3.29 , it yields that

$$\begin{aligned}
\|h_i(\tilde{f} - p_{f,i})\|_{L^\infty(\Omega_{m_*,i})} &\le \|\tilde{f} - p_{f,i}\|_{L^\infty(\Omega_{m_*,i})} \le C_1(d) C_E / K, \\
\|h_i(\tilde{f} - p_{f,i})\|_{W^{1,\infty}(\Omega_{m_*,i})} &\le \|\tilde{f} - p_{f,i}\|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(d) C_E.
\end{aligned}$$

Step 3. Get global estimates. To deduce the global estimates, we start with defining f_{K,m_*} over Ω_{m_*} by

$$f_{K,m_*} := \sum\nolimits_{i \in \{0,1,\cdots,K\}^d} h_i p_{f,i}.$$

The error bounds follow that

$$\begin{aligned}
\|f_{K,m_*} - f\|_{L^\infty(\Omega_{m_*})} &\le \max_{i \in \{0,1,\cdots,K\}^d} \|h_i(\tilde{f} - p_{f,i})\|_{L^\infty(\Omega_{m_*,i})} \le C_1(d) C_E / K, \\
\|f_{K,m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} &\le \max_{i \in \{0,1,\cdots,K\}^d} \|h_i(\tilde{f} - p_{f,i})\|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(d) C_E.
\end{aligned}$$

This completes the proof. □

Lemma 3.56. *Given any $f \in W^{1,\infty}((0,1)^d)$ with $\|f\|_{W^{1,\infty}((0,1)^d)} \le 1$, for any $N, L \in \mathbb{N}$ and any $m \in \{1,2\}^d$, there exists a deep ReLU network ψ_m with width $\mathcal{O}(N \log N)$ and depth $\mathcal{O}(L \log L)$ such that*

$$\|\psi_m - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim (NL)^{-2/d},$$

where we omit constants in d.

Proof. The idea of proof is similar to those of Hon and Yang [2022, Theorem 3.1] and Yang et al. [2023, Theorem 7]. For completeness, we provide a concrete proof in the

following. Without loss of generality, we consider $m = m_* := [1, 1, \cdots, 1]^\top$. Given $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, by Lemma 3.55, we have

$$\|f_{K,m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \lesssim 1,$$
$$\|f_{K,m_*} - f\|_{L^\infty(\Omega_{m_*})} \lesssim 1/K \lesssim (NL)^{-2/d},$$

where f_{K,m_*} is a constant function for $x \in \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right]$ and $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K-1\}^d$. The insight is to approximate f_{K,m_*} with deep ReLU networks. Let $\delta = 1/(4K) \le 1/(3K)$ in Lemma 3.81. Then by Lemma 3.81, there exists a deep ReLU network $\phi_1(x)$ with width $4N + 5$ and depth $4L + 4$ such that

$$\phi_1(x) = k, \quad x \in \left[\frac{k}{K}, \frac{k+1}{K} - \frac{1}{4K}\right], \quad k = 0, 1, \cdots, K-1.$$

We further define

$$\phi_2(x) = \left[\frac{\phi_1(x_1)}{K}, \frac{\phi_1(x_2)}{K}, \cdots, \frac{\phi_1(x_d)}{K}\right]^\top.$$

For each $p = 0, 1, \cdots, K^d - 1$, there exists a bijection

$$\eta(p) = [\eta_1, \eta_2, \cdots, \eta_d]^\top \in \{0, 1, \cdots, K-1\}^d$$

satisfying $\sum_{j=1}^d \eta_j K^{j-1} = p$. We also define

$$\xi_p = \frac{f_{K,m_*}(\eta(p)/K) + C_2(d)}{2C_2(d)} \in [0, 1],$$

where $|f_{K,m_*}| < C_2(d) := 1 + C_1(d)C_E$. Then, due to Lemma 3.82, there exists a deep ReLU network $\tilde{\phi}$ with width $16(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that $|\tilde{\phi}(p) - \xi_p| \le (NL)^{-2}$ for $p = 0, 1, \cdots, K^d - 1$. Let us define

$$\phi(x) := 2C_2(d)\tilde{\phi}\left(\sum_{j=1}^d \eta_j K^j\right) - C_2(d).$$

Then it is clear that

$$|\phi(\eta(p)/N) - f_{K,m_*}(\eta(p)/N)| = 2C_2(d)|\tilde{\phi}(x) - \xi_p| \le 2C_2(d)(NL)^{-2}.$$

Furthermore, let $\psi_{m_*}(x) := \phi \circ \phi_2(x)$ for any $x \in \Omega_{m_*}$. Since $\psi_{m_*} - f_{K,m_*}$ is a step function whose first-order weak derivative is 0 over Ω_{m_*}, then it implies that

$$\|\psi_{m_*} - f_{K,m_*}\|_{W^{1,\infty}(\Omega_{m_*})} = \|\psi_{m_*} - f_{K,m_*}\|_{L^\infty(\Omega_{m_*})} \le 2C_2(d)(NL)^{-2}.$$

By the triangle inequalities for $\|\cdot\|_{L^\infty(\Omega_{m_*})}$ and $\|\cdot\|_{W^{1,\infty}(\Omega_{m_*})}$, it is easy to derive that

$$\|\psi_{m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \le \|\psi_{m_*} - f_{K,m_*}\|_{W^{1,\infty}(\Omega_{m_*})} + \|f_{K,m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \lesssim 1,$$
$$\|\psi_{m_*} - f\|_{L^\infty(\Omega_{m_*})} \le \|\psi_{m_*} - f_{K,m_*}\|_{L^\infty(\Omega_{m_*})} + \|f_{K,m_*} - f\|_{L^\infty(\Omega_{m_*})} \lesssim (NL)^{-2/d}.$$

Lastly, we calculate the width and depth of the deep ReLU network to implement $\psi_{m_*} = \phi \circ \phi_2$. Because that ϕ has width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$ and ϕ_2 has width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$, the deep ReLU network of ψ_{m_*} is constructed with width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$. This completes the proof. □

Proof of Lemma 3.49. We proceed in a similar way as the proof of Yang et al. [2023, Theorem 3]. By Lemma 3.54, there exists a sequence of deep ReLU networks $\{\phi_m\}_{m\in\{1,2\}^d}$ such that for any $m \in \{1,2\}^d$,

$$\|\phi_m - g_m\|_{W^{1,\infty}((0,1)^d)} \le 50 d^{5/2}(N+1)^{-4dL}.$$

Each ϕ_m is implemented by a deep ReLU network with width $\mathcal{O}(dN)$ and depth $\mathcal{O}(d^2L)$. By Lemma 3.56, there exists a collection of deep ReLU networks $\{\psi_m\}_{m\in\{1,2\}^d}$ such that for any $m \in \{1,2\}^d$,

$$\|\psi_m - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim (NL)^{-2/d},$$

where we omit constants in d. Each ψ_m is implemented by a deep ReLU network with width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$. Before proceeding, it is useful to estimate $\|\phi_m\|_{L^\infty(\Omega_m)}$, $\|\phi_m\|_{W^{1,\infty}(\Omega_m)}$, $\|\psi_m\|_{L^\infty(\Omega_m)}$, and $\|\psi_m\|_{W^{1,\infty}(\Omega_m)}$ as follows

$$\begin{aligned}\|\phi_m\|_{L^\infty(\Omega_m)} &\le \|\phi_m\|_{L^\infty([0,1]^d)} \le \|g_m\|_{L^\infty([0,1]^d)} + \|\phi_m - g_m\|_{L^\infty([0,1]^d)} \\ &\le 1 + 50d^{5/2} \lesssim d^{5/2},\end{aligned}$$

$$\begin{aligned}\|\phi_m\|_{W^{1,\infty}(\Omega_m)} &\le \|\phi_m\|_{W^{1,\infty}([0,1]^d)} \le \|g_m\|_{W^{1,\infty}([0,1]^d)} + \|\phi_m - g_m\|_{W^{1,\infty}([0,1]^d)} \\ &\le 4\lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor + 50d^{5/2},\end{aligned}$$

$$\|\psi_m\|_{L^\infty(\Omega_m)} \le \|f\|_{L^\infty(\Omega_m)} + \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim 1,$$

$$\|\psi_m\|_{W^{1,\infty}(\Omega_m)} \le \|f\|_{W^{1,\infty}([0,1]^d)} + \|\psi_m - f\|_{W^{1,\infty}([0,1]^d)} \lesssim 1.$$

Let $B_1 := \max_{m\in\{1,2\}^d}\{\|\phi_m\|_{L^\infty(\Omega_m)}, \|\psi_m\|_{L^\infty(\Omega_m)}\}$, then it yields that $B_1 \lesssim d^{5/2}$ by the estimates of $\|\phi_m\|_{L^\infty(\Omega_m)}$ and $\|\psi_m\|_{W^{1,\infty}(\Omega_m)}$. Let $B_2 := \max_{m\in\{1,2\}^d}\{\|\phi_m\|_{W^{1,\infty}(\Omega_m)}, \|\psi_m\|_{W^{1,\infty}(\Omega_m)}\}$.

Similarly, it yields that $B_2 \lesssim (NL)^{2/d} + d^{5/2}$. By Lemma 3.75, for any $N, L \in \mathbb{N}$, there exists a deep ReLU network $\phi_{\times,B_1}$ with width $15(N+1)$ and depth $16L$ such that $\|\phi_{\times,B_1}\|_{W^{1,\infty}((-B_1,B_1)^2)} \le 12B_1^2$ and

$$\|\phi_{\times,B_1}(x,y) - xy\|_{W^{1,\infty}((-B_1,B_1)^2)} \le 6B_1^2(N+1)^{-8L}.$$

To obtain a global estimate on $[0,1]^d$, we combine the local estimate $\{\psi_m\}_{m\in\{1,2\}^d}$ and the approximate partition of unity $\{\phi_m\}_{m\in\{1,2\}^d}$. Let us construct the global approximation function ϕ by

$$\phi(x) := \sum_{m\in\{1,2\}^d} \phi_{\times,B_1}(\phi_m(x), \psi_m(x)). \tag{3.30}$$

Next, we bound the error of the global approximation estimate by

$$\begin{aligned}
\|f-\phi\|_{L^\infty([0,1]^d)} =& \|\textstyle\sum_{m\in\{1,2\}^d} g_m f - \phi\|_{W^{1,\infty}((0,1)^d)} \\
\le& \underbrace{\|\textstyle\sum_{m\in\{1,2\}^d} [g_m f - \phi_m\psi_m]\|_{L^\infty([0,1]^d)}}_{=:\mathcal{R}_1} \\
&+ \underbrace{\|\textstyle\sum_{m\in\{1,2\}^d} [\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{L^\infty([0,1]^d)}}_{=:\mathcal{R}_2}
\end{aligned}$$

and

$$\begin{aligned}
\|f-\phi\|_{W^{1,\infty}((0,1)^d)} =& \|\textstyle\sum_{m\in\{1,2\}^d} g_m f - \phi\|_{W^{1,\infty}((0,1)^d)} \\
\le& \underbrace{\|\textstyle\sum_{m\in\{1,2\}^d} [g_m f - \phi_m\psi_m]\|_{W^{1,\infty}((0,1)^d)}}_{=:\mathcal{R}_3} \\
&+ \underbrace{\|\textstyle\sum_{m\in\{1,2\}^d} [\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{W^{1,\infty}((0,1)^d)}}_{=:\mathcal{R}_4}.
\end{aligned}$$

It remains to bound $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$, and $\mathcal{R}_4$, respectively. For the term $\mathcal{R}_1$, it holds

$$
\begin{aligned}
\mathcal{R}_1 &\leq \sum_{m\in\{1,2\}^d} \|g_m f - \phi_m \psi_m\|_{L^\infty([0,1]^d)} \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{L^\infty([0,1]^d)} + \|\phi_m (f - \psi_m)\|_{L^\infty([0,1]^d)}\right] \\
&= \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{L^\infty([0,1]^d)} + \|\phi_m (f - \psi_m)\|_{L^\infty(\Omega_m)}\right] \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|g_m - \phi_m\|_{L^\infty([0,1]^d)} \|f\|_{L^\infty([0,1]^d)} + \|\phi_m\|_{L^\infty(\Omega_m)} \|f - \psi_m\|_{L^\infty(\Omega_m)}\right] \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|g_m - \phi_m\|_{W^{1,\infty}([0,1]^d)} \|f\|_{W^{1,\infty}([0,1]^d)} + \|\phi_m\|_{L^\infty(\Omega_m)} \|f - \psi_m\|_{L^\infty(\Omega_m)}\right] \\
&\leq 2^d [50 d^{5/2} (N+1)^{-4dL} + (1 + 50 d^{5/2}) (NL)^{-2/d}] \\
&\lesssim (NL)^{-2/d},
\end{aligned}
$$

where we use $(NL)^{2/d} \leq (N+1)^{4dL}$ to derive the last inequality and omit a prefactor in d. For the term $\mathcal{R}_3$, it holds

$$
\begin{aligned}
\mathcal{R}_3 &\leq \sum_{m\in\{1,2\}^d} \|g_m f - \phi_m \psi_m\|_{W^{1,\infty}((0,1)^d)} \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{W^{1,\infty}((0,1)^d)} + \|\phi_m (f - \psi_m)\|_{W^{1,\infty}((0,1)^d)}\right] \\
&= \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{W^{1,\infty}((0,1)^d)} + \|\phi_m (f - \psi_m)\|_{W^{1,\infty}(\Omega_m)}\right] \\
&\leq \sum_{m\in\{1,2\}^d} \Big[\|g_m - \phi_m\|_{W^{1,\infty}((0,1)^d)} \|f\|_{W^{1,\infty}((0,1)^d)} \\
&\qquad + \|\phi_m\|_{W^{1,\infty}(\Omega_m)} \|f - \psi_m\|_{L^\infty(\Omega_m)} + \|\phi_m\|_{L^\infty(\Omega_m)} \|f - \psi_m\|_{W^{1,\infty}(\Omega_m)}\Big] \\
&\lesssim 2^d [50 d^{5/2} (N+1)^{-4dL} + (4 \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor + 50 d^{5/2}) (NL)^{-2/d} + (1 + 50 d^{5/2})] \\
&\lesssim 1,
\end{aligned}
$$

where we use $(NL)^{2/d} \leq (N+1)^{4dL}$ to derive the last inequality and omit a prefactor in

d. For the terms $\mathcal{R}_2$ and $\mathcal{R}_4$, it holds

$$
\begin{aligned}
\mathcal{R}_2 \le \mathcal{R}_4 &\\
&\le \sum_{m\in\{1,2\}^d} \|[\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{W^{1,\infty}((0,1)^d)} \\
&\le \sum_{m\in\{1,2\}^d} \|[\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{W^{1,\infty}(\Omega_m)} \\
&\le \sum_{m\in\{1,2\}^d} 2\sqrt{d}\max\Big\{\|\phi_{\times,B_1}(x,y)-xy\|_{L^\infty((-B_1,B_1)^2)}, \\
&\quad |\phi_{\times,B_1}(x,y)-xy|_{W^{1,\infty}((-B_1,B_1)^2)} \times \max\{|\phi_m|_{W^{1,\infty}(\Omega_m)}, |\psi_m|_{W^{1,\infty}(\Omega_m)}\}\Big\} \\
&\le \sum_{m\in\{1,2\}^d} 2\sqrt{d}\|\phi_{\times,B_1}(x,y)-xy\|_{W^{1,\infty}((-B_1,B_1)^2)} \\
&\qquad \times \max\{\|\phi_m\|_{W^{1,\infty}(\Omega_m)}, \|\psi_m\|_{W^{1,\infty}(\Omega_m)}\} \\
&\le \sum_{m\in\{1,2\}^d} 12\sqrt{d}B_1^2(N+1)^{-8L}B_2 \\
&\lesssim 2^d\sqrt{d}d^5(N+1)^{-8L}((NL)^{2/d}+d^{5/2}) \\
&\lesssim 2^d\sqrt{d}d^5(d^{5/2}(NL)^{2/d})(N+1)^{-8L} \\
&\lesssim (NL)^{2/d}(N+1)^{-8L} \\
&\lesssim (NL)^{-2/d},
\end{aligned}
$$

where we use $(NL)^{2/d} \le (N+1)^{4L}$ in the last inequality and omit constants depending only on d. Combining the estimates of $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$, and $\mathcal{R}_4$, we have

$$\|f-\phi\|_{L^\infty([0,1]^d)} \le \mathcal{R}_1 + \mathcal{R}_2 \lesssim (NL)^{-2/d}$$

and

$$\|f-\phi\|_{W^{1,\infty}([0,1]^d)} \le \mathcal{R}_3 + \mathcal{R}_4 \lesssim 1 + (NL)^{-2/d} \lesssim 1.$$

It is easy to see

$$\|\phi\|_{W^{1,\infty}([0,1]^d)} \le \|f\|_{W^{1,\infty}([0,1]^d)} + \|f-\phi\|_{W^{1,\infty}([0,1]^d)} \lesssim 1.$$

Lastly, we calculate the complexity of the constructed deep ReLU network ϕ in (3.30). By the definition of ϕ in (3.30), we know that ϕ consists of $\mathcal{O}(2^d)$ parallel subnetworks listed as follows:

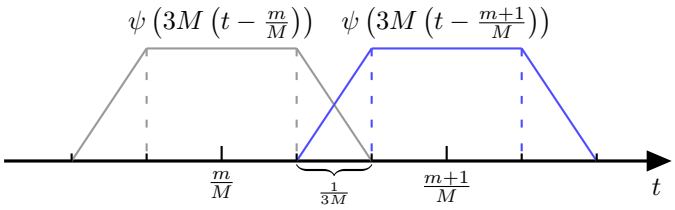

Figure 3.2. Functions $\phi_m(t)$ and $\phi_{m+1}(t)$ for defining a partition of unity.

- $\phi_{\times,B_1}$ with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$;
- ϕ_m with width $\mathcal{O}(dN)$ and depth $\mathcal{O}(d^2L)$;
- ψ_m with width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$.

Hence, the deep ReLU network implementing the function ϕ has width $\mathcal{O}(2^d dN\log N)$ and depth $\mathcal{O}(d^2L\log L)$. □

Proof of Corollary 3.50. The proof is completed by employing Lemma 3.49 on

$$\bar{f} := f/\|f\|_{W^{1,\infty}((0,1)^d)}.$$

□

Approximation in time with Lipschitz regularity

To handle the singularity of the velocity field in time, we develop a new approximation result to approximate the velocity field in time.

Lemma 3.57. *Given any $f \in W^{1,\infty}((0,1))$ with $\|f\|_{W^{1,\infty}((0,1))} \le \infty$, for any $M \in \mathbb{N}$, there exists a function ξ implemented by a deep ReLU network with width $\mathcal{O}(M)$ and depth $\mathcal{O}(1)$ such that $|\xi|_{W^{1,\infty}((0,1))} \lesssim |f|_{W^{1,\infty}((0,1))}$ and*

$$\|\xi - f\|_{L^\infty([0,1])} \lesssim |f|_{W^{1,\infty}((0,1))}/M.$$

Proof. The proof consists of two steps. We start with the construction of a continuous piecewise linear function for approximating 1-Lipschitz functions, which shall be implemented by a deep ReLU network. After that, we establish the global Lipschitz continuity of the constructed deep ReLU network, in addition to the approximation bounds in the $L^\infty([0,1])$ norm.

Step 1. We construct a partition of unity following Yarotsky [2017, Proof of Theorem 1]. Let $M \in \mathbb{N}$ and $m \in \{0, 1, \cdots, M\}$. We collect a set of functions $\{\phi_m\}_{m=0}^M$ that are defined as follows: for any $t \in [0, 1]$, let

$$\phi_m(t) := \psi\left(3M\left(t - \frac{m}{M}\right)\right) \quad \text{with} \quad \psi(z) = \begin{cases} 1, & |z| < 1, \\ 0, & |z| > 2, \\ 2 - |z|, & 1 \le |z| \le 2, \end{cases} \tag{3.31}$$

that satisfies $\sum_{m=0}^M \phi_m(t) = 1$. It implies that $\{\phi_m\}_{m=0}^M$ forms a partition of unity on the domain $[0, 1]$. We plot ϕ_m and ϕ_{m+1} in Figure 3.2. As in Chen et al. [2020, Proof of Lemma 10], for each $m \in \{0, 1, \cdots, M\}$, we consider a piecewise constant function $f_m := f(m/M)$. Actually, the piecewise constant approximation is specially the zero-degree Taylor polynomial for the function f at $x = m/M$ in Yarotsky [2017, Proof of Theorem 1]. We claim that

$$\tilde{f}(t) := \sum_{m=0}^{M} \phi_m(t) f_m \tag{3.32}$$

provides an approximation of f, and the approximation error is evaluated by

$$\begin{aligned} \|\tilde{f} - f\|_{L^\infty([0,1])} &= \sup_{t\in[0,1]} \left| \sum\nolimits_{m=0}^{M} \phi_m(t)[f_m - f(t)] \right| \\ &= \sup_{t\in[0,1]} \left| \sum\nolimits_{|t-\frac{m}{M}|\le\frac{2}{3M}} \phi_m(t)[f(m/M) - f(t)] \right| \\ &\le \frac{2}{3M} |f|_{W^{1,\infty}((0,1))}, \end{aligned}$$

where the Lipschitz continuity of f is used in the inequality. It is clear that $\tilde{f}$ can be implemented with a deep ReLU network.

Step 2. We establish the global Lipschitz continuity of $\tilde{f}$. Notice that for any $t, s \in [0, 1]$,

$$\begin{aligned} |\tilde{f}(t) - \tilde{f}(s)| &\le |\tilde{f}(t) - f(t)| + |f(t) - f(s)| + |f(s) - \tilde{f}(s)| \\ &\le 2\|\tilde{f} - f\|_{L^\infty([0,1])} + |f|_{W^{1,\infty}((0,1))}|t - s| \\ &\le \frac{4}{3M} |f|_{W^{1,\infty}((0,1))} + |f|_{W^{1,\infty}((0,1))}|t - s|. \end{aligned}$$

(1) If $|t - s| \ge \frac{1}{3M}$, it is clear that $|\tilde{f}(t) - \tilde{f}(s)| \le 5|f|_{W^{1,\infty}((0,1))}|t - s|$.

(2) If $|t-s| < \frac{1}{3M}$, we try to directly bound the difference

$$\begin{aligned}|\tilde{f}(t)-\tilde{f}(s)| &= \Big|\sum\nolimits_{m=0}^{M}[\phi_m(t)-\phi_m(s)]f_m\Big| \\ &= \Big|\sum\nolimits_{m=0}^{M}[\psi(3Mt-3m)-\psi(3Ms-3m)]f_m\Big| =: \mathcal{E}.\end{aligned}$$

Next, we focus on bounding $\mathcal{E}$. Without loss of generality, we assume $s > t$. Considering $|t-s| < \frac{1}{3M}$, we deduce that $s \in (t, t+\frac{1}{3M})$. From Figure 3.2, we can observe that there exist at most two numbers $m = \tilde{m} \in \{0,1,\cdots,M\}$ or $m = \bar{m} := \tilde{m}+1$ such that $\psi(3Mt-3m) \not\equiv 0$ or $\psi(3Ms-3m) \not\equiv 0$. It follows that

$$\begin{aligned}\mathcal{E} &= \Big|[\psi(3Mt-3\tilde{m})-\psi(3Ms-3\tilde{m})]f_{\tilde{m}} \\ &\quad + [\psi(3Mt-3\bar{m})-\psi(3Ms-3\bar{m})]f_{\bar{m}}\Big| \\ &= \Big|[\psi(3Mt-3\tilde{m})-\psi(3Ms-3\tilde{m})]f_{\tilde{m}} \\ &\quad + [(1-\psi(3Mt-3\tilde{m}))-(1-\psi(3Ms-3\tilde{m}))]f_{\bar{m}}\Big| \\ &= \Big|[\psi(3Mt-3\tilde{m})-\psi(3Ms-3\tilde{m})](f_{\tilde{m}}-f_{\bar{m}})\Big| \\ &\le |f_{\tilde{m}}-f_{\bar{m}}|\cdot\Big|\psi(3Mt-3\tilde{m})-\psi(3Ms-3\tilde{m})\Big| \\ &\le \frac{1}{M}|f|_{W^{1,\infty}((0,1))}\Big|\psi(3Mt-3\tilde{m})-\psi(3Ms-3\tilde{m})\Big| \\ &= 3|f|_{W^{1,\infty}((0,1))}|t-s|.\end{aligned}$$

Hence, if $|t-s| < \frac{1}{3M}$, it holds that $|\tilde{f}(t)-\tilde{f}(s)| \le 3|f|_{W^{1,\infty}((0,1))}|t-s|$.

To sum up, for any $t, s \in [0,1]$, it holds that $|\tilde{f}(t)-\tilde{f}(s)| \le 5|f|_{W^{1,\infty}((0,1))}|t-s|$. It is easy to see from (3.32) that the deep ReLU network implementing $\tilde{f}$ has width $\mathcal{O}(M)$ and depth $\mathcal{O}(1)$. Then we complete the proof. □

Time-space approximation

In the subsection, we construct a time-space approximation while keeping the Lipschitz regularity both in the space variable and in the time variable.

Lemma 3.58 (Clipping functions). *Given $A > 0$, we define $\beta_A : \mathbb{R} \to [-A, A]$ by*

$$\beta_A(z) := \begin{cases} -A, & z \in (-\infty, -A), \\ z, & z \in [-A, A], \\ A, & z \in (A, \infty). \end{cases}$$

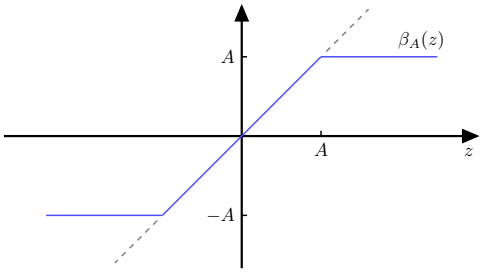

Figure 3.3. The clipping function β_A.

There exists a clipping function $\mathcal{C}_A : \mathbb{R}^d \to [-A, A]^d$ *at level* A *implemented by a deep ReLU network with width* $\mathcal{O}(d)$ *and depth* $\mathcal{O}(1)$ *such that for any* $x = [x_1, x_2, \cdots, x_d]^\top \in \mathbb{R}^d$,

$$\mathcal{C}_A(x) = [\beta_A(x_1), \beta_A(x_2), \cdots, \beta_A(x_d)]^\top.$$

Proof. It is clear that $\mathcal{C}_A(x) = \rho(x + A\mathbf{1}_d) - \rho(x - A\mathbf{1}_d) - A\mathbf{1}_d$ where $\rho : \mathbb{R}^d \to \mathbb{R}^d$ is the ReLU function. This expression implies that the clipping function $\mathcal{C}_A$ can be implemented by a deep ReLU network with width $\mathcal{O}(d)$ and depth $\mathcal{O}(1)$. □

The main idea of the time-space approximation on $[0, 1-\underline{t}]\times\mathbb{R}^d$ is based on Lemmas 3.49, 3.57, and 3.58.

Proof of Theorem 3.31. We derive a time-space approximation $\bar{v}$ of the velocity field v^* on the domain $\Omega_{\underline{t},A} = [0, 1-\underline{t}]\times[-A, A]^d$ and bound the Lipschitz constants of $\bar{v}$ on the domain $[0, 1-\underline{t}]\times\mathbb{R}^d$.

First of all, we use the clipping function $\mathcal{C}$ defined in Lemma 3.58 to clip the support of the space variable, that is, for each $x \in \mathbb{R}^d$, we have $\mathcal{C}(x) \in [-A, A]^d$. We only need to consider approximation in x on the domain $[-A, A]^d$.

Then we can employ the mappings $\tilde{t} = \mathcal{T}_1(t) := t/(1-\underline{t})$ and $\tilde{x} = \mathcal{T}_2(x) := (x + A\mathbf{1}_d)/(2A)$ to transform the domain $\Omega_{\underline{t},A}$ into the domain $[0, 1]^{d+1}$. When the domain $[0, 1-\underline{t}]\times\mathbb{R}^d$ is considered, the transformed domain is $[0, 1]\times\mathbb{R}^d$. Notice that both mappings are invertible and can be implemented by deep ReLU networks. We denote their inverse functions as $t = \mathcal{T}_1^{-1}(\tilde{t})$ and $x = \mathcal{T}_2^{-1}(\tilde{x})$. We further define a new velocity field $v^\diamond$ by $v^\diamond(\tilde{t}, \tilde{x}) := v^*(\mathcal{T}_1^{-1}(\tilde{t}), \mathcal{T}_2^{-1}(\tilde{x}))$ for any $(\tilde{t}, \tilde{x}) \in [0, 1]\times\mathbb{R}^d$. It is clear that $v^*(t, x) = v^\diamond(\mathcal{T}_1(t), \mathcal{T}_2(x))$ for any $(t, x) \in [0, 1-\underline{t}]\times\mathbb{R}^d$. According to Theorem 3.26, the new velocity field $v^\diamond$ satisfies

(1) For any $\tilde{s}, \tilde{t} \in [0,1]$ and $\tilde{x} \in \mathbb{R}^d$, $\|v^\diamond(\tilde{t},\tilde{x}) - v^\diamond(\tilde{s},\tilde{x})\|_\infty \lesssim \underline{t}^{-2}|\tilde{t}-\tilde{s}|$;

(2) For any $\tilde{x}, \tilde{y} \in \mathbb{R}^d$ and $\tilde{t} \in [0,1]$, $\|v^\diamond(\tilde{t},\tilde{x}) - v^\diamond(\tilde{t},\tilde{y})\|_\infty \lesssim A\|\tilde{x}-\tilde{y}\|_\infty$;

(3) $\|v^\diamond\|_{L^\infty([0,1]^{d+1})} \lesssim A$,

where we omit constants in $d, \kappa, \beta, \sigma, R$.

In the following, we construct a time-space approximation of the new velocity field $v^\diamond$ on the transformed domain $[0,1]^{d+1}$ using deep ReLU networks. Let $v^\diamond = [v_1^\diamond, v_2^\diamond, \cdots, v_d^\diamond]^\top$. Given $M \in \mathbb{N}$, we uniformly partition the unit interval $[0,1]$ into M non-overlapping sub-intervals with length $1/M$. Let $\{\phi_j(\tilde{t})\}_{j=0}^M$ form a partition of unity on $[0,1]$ with the same definition as (3.31) in the proof of Lemma 3.57. For each $i \in \{1,2,\cdots,d\}$, we define a time approximation of $v_i^\diamond$ by

$$\tilde{v}_i(\tilde{t},\tilde{x}) := \sum_{j=0}^M v_i^\diamond(j/M,\tilde{x})\,\phi_j(\tilde{t}).$$

Let $\tilde{v} := [\tilde{v}_1, \tilde{v}_2, \cdots, \tilde{v}_d]^\top$. Due to Lemma 3.57, for any $\tilde{x} \in [0,1]^d$, it holds that

$$|\tilde{v}(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} \lesssim |v^\diamond(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} \lesssim \underline{t}^{-2}.$$

For $i = 1,2,\cdots,d$ and $j = 0,1,\cdots,M$, let $\zeta_{ij}(\tilde{x})$ be a space approximation of $v_i^\diamond(j/M,\tilde{x})$ implemented by a deep ReLU network constructed in Lemma 3.49. Then it holds that $\max_{i,j}\|\zeta_{ij}\|_{W^{1,\infty}((0,1)^d)} \lesssim A$. By Lemma 3.76, we can construct a deep ReLU network $\phi_{\times,(B_3,B_4)}$ with width $15(N+1)$ and depth $8L$ to approximate the product function such that $\|\phi_{\times,(B_3,B_4)}\|_{W^{1,\infty}((-B_3,B_3)\times(-B_4,B_4))} \le 12B_3B_4$,

$$\|\phi_{\times,(B_3,B_4)}(x,y) - xy\|_{W^{1,\infty}((-B_3,B_3)\times(-B_4,B_4))} \le 6B_3B_4(N+1)^{-4L}, \tag{3.33}$$

and

$$\phi_{\times,(B_3,B_4)}(x,0) = \frac{\partial \phi_{\times,(B_3,B_4)}(x,0)}{\partial x} = 0 \text{ for } x \in (-B_3,B_3). \tag{3.34}$$

Using the same partition of unity $\{\phi_j(\tilde{t})\}_{j=0}^M$ on $[0,1]$, we define a time-space approximation of $v_i^\diamond$ for each $i \in \{1,2,\cdots,d\}$ by

$$v_i^\natural(\tilde{t},\tilde{x}) := \sum_{j=0}^M \phi_{\times,(B_3,B_4)}\Big(\zeta_{ij}(\tilde{x}), \phi_j(\tilde{t})\Big), \tag{3.35}$$

which can be implemented with a deep ReLU network. We choose the parameters B_3, B_4 such that $B_3 \asymp \max_{i,j}\|\zeta_{ij}\|_{L^\infty([0,1]^d)} \lesssim A$ and $B_4 \asymp \max_j \|\phi_j\|_{L^\infty([0,1])} \lesssim 1$.

Claim 3.59. There are at most two nonzero terms in the summation (3.35) defining the time-space approximation function $v_i^\natural$.

This claim holds because for any $\tilde{t} \in [0,1]$, there are at most two indexes j's from $\{0,1,2,\cdots,M\}$ such that $\phi_j(\tilde{t})$ is nonzero according to the definition of the partition of unity $\{\phi_j(\tilde{t})\}_{j=0}^M$. Then our claim follows from the property (3.34) of the approximation product function $\phi_{\times,(B_3,B_4)}$.

Before we study the properties of $v_i^\natural$, we introduce a surrogate function $\check{v}_i$ defined by

$$\check{v}_i(\tilde{t},\tilde{x}) := \sum_{j=0}^{M} \zeta_{ij}(\tilde{x})\phi_j(\tilde{t}).$$

The function $\check{v}_i$ will be useful to study the approximation capacity and the regularity of $v_i^\natural$. We derive the approximation rate and the regularity properties of $\check{v}_i$ in the following. Due to Lemma 3.49, for any $\tilde{x} \in [0,1]^d$, $i = 1,2,\cdots,d$, and $j = 0,1,\cdots,M$, we have

$$|\zeta_{ij}(\tilde{x}) - v_i^\diamond(j/M,\tilde{x})| \lesssim A(NL)^{-2/d}.$$

We evaluate the approximation error of $\check{v}_i$ by the following error decomposition:

$$\|\check{v}_i - v_i^\diamond\|_{L^\infty([0,1]^{d+1})} \le \underbrace{\|\check{v}_i - \tilde{v}_i\|_{L^\infty([0,1]^{d+1})}}_{=:\mathcal{E}_i^1} + \underbrace{\|\tilde{v}_i - v_i^\diamond\|_{L^\infty([0,1]^{d+1})}}_{=:\mathcal{E}_i^2}. \tag{3.36}$$

By Lemma 3.49, we bound $\mathcal{E}_i^1$ by

$$\begin{aligned}
\mathcal{E}_i^1 &\le \Big\| \sum_{j=0}^{M} \Big[\zeta_{ij}(\tilde{x}) - v_i^\diamond(j/M,\tilde{x})\Big]\phi_j(\tilde{t}) \Big\|_{L^\infty([0,1]^{d+1})} \\
&\le \max_{0\le j\le M} \|\zeta_{ij}(\tilde{x}) - v_i^\diamond(j/M,\tilde{x})\|_{L^\infty([0,1]^d)} \\
&\lesssim \max_{0\le j\le M} \|v_i^\diamond(j/M,\tilde{x})\|_{W^{1,\infty}((0,1)^d)} (NL)^{-2/d} \\
&\lesssim A(NL)^{-2/d}.
\end{aligned} \tag{3.37}$$

By Lemma 3.57, we bound $\mathcal{E}_i^2$ by

$$\mathcal{E}_i^2 \lesssim \sup_{\tilde{x}\in[0,1]^d} |v_i^\diamond(\cdot,\tilde{x})|_{W^{1,\infty}((0,1))}/M \lesssim \underline{t}^{-2}M^{-1}. \tag{3.38}$$

Combining (3.36), (3.37), and (3.38), we have

$$\|\check{v}_i - v_i^\diamond\|_{L^\infty([0,1]^{d+1})} \lesssim A(NL)^{-2/d} + \underline{t}^{-2}M^{-1}.$$

Suppose that $(NL)^{2/d} \asymp \underline{t}^2 M$, and it yields that for $i = 1, 2, \cdots, d$,

$$\|\check{v}_i - v_i^{\diamond}\|_{L^\infty([0,1]^{d+1})} \lesssim A(NL)^{-2/d}.$$

Let $\check{v} := [\check{v}_1, \check{v}_2, \cdots, \check{v}_d]^\top$. We have the approximation power of $\check{v}$ evaluated by

$$\|\check{v} - v^{\diamond}\|_{L^\infty([0,1]^{d+1})} \lesssim A(NL)^{-2/d}. \tag{3.39}$$

Moreover, the Lipschitz continuity of $\check{v}$ in $\tilde{t}$ and $\tilde{x}$ can be verified. Concretely, we have the Lipschitz estimate in the space variable $\tilde{x}$: for any $\tilde{x}, \tilde{y} \in [0,1]^d$ and $\tilde{t} \in [0,1]$,

$$\begin{aligned}
\|\check{v}(\tilde{t}, \tilde{x}) - \check{v}(\tilde{t}, \tilde{y})\|_\infty &\le \max\nolimits_{1\le i\le d} \Big\| \sum\nolimits_{j=0}^{M} [\zeta_{ij}(\tilde{x}) - \zeta_{ij}(\tilde{y})]\phi_j(\tilde{t}) \Big\|_\infty \\
&\le \max\nolimits_{1\le i\le d,\ 0\le j\le M} \|\zeta_{ij}(\tilde{x}) - \zeta_{ij}(\tilde{y})\|_\infty \\
&\le \max\nolimits_{1\le i\le d,\ 0\le j\le M} \|\zeta_{ij}\|_{W^{1,\infty}((0,1)^d)} \cdot \|\tilde{x} - \tilde{y}\|_\infty \\
&\lesssim \|v^{\diamond}\|_{W^{1,\infty}((0,1)^d;\mathbb{R}^d)} \cdot \|\tilde{x} - \tilde{y}\|_\infty \\
&\lesssim A\|\tilde{x} - \tilde{y}\|_\infty,
\end{aligned}$$

It is somewhat tedious to derive the Lipschitz estimate in the time variable $\tilde{t}$. For any $\tilde{s}, \tilde{t} \in [0,1]$ and $\tilde{x} \in [0,1]^d$,

$$\begin{aligned}
&\|\check{v}(\tilde{t}, \tilde{x}) - \check{v}(\tilde{s}, \tilde{x})\|_\infty \\
\le\ &\|\check{v}(\tilde{t}, \tilde{x}) - \tilde{v}(\tilde{t}, \tilde{x})\|_\infty + \|\tilde{v}(\tilde{t}, \tilde{x}) - \tilde{v}(\tilde{s}, \tilde{x})\|_\infty + \|\tilde{v}(\tilde{s}, \tilde{x}) - \check{v}(\tilde{s}, \tilde{x})\|_\infty \\
\le\ &2 \sup_{\vartheta\in[0,1]} \|\check{v}(\vartheta, \tilde{x}) - \tilde{v}(\vartheta, \tilde{x})\|_\infty + |\tilde{v}(\cdot, \tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} |\tilde{t} - \tilde{s}| \\
\le\ &2 \max\nolimits_{1\le i\le d} \mathcal{E}_i^1 + |\tilde{v}(\cdot, \tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} |\tilde{t} - \tilde{s}| \\
\lesssim\ &A(NL)^{-2/d} + \underline{t}^{-2}|\tilde{t} - \tilde{s}|.
\end{aligned}$$

Considering $(NL)^{2/d} \asymp \underline{t}^2 M$, we deduce that

$$\|\check{v}(\tilde{t}, \tilde{x}) - \check{v}(\tilde{s}, \tilde{x})\|_\infty \lesssim A\underline{t}^{-2}M^{-1} + \underline{t}^{-2}|\tilde{t} - \tilde{s}|.$$

Then we consider two cases for bounding $\|\check{v}(\tilde{t}, \tilde{x}) - \check{v}(\tilde{s}, \tilde{x})\|_\infty$.

Case 1. If $|\tilde{t} - \tilde{s}| \ge \frac{1}{3M}$, it is clear that $\|\check{v}(\tilde{t}, \tilde{x}) - \check{v}(\tilde{s}, \tilde{x})\|_\infty \lesssim A\underline{t}^{-2}|\tilde{t} - \tilde{s}|$.

Case 2. If $|\tilde{t} - \tilde{s}| < \frac{1}{3M}$, for any $i \in \{1, 2, \cdots, d\}$, we try to bound the difference

$$\begin{aligned}
&|\check{v}_i(\tilde{t}, \tilde{x}) - \check{v}_i(\tilde{s}, \tilde{x})| \\
=\ &\Big| \sum\nolimits_{j=0}^{M} \zeta_{ij}(\tilde{x}) \Big[\phi_j(\tilde{t}) - \phi_j(\tilde{s}) \Big] \Big| \\
=\ &\Big| \sum\nolimits_{j=0}^{M} \zeta_{ij}(\tilde{x}) [\psi(3M\tilde{t} - 3j) - \psi(3M\tilde{s} - 3j)] \Big| =: \mathcal{E}_3.
\end{aligned}$$

Then, we focus on bounding $\mathcal{E}_3$. Without loss of generality, we assume $\tilde{t} < \tilde{s}$. The remaining calculation is similar to the proof of Lemma 3.57. Let $m = \tilde{m} \in \{0, 1, \cdots, M\}$ or $m = \bar{m} := \tilde{m}+1$ be two possible numbers satisfying $\psi(3M\tilde{t}-3m) \not\equiv 0$ or $\psi(3M\tilde{s}-3m) \not\equiv 0$. Then it holds that

$$\begin{aligned}
\mathcal{E}_3 &= \Big|[\psi(3M\tilde{t}-3\tilde{m})-\psi(3M\tilde{s}-3\tilde{m})]\zeta_{i\tilde{m}}(\tilde{x}) \\
&\quad + [\psi(3M\tilde{t}-3\bar{m})-\psi(3M\tilde{s}-3\bar{m})]\zeta_{i\bar{m}}(\tilde{x})\Big| \\
&= \Big|[\psi(3M\tilde{t}-3\tilde{m})-\psi(3M\tilde{s}-3\tilde{m})]\zeta_{i\tilde{m}}(\tilde{x}) \\
&\quad + [(1-\psi(3M\tilde{t}-3\tilde{m}))-(1-\psi(3M\tilde{s}-3\tilde{m}))]\zeta_{i\bar{m}}(\tilde{x})\Big| \\
&= \Big|[\psi(3M\tilde{t}-3\tilde{m})-\psi(3M\tilde{s}-3\tilde{m})][\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})]\Big| \\
&\le |\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})|\Big|\psi(3M\tilde{t}-3\tilde{m})-\psi(3M\tilde{s}-3\tilde{m})\Big| \\
&\le 3M|\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})|\cdot|\tilde{t}-\tilde{s}|.
\end{aligned}$$

We bound the term $|\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})|$ by

$$\begin{aligned}
&|\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})| \\
\le\ &|\zeta_{i\tilde{m}}(\tilde{x})-v_i^\diamond(\tilde{m}/M,\tilde{x})|+|v_i^\diamond(\tilde{m}/M,\tilde{x})-v_i^\diamond(\bar{m}/M,\tilde{x})| \\
&+|v_i^\diamond(\bar{m}/M,\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})| \\
\lesssim\ &A(NL)^{-2/d}+\underline{t}^{-2}M^{-1}.
\end{aligned}$$

Recall that $(NL)^{2/d} \asymp \underline{t}^2 M$. It implies that $|\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})| \lesssim A\underline{t}^{-2}M^{-1}$. Therefore, if $|\tilde{t}-\tilde{s}| < \frac{1}{3M}$, it holds that

$$|\check{v}_i(\tilde{t},\tilde{x})-\check{v}_i(\tilde{s},\tilde{x})| \le 3M|\zeta_{i\tilde{m}}(\tilde{x})-\zeta_{i\bar{m}}(\tilde{x})||\tilde{t}-\tilde{s}| \lesssim A\underline{t}^{-2}|\tilde{t}-\tilde{s}|.$$

We summarize the Lipschitz properties of $\check{v}$ as follows:

$$|\check{v}(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} \lesssim A\underline{t}^{-2}, \quad |\check{v}(\tilde{t},\cdot)|_{W^{1,\infty}((0,1)^d;\mathbb{R}^d)} \lesssim A.$$

Let $v^\natural := [v_1^\natural, v_2^\natural, \cdots, v_d^\natural]^\top$. We use the approximation rate of $\check{v}$ to derive that of $v^\natural$. By the triangle inequality, it holds that

$$\begin{aligned}
&\|v^\natural - v^\diamond\|_{L^\infty([0,1]^{d+1})} \\
\le&\|v^\natural-\check{v}\|_{L^\infty([0,1]^{d+1})}+\|\check{v}-v^\diamond\|_{L^\infty([0,1]^{d+1})} \\
\lesssim&A(N+1)^{-4L}+A(NL)^{-2/d} \quad \text{(By Claim 3.59, Eq. (3.33), and Eq. (3.39))} \\
\lesssim&A(NL)^{-2/d} \quad \text{(By } (NL)^{2/d} \le (N+1)^{4L} \text{ for any } N,L\in\mathbb{N}).
\end{aligned}$$

Thus, the approximation rate of $v^\natural$ is given by

$$\|v^\natural - v^\diamond\|_{L^\infty([0,1]^{d+1})} \lesssim A(NL)^{-2/d}. \tag{3.40}$$

Then we study the Lipschitz properties of $v^\natural$. By Lemma 3.73 and Claim 3.59, it holds that

$$\begin{aligned}
&|v^\natural(\cdot,\tilde{x}) - \check{v}(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} \\
=&\max_i |v_i^\natural(\cdot,\tilde{x}) - \check{v}_i(\cdot,\tilde{x})|_{W^{1,\infty}((0,1))} \\
=&\max_i \Big| \sum_{j=0}^M \phi_{\times,(B_3,B_4)}\Big(\zeta_{ij}(\tilde{x}), \phi_j(\cdot)\Big) - \sum_{j=0}^M \zeta_{ij}(\tilde{x})\phi_j(\cdot)\Big|_{W^{1,\infty}((0,1))} \\
\lesssim&\max_i |\phi_{\times,(B_3,B_4)}(x,y) - xy|_{W^{1,\infty}((-B_3,B_3)\times(-B_4,B_4))} \cdot |\phi_j|_{W^{1,\infty}((0,1))} \\
\lesssim&AM(N+1)^{-4L} \lesssim A\underline{t}^{-2}(NL)^{2/d}(N+1)^{-4L} \quad (\text{By } (NL)^{2/d} \asymp \underline{t}^2 M) \\
\lesssim&A\underline{t}^{-2} \quad (\text{By } (NL)^{2/d} \le (N+1)^{4L} \text{ for any } N,L \in \mathbb{N}).
\end{aligned}$$

By the triangle inequality, the Lipschitz property of $v^\natural$ in the time variable $\tilde{t}$ is evaluated by

$$\begin{aligned}
&|v^\natural(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} \\
\le&|v^\natural(\cdot,\tilde{x}) - \check{v}(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} + |\check{v}(\cdot,\tilde{x})|_{W^{1,\infty}((0,1);\mathbb{R}^d)} \\
\lesssim&A\underline{t}^{-2} + A\underline{t}^{-2} \lesssim A\underline{t}^{-2}.
\end{aligned}$$

By Lemma 3.73 and Claim 3.59, we derive the Lipschitz property of $v^\natural$ in the space variable $\tilde{x}$ as follows

$$\begin{aligned}
&|v^\natural(\tilde{t},\cdot)|_{W^{1,\infty}((0,1)^d)} \\
=&\max_i |v_i^\natural(\tilde{t},\cdot)|_{W^{1,\infty}((0,1)^d)} \\
=&\max_i \Big| \sum_{j=0}^M \phi_{\times,(B_3,B_4)}\Big(\zeta_{ij}(\cdot), \phi_j(\tilde{t})\Big)\Big|_{W^{1,\infty}((0,1)^d)} \\
\lesssim&\max_i \Big\{ |\phi_{\times,(B_3,B_4)}(\cdot,y)|_{W^{1,\infty}((-B_3,B_3))} \cdot \max_j |\zeta_{ij}|_{W^{1,\infty}((0,1)^d)} \Big\} \\
\lesssim&A^2.
\end{aligned}$$

We claim that $\bar{v}(t,x) := v^\natural(\mathcal{T}_1(t), \mathcal{T}_2 \circ \mathcal{C}_A(x))$ provides a good approximation of the velocity field v^* on the domain $\Omega_{\underline{t},A}$, and $\bar{v}$ can be implemented by a deep ReLU network.

According to the error bound (3.40), the approximation rate of $\bar{v}$ is given by

$$\|\bar{v}(t,x) - v^*(t,x)\|_{L^\infty(\Omega_{\underline{t},A})} \lesssim A^2(NL)^{-2/d},$$

where we omit constants in $d, \kappa, \beta, \sigma, R$. Furthermore, we need to estimate the Lipschitz constants of $\bar{v}$. Here, we use Lemma 3.73 to calculate the Sobolev semi-norms of the composite functions:

$$|\bar{v}(\cdot,x)|_{W^{1,\infty}((0,1-\underline{t});\mathbb{R}^d)} \lesssim |v^\natural(\cdot, \mathcal{T}_2 \circ \mathcal{C}_A(x))|_{W^{1,\infty}((0,1);\mathbb{R}^d)} |\mathcal{T}_1|_{W^{1,\infty}((0,1-\underline{t});(0,1))} \lesssim A\underline{t}^{-2},$$

$$|\bar{v}(t,\cdot)|_{W^{1,\infty}(\mathbb{R}^d;\mathbb{R}^d)} \lesssim |v^\natural(\mathcal{T}_1(t),\cdot)|_{W^{1,\infty}((0,1)^d;\mathbb{R}^d)} |\mathcal{T}_2|_{W^{1,\infty}((-A,A)^d;\mathbb{R}^d)} |\mathcal{C}_A|_{W^{1,\infty}(\mathbb{R}^d;[-A,A]^d)} \lesssim A.$$

In addition, we have the L^∞ bound $\|\bar{v}\|_{L^\infty([1-\underline{t}]\times\mathbb{R}^d)} \lesssim A$.

In the end, it remains to calculate the complexity of the deep ReLU network implementing $\bar{v}$. By the definition of $v^\natural$ in (3.35), we know that $v^\natural$ consists of $\mathcal{O}(\underline{t}^{-2}d(NL)^{2/d})$ parallel subnetworks listed as follows:

- $\phi_{\times,(B_3,B_4)}$ with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$;
- ζ_{ij} with width $\mathcal{O}(2^d dN\log N)$ and depth $\mathcal{O}(d^2 L\log L)$;
- ϕ_j with width $\mathcal{O}(1)$ and depth $\mathcal{O}(1)$.

Hence, the deep ReLU network implementing $v^\natural$ has width $\mathcal{O}(\underline{t}^{-2}2^d d^2(NL)^{2/d}N\log N)$ and depth $\mathcal{O}(d^2 L\log L)$. By omitting polynomial prefactors in d, we obtain that the deep ReLU network implementing the function $\bar{v}$ has width $\mathcal{O}(\underline{t}^{-2}2^d(NL)^{2/d}N\log N)$, depth $\mathcal{O}(L\log L)$, and size $\mathcal{O}(\underline{t}^{-2}4^d\ (NL)^{2/d}(N\log N)^2 L\log L)$. □

3.8.3 Error analysis of flow matching

In the section, we present proofs for error analyses of flow matching.

Basic error decomposition

We present the proof of Lemma 3.29.

Proof of Lemma 3.29. We follow the proof of Jiao et al. [2023a, Lemma 3.1]. Due to that v^* is the minimizer of $\mathcal{L}$, direct calculation implies

$$\mathbb{E}_{\mathbb{D}_n}\mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|\hat{v}_n(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2 = \mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(\hat{v}_n) - \mathcal{L}(v^*)].$$

Since $\hat{v}_n$ is the minimizer of the empirical risk, for any $v^\dagger \in \arg\inf_{v\in\mathcal{F}_n} \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|v(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2$, it holds that

$$\mathcal{L}_n(\hat{v}_n) - \mathcal{L}_n(v^*) \le \mathcal{L}_n(v^\dagger) - \mathcal{L}_n(v^*).$$

Taking expectations over $\mathbb{D}_n$ on both sides, it yields that

$$\mathbb{E}_{\mathbb{D}_n}[\mathcal{L}_n(\hat{v}_n) - \mathcal{L}(v^*)] \le \mathcal{L}(v^\dagger) - \mathcal{L}(v^*) = \inf_{v\in\mathcal{F}_n} \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|v(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2. \tag{3.41}$$

Using the inequality equation 3.41, we deduce that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{D}_n}\mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|\hat{v}_n(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2 = \mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(\hat{v}_n) - \mathcal{L}(v^*)] \\
\le\ &\mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(\hat{v}_n) - \mathcal{L}(v^*)] - 2\mathbb{E}_{\mathbb{D}_n}[\mathcal{L}_n(\hat{v}_n) - \mathcal{L}(v^*)] + 2\inf_{v\in\mathcal{F}_n} \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|v(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2 \\
\le\ &\mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(v^*) - 2\mathcal{L}_n(\hat{v}_n) + \mathcal{L}(\hat{v}_n)] + 2\inf_{v\in\mathcal{F}_n} \mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|v(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2.
\end{aligned}$$

This completes the proof. □

Truncation error

The truncation error is well controlled by the fast-decaying tail probability of $\mathsf{X}_t \sim p_t$. We bound the tail probability in Lemma 3.35 and the truncation error in Lemma 3.36. For a sub-Gaussian random variable X, we use $\|\mathsf{X}\|_{\psi_2}$ to denote its sub-Gaussian norm.

Proof of Lemma 3.35. Let $\mathsf{X}_t = [\mathsf{X}_t^1, \mathsf{X}_t^2, \cdots, \mathsf{X}_t^d]^\top$. Similarly, let $\mathsf{Z} = [\mathsf{Z}^1, \mathsf{Z}^2, \cdots, \mathsf{Z}^d]^\top$ and $\mathsf{X}_1 = [\mathsf{X}_1^1, \mathsf{X}_1^2, \cdots, \mathsf{X}_1^d]^\top$. By the general Hoeffding inequality [Vershynin, 2018, Theorem 2.6.3], for any $1 \le i \le d$, we bound the tail probability of X_t^i by

$$\mathbb{P}(|\mathsf{X}_t^i| > A) = \mathbb{P}(|(1-t)\mathsf{Z}^i + t\mathsf{X}_1^i| > A) \le 2\exp\left(-\frac{C_1 A^2}{K_1^2}\right),$$

where C_1 is a universal constant and $K_1 := \|\mathsf{Z}^1\|_{\psi_2} \vee \max_{1\le i\le d}\|\mathsf{X}_1^i\|_{\psi_2}$ with $\mathsf{Z}^1 \sim \gamma_1$. According to Remark 3.11, $K_1 \asymp \sqrt{C_{\mathrm{LSI}}}$ is finite with dependence on parameters in $\mathcal{S}_1$. By the union bound, it further yields

$$\mathbb{P}(\mathsf{X}_t \in \Omega_A^c) = \mathbb{P}(\exists 1 \le i \le d : |\mathsf{X}_t^i| > A) \le \sum_{i=1}^d \mathbb{P}(|\mathsf{X}_t^i| > A) \le 2d\exp\left(-\frac{C_2 A^2}{C_{\mathrm{LSI}}}\right),$$

where C_2 is a universal constant and C_{LSI} depends on parameters in $\mathcal{S}_1$. This tail probability bound holds uniformly for $t \in [0,1]$. □

Proof of Lemma 3.36. We decompose the truncation error by

$$\begin{aligned}\mathcal{E}_{\text{trunc}} &= \mathbb{E}_{(t,X_t)}\|[\bar{v}(t,X_t) - v^*(t,X_t)]\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2 \\ &\lesssim \underbrace{\mathbb{E}_{(t,X_t)}\|\bar{v}(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2}_{=:\mathcal{E}^1_{\text{trunc}}} + \underbrace{\mathbb{E}_{(t,X_t)}\|v^*(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2}_{=:\mathcal{E}^2_{\text{trunc}}}. \end{aligned} \tag{3.42}$$

First, we bound $\mathcal{E}^1_{\text{trunc}}$. For any $A > 0$ and $t \in [0, 1-\underline{t}]$, it holds that

$$\begin{aligned}\mathbb{E}_{X_t}\|\bar{v}(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2 &= \mathbb{E}_{X_t}[\|\bar{v}(t,X_t)\|_2^2\,\mathrm{Id}_{\Omega_A^c}(X_t)] \\ &\le \left(\mathbb{E}_{X_t}[\|\bar{v}(t,X_t)\|_2^4]\cdot\mathbb{P}(X_t\in\Omega_A^c)\right)^{1/2} \\ &\lesssim A^2\mathbb{P}(X_t\in\Omega_A^c)^{1/2}, \end{aligned} \tag{3.43}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the second inequality is due to $\|\bar{v}(t,x)\|_{L^\infty([0,1-\underline{t}]\times\mathbb{R}^d)} \lesssim A$ given in Theorem 3.31. Combining (3.20) in Lemma 3.35 and (3.43) above, it follows

$$\mathcal{E}^1_{\text{trunc}} = \mathbb{E}_{(t,X_t)}\|\bar{v}(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2 \lesssim \sqrt{d}A^2\exp\left(-\frac{C_3A^2}{C_{\text{LSI}}}\right), \tag{3.44}$$

where C_3 is a universal constant.

Then, we bound $\mathcal{E}^2_{\text{trunc}}$. Due to $v^*(t,x) = \mathbb{E}[X_1 - Z|X_t = x]$, it holds

$$\begin{aligned}\mathbb{E}_{X_t}\|v^*(t,X_t)\|_2^4 &= \mathbb{E}_{X_t}\|\mathbb{E}[X_1 - Z|X_t = x]\|_2^4 \\ &\le \mathbb{E}_{X_t}\mathbb{E}[\|X_1 - Z\|_2^4|X_t = x] \\ &\le \mathbb{E}[\|X_1 - Z\|_2^4] \\ &\le 8\mathbb{E}[\|Z\|_2^4] + 8\mathbb{E}[\|X_1\|_2^4], \end{aligned}$$

where the fourth moments in the last expression are finite by the property of the Gaussian distribution and the sub-Gaussian property of X_1. For any $A > 0$ and $t \in [0, 1-\underline{t}]$, we further bound $\mathbb{E}_{X_t}\|v^*(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2$ by

$$\begin{aligned}\mathbb{E}_{X_t}\|v^*(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2 &= \mathbb{E}_{X_t}[\|v^*(t,X_t)\|_2^2\,\mathrm{Id}_{\Omega_A^c}(X_t)] \\ &\le \left(\mathbb{E}_{X_t}\|[v^*(t,X_t)\|_2^4]\cdot\mathbb{P}(X_t\in\Omega_A^c)\right)^{1/2} \\ &\lesssim \mathbb{E}[\|X_1\|_2^4]^{1/2}\cdot\mathbb{P}(X_t\in\Omega_A^c)^{1/2}. \end{aligned} \tag{3.45}$$

Combining (3.20) in Lemma 3.35 and (3.45) above, it follows

$$\mathcal{E}^2_{\text{trunc}} = \mathbb{E}_{(t,X_t)}\|v^*(t,X_t)\,\mathrm{Id}_{\Omega_A^c}(X_t)\|_2^2 \lesssim \sqrt{d}\exp\left(-\frac{C_3A^2}{C_{\text{LSI}}}\right), \tag{3.46}$$

where we omit the dependence on the fourth moment of the target X_1.
Finally, combining (3.42), (3.44), and (3.46), we get

$$\mathcal{E}_{\text{trunc}} \lesssim \sqrt{d} A^2 \exp\left(-\frac{C_3 A^2}{C_{\text{LSI}}}\right),$$

where we omit the dependence on the fourth moment of the target X_1. This completes the proof. □

Stochastic error

The stochastic error is known as generalization error in statistical machine learning. In this part, we study the stochastic error of flow matching with techniques in empirical processes and present the proof of Lemma 3.38. Before that, we show necessary definitions from the content of empirical processes for establishing bounds of the stochastic error.

Definition 3.60 (Uniform and empirical covering numbers)**.** Given the samples $\mathbb{X}_n := \{\mathsf{X}_i\}_{i=1}^n$, we define the empirical L^∞ pseudometric $\|\cdot\|_{L^\infty(\mathbb{X}_n)}$ on the samples $\mathbb{X}_n$ by

$$\|f\|_{L^\infty(\mathbb{X}_n)} := \max_{1\le i\le n} |f(\mathsf{X}_i)|.$$

A set $\mathcal{F}_\delta$ is called an empirical L^∞ δ-cover of the function class $\mathcal{F}$ on the samples $\mathbb{X}_n$ if for each $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}_\delta$ such that $\|f - f'\|_{L^\infty(\mathbb{X}_n)} \le \delta$. Furthermore,

$$\mathcal{N}_\infty(\delta, \mathcal{F}, \mathbb{X}_n) := \inf\left\{|\mathcal{F}_\delta| : \mathcal{F}_\delta \text{ is an empirical } L^\infty \text{ } \delta\text{-cover of } \mathcal{F} \text{ on } \mathbb{X}_n\right\}$$

is called the empirical L^∞ δ-covering number of $\mathcal{F}$ on $\mathbb{X}_n$. Given n, the largest L^∞ δ-covering number over samples $\mathbb{X}_n$ is referred to as the uniform L^∞ δ-covering number $\mathcal{N}_\infty(\delta, \mathcal{F}, n) := \sup_{\mathbb{X}_n} \mathcal{N}_\infty(\delta, \mathcal{F}, \mathbb{X}_n)$.

Definition 3.61. Let $\mathcal{F}$ be a class of functions from a set $\mathcal{Z}$ to $\mathbb{R}$. A set $\{\mathsf{Z}_1, \cdots, \mathsf{Z}_m\} \subset \mathcal{X}$ is said to be shattered by $\mathcal{F}$ if there exist $t_1, t_2, \cdots, t_m \in \mathbb{R}$ such that, for each $b \in \{0,1\}^m$, there exist a function $f_b \in \mathcal{F}$ satisfying $\operatorname{sgn}(f_b(\mathsf{Z}_i) - t_i) = b_i$ for $1 \le i \le m$. We say that the threshold values $t_1, t_2, \cdots, t_m$ witness the shattering.

Definition 3.62 (Pseudo-dimension)**.** Let $\mathcal{F}$ be a class of functions from a set Ω to $\mathbb{R}$. The pseudo-dimension of $\mathcal{F}$, denoted by $\operatorname{Pdim}(\mathcal{F})$, is the maximum cardinality of a subset of Ω shattered by $\mathcal{F}$.

Proof of Lemma 3.38. Let $\mathbb{D}_n = \{\mathsf{S}_i := (\mathsf{Z}_i, \mathsf{X}_{1,i}, \mathsf{t}_i)\}_{i=1}^n$ be a random sample from the distribution of $\mathsf{Z}, \mathsf{X}_1, \mathsf{t}$ and $\mathbb{D}'_n := \{\mathsf{S}'_i := (\mathsf{Z}'_i, \mathsf{X}'_{1,i}, \mathsf{t}'_i)\}_{i=1}^n$ be another ghost sample independent of $\mathbb{D}_n$. We denote that $\mathsf{X}_{\mathsf{t}_i} := (1-\mathsf{t}_i)\mathsf{Z}_i + \mathsf{t}_i\mathsf{X}_{1,i}$, $\mathsf{X}'_{\mathsf{t}_i} := (1-\mathsf{t}'_i)\mathsf{Z}'_i + \mathsf{t}'_i\mathsf{X}'_{1,i}$, $\mathsf{Y}_i := \mathsf{X}_{1,i} - \mathsf{Z}_i$, and $\mathsf{Y}'_i := \mathsf{X}'_{1,i} - \mathsf{Z}'_i$. Define $\mathcal{D}(v, \mathsf{S}_i) := \|v(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathsf{Y}_i\|_2^2 - \|v^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathsf{Y}_i\|_2^2$ for any $v \in \mathcal{F}_n$ and S_i. Notice that

$$\mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(v^*) - 2\mathcal{L}_n(\hat{v}_n) + \mathcal{L}(\hat{v}_n)] = \mathbb{E}_{\mathbb{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}_{\mathbb{D}'_n}\mathcal{D}(\hat{v}_n, \mathsf{S}'_i) - 2\mathcal{D}(\hat{v}_n, \mathsf{S}_i)\right)\right]. \quad (3.47)$$

It is clear that the right-hand side of (3.47) defines an asymmetric empirical process. Let $\mathcal{G}(v, \mathsf{S}_i) := \mathbb{E}_{\mathbb{D}'_n}\mathcal{D}(v, \mathsf{S}'_i) - 2\mathcal{D}(v, \mathsf{S}_i)$ for any $v \in \mathcal{F}_n$. Then we have

$$\mathbb{E}_{\mathbb{D}_n}[\mathcal{L}(v^*) - 2\mathcal{L}_n(\hat{v}_n) + \mathcal{L}(\hat{v}_n)] = \mathbb{E}_{\mathbb{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{G}(\hat{v}_n, \mathsf{S}_i)\right].$$

Let $B_n \geq \mathsf{B} \geq B \geq 1$ be a positive number that may depend on the sample size n. We construct a clipping function $\mathcal{C}_{B_n}$ at level B_n following the definition of clipping functions in Lemma 3.58. Let $v_{B_n}(t,x) := \mathbb{E}[\mathcal{C}_{B_n}(\mathsf{Y})|\mathsf{X}_t = x]$ be the regression function of the truncated Y. Similar to the definitions of $\mathcal{D}(v, \mathsf{S}_i)$ and $\mathcal{G}(v, \mathsf{S}_i)$, we define $\mathcal{D}_{B_n}(v, \mathsf{S}_i) := \|v(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathcal{C}_{B_n}(\mathsf{Y}_i)\|_2^2 - \|v_{B_n}(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathcal{C}_{B_n}(\mathsf{Y}_i)\|_2^2$ and $\mathcal{G}_{B_n}(v, \mathsf{S}_i) := \mathbb{E}_{\mathbb{D}'_n}\mathcal{D}_{B_n}(v, \mathsf{S}'_i) - 2\mathcal{D}_{B_n}(v, \mathsf{S}_i)$. Then for any $v \in \mathcal{F}_n$ we have

$$\begin{aligned}
&|\mathcal{D}(v, \mathsf{S}_i) - \mathcal{D}_{B_n}(v, \mathsf{S}_i)| \\
=&\Big|2\langle v(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - v^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}), \mathcal{C}_{B_n}(\mathsf{Y}_i) - \mathsf{Y}_i\rangle \\
&+ \|v_{B_n}(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathcal{C}_{B_n}(\mathsf{Y}_i)\|_2^2 - \|v^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - \mathcal{C}_{B_n}(\mathsf{Y}_i)\|_2^2\Big| \\
\leq& 2\Big|\Big\langle v(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - v^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}), \mathcal{C}_{B_n}(\mathsf{Y}_i) - \mathsf{Y}_i\Big\rangle\Big| \\
&+ \Big|\Big\langle v_{B_n}(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - v^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}), v_{B_n}(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) + v^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - 2\mathcal{C}_{B_n}(\mathsf{Y}_i)\Big\rangle\Big|.
\end{aligned}$$

By considering coordinate-wise scalar expressions of the risks, we get

$$\begin{aligned}
&|\mathcal{D}(v, \mathsf{S}_i) - \mathcal{D}_{B_n}(v, \mathsf{S}_i)| \\
\leq& \sum_{j=1}^d \Big\{2\Big|[v_j(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - v_j^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i})] \cdot [(\mathcal{C}_{B_n}(\mathsf{Y}_i))_j - (\mathsf{Y}_i)_j]\Big| \\
&+ \Big|[(v_{B_n})_j(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - v_j^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i})] \cdot [(v_{B_n})_j(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) + v_j^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - 2(\mathcal{C}_{B_n}(\mathsf{Y}_i))_j]\Big|\Big\} \\
\leq& \sum_{j=1}^d \Big\{4\mathsf{B}\Big|(\mathcal{C}_{B_n}(\mathsf{Y}_i))_j - (\mathsf{Y}_i)_j\Big| + 4B_n\Big|(v_{B_n})_j(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - v_j^*(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i})\Big|\Big\} \\
\leq& \sum_{j=1}^d \Big\{4\mathsf{B}\Big|(\mathcal{C}_{B_n}(\mathsf{Y}_i))_j - (\mathsf{Y}_i)_j\Big| + 4B_n\mathbb{E}[|(\mathcal{C}_{B_n}(\mathsf{Y}_i))_j - (\mathsf{Y}_i)_j|\big|\mathsf{X}_{t_i} = \mathsf{X}_{\mathsf{t}_i}]\Big\}.
\end{aligned}$$

Note that $|(\mathcal{C}_{B_n}(\mathsf{Y}_i))_j - (\mathsf{Y}_i)_j| \le |(\mathsf{Y}_i)_j| \mathrm{Id}_{\{|(\mathsf{Y}_i)_j| \ge B_n\}}$ and $B_n \ge \mathsf{B}$. Then it follows that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{D}_n} |\mathcal{D}(v, \mathsf{S}_i) - \mathcal{D}_{B_n}(v, \mathsf{S}_i)| \\
\le &\mathbb{E}_{\mathbb{D}_n} \Bigg[\sum_{j=1}^{d} \left\{ 4\mathsf{B} |(\mathsf{Y}_i)_j| \mathrm{Id}_{\{|(\mathsf{Y}_i)_j| \ge B_n\}} + 4B_n \mathbb{E}[|(\mathsf{Y}_i)_j| \mathrm{Id}_{\{|(\mathsf{Y}_i)_j| \ge B_n\}} \big| \mathsf{X}_{t_i} = \mathsf{X}_{\mathsf{t}_i}] \right\} \Bigg] \\
\le &\sum_{j=1}^{d} 8B_n \mathbb{E}_{\mathbb{D}_n} \Big[|(\mathsf{Y}_i)_j| \mathrm{Id}_{\{|(\mathsf{Y}_i)_j| \ge B_n\}} \Big] \\
\le &\sum_{j=1}^{d} 8B_n \mathbb{E}_{\mathbb{D}_n} \Big[|(\mathsf{Y}_i)_j|^2 \Big] \cdot \mathbb{P}\{|(\mathsf{Y}_i)_j| \ge B_n\}.
\end{aligned}$$

By Assumption 3.8 and Remark 3.11, the law of $\mathsf{Y} = \mathsf{X} - \mathsf{Z}$ is sub-Gaussian. Then there exist two constants K_1 and K_2 such that for each $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, d$,

$$\mathbb{P}\{|(\mathsf{Y}_i)_j| \ge B_n\} \le 2 \exp\left(-\frac{B_n^2}{K_1^2} \right), \quad \mathbb{E}_{\mathbb{D}_n} \Big[|(\mathsf{Y}_i)_j|^2 \Big] \le 2K_2^2.$$

The bounds above further imply that

$$\begin{aligned}
\mathbb{E}_{\mathbb{D}_n} \mathcal{D}(v, \mathsf{S}_i) \le &\mathbb{E}_{\mathbb{D}_n} \mathcal{D}_{B_n}(v, \mathsf{S}_i) + \sum_{j=1}^{d} 8B_n \mathbb{E}_{\mathbb{D}_n} \Big[|(\mathsf{Y}_i)_j|^2 \Big] \mathbb{P}(|(\mathsf{Y}_i)_j| \ge B_n) \\
\le &\mathbb{E}_{\mathbb{D}_n} \mathcal{D}_{B_n}(v, \mathsf{S}_i) + 32 d B_n K_2^2 \exp\left(-\frac{B_n^2}{K_1^2} \right).
\end{aligned}$$

Therefore, we conclude that

$$\mathbb{E}_{\mathbb{D}_n} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{G}(\hat{v}_n, \mathsf{S}_i) \right] \le \mathbb{E}_{\mathbb{D}_n} \left[\frac{1}{n} \sum_{i=1}^{n} \mathcal{G}_{B_n}(\hat{v}_n, \mathsf{S}_i) \right] + K_3 B_n \exp\left(-\frac{B_n^2}{K_1^2} \right), \tag{3.48}$$

where the constant K_3 does not depend on n and B_n.

Next, we consider bounding a tail probability of the empirical process. Before proceeding, we define $(\mathcal{D}_{B_n})_j(v, \mathsf{S}_i) := [v_j(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - (\mathcal{C}_{B_n}(\mathsf{Y}_i))_j]^2 - [(v_{B_n})_j(\mathsf{t}_i, \mathsf{X}_{\mathsf{t}_i}) - (\mathcal{C}_{B_n}(\mathsf{Y}_i))_j]^2$ for any $j \in \{1, 2, \cdots, d\}$. It is clear that $\mathcal{D}_{B_n}(v, \mathsf{S}_i) = \sum_{j=1}^{d} (\mathcal{D}_{B_n})_j(v, \mathsf{S}_i)$, and we have the following tail probability bounds

$$\begin{aligned}
&\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{G}_{B_n}(\hat{v}_n, \mathsf{S}_i) > t \right\} \\
\le &\mathbb{P}\left\{ \exists v \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^{n} \mathcal{G}_{B_n}(v, \mathsf{S}_i) > t \right\} \\
= &\mathbb{P}\left\{ \exists v \in \mathcal{F}_n : \mathbb{E}_{\mathbb{D}_n'} \mathcal{D}_{B_n}(v, \mathsf{S}_i') - \frac{2}{n} \sum_{i=1}^{n} \mathcal{D}_{B_n}(v, \mathsf{S}_i) > t \right\} \\
\le &\mathbb{P}\left\{ \exists v \in \mathcal{F}_n \text{ and } \exists 1 \le j \le d : \mathbb{E}_{\mathbb{D}_n'} (\mathcal{D}_{B_n})_j(v, \mathsf{S}_i') - \frac{2}{n} \sum_{i=1}^{n} (\mathcal{D}_{B_n})_j(v, \mathsf{S}_i) > \frac{t}{d} \right\}.
\end{aligned}$$

Note that $|(\mathcal{C}_{B_n}(\mathsf{Y}))_j| \le B_n$, $\|(v_{B_n})_j\|_\infty \le B_n$, and $B_n \ge \mathsf{B} \ge 1$. By Theorem 11.4 of Györfi et al. [2002] and letting $\epsilon = 1/2, \alpha = \beta = t/(2d)$ in Györfi et al. [2002, Theorem 11.4], it yields that for each $n \ge 1$,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \mathcal{G}_{B_n}(\hat{v}_n, \mathsf{S}_i) > t\right\} \\
\le&\mathbb{P}\left\{\exists v \in \mathcal{F}_n \text{ and } \exists 1 \le j \le d : \mathbb{E}_{\mathbb{D}_n'}(\mathcal{D}_{B_n})_j(v, \mathsf{S}_i') - \frac{2}{n}\sum_{i=1}^n (\mathcal{D}_{B_n})_j(v, \mathsf{S}_i) > \frac{t}{d}\right\} \\
\le&14\mathcal{N}_\infty(t/(80dB_n), \mathcal{F}_n, n)\exp\left(-\frac{tn}{5136dB_n^4}\right). \qquad (3.49)
\end{aligned}$$

Then we use the tail probability bound (3.49) to bound the stochastic error. For any $\alpha_n > 0$,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{G}_{B_n}(\hat{v}_n, \mathsf{S}_i)\right] \\
\le&\alpha_n + \int_{\alpha_n}^\infty \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \mathcal{G}_{B_n}(\hat{v}_n, \mathsf{S}_i) > t\right\}\mathrm{d}t \\
\le&\alpha_n + \int_{\alpha_n}^\infty 14\mathcal{N}_\infty(t/(80dB_n), \mathcal{F}_n, n)\exp\left(-\frac{tn}{5136dB_n^4}\right)\mathrm{d}t \\
\le&\alpha_n + \int_{\alpha_n}^\infty 14\mathcal{N}_\infty(\alpha_n/(80dB_n), \mathcal{F}_n, n)\exp\left(-\frac{tn}{5136dB_n^4}\right)\mathrm{d}t \\
\le&\alpha_n + 14\mathcal{N}_\infty(\alpha_n/(80dB_n), \mathcal{F}_n, n)\exp\left(-\frac{\alpha_n n}{5136dB_n^4}\right)\frac{5136dB_n^4}{n}.
\end{aligned}$$

By choosing $\alpha_n = \log(14\mathcal{N}_\infty(1/n, \mathcal{F}_n, n)) \cdot 5136dB_n^4/n$ and noticing that $\alpha_n/(80dB_n) \ge 1/n$ and $\mathcal{N}_\infty(1/n, \mathcal{F}_n, n) \ge \mathcal{N}_\infty(\alpha_n/(80dB_n), \mathcal{F}_n, n)$, we obtain that

$$\mathbb{E}_{\mathbb{D}_n}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{G}_{B_n}(\hat{v}_n, \mathsf{S}_i)\right] \le \frac{5136dB_n^4(\log(14\mathcal{N}_\infty(1/n, \mathcal{F}_n, n)) + 1)}{n}. \qquad (3.50)$$

Setting $B_n \asymp \mathsf{B}\log n$ and combining (3.48) and (3.50), the stochastic error is upper bounded by the covering number of the hypothesis class $\mathcal{F}_n \subseteq \mathcal{NN}(\mathsf{S}, \mathsf{W}, \mathsf{D}, \mathsf{B}, d+1, d)$ as

$$\mathcal{E}_{\text{stoc}} \lesssim \frac{d}{n}(\log n)^4 \mathsf{B}^4 \log \mathcal{N}_\infty(1/n, \mathcal{F}_n, n) \qquad (3.51)$$

where we omit a constant not depending on n or B. By the relationship between the uniform covering number and the pseudo-dimension of the deep ReLU network class

$\mathcal{F}_n$ [Anthony and Bartlett, 1999, Theorem 12.2], it yields that for $n \geq \mathrm{Pdim}(\mathcal{F}_n)$,

$$\mathcal{N}_\infty(1/n, \mathcal{F}_n, n) \leq \left(\frac{2e\mathsf{B}n^2}{\mathrm{Pdim}(\mathcal{F}_n)}\right)^{\mathrm{Pdim}(\mathcal{F}_n)}, \tag{3.52}$$

where $\mathrm{Pdim}(\mathcal{F}_n)$ denotes the pseudo-dimension of $\mathcal{F}_n$. By Theorems 3 and 7 of Bartlett et al. [2019], the pseudo-dimension of the deep ReLU network class $\mathcal{F}_n \subseteq \mathcal{NN}(\mathsf{S}, \mathsf{W}, \mathsf{D}, \mathsf{B}, d+1, d)$ satisfies

$$\mathsf{SD}\log(\mathsf{S}/\mathsf{D}) \lesssim \mathrm{Pdim}(\mathcal{F}_n) \lesssim \mathsf{SD}\log(\mathsf{S}). \tag{3.53}$$

Combining (3.51), (3.52), (3.53), and $\mathsf{B} \lesssim A$, we complete the proof by showing

$$\mathcal{E}_{\mathrm{stoc}} \lesssim \frac{d}{n}(\log n)^4 A^4 \mathsf{SD}\log(\mathsf{S})\log(An^2).$$

□

Balancing errors

We present the proof of Theorem 3.40 for balancing the approximation error and the stochastic error of flow matching.

Proof of Theorem 3.40. According to Corollary 3.37 and Corollary 3.39, it holds that

$$\mathcal{E}_{\mathrm{stoc}} \lesssim \frac{1}{n}A^4\underline{t}^{-2}(NL)^{2+2/d}, \quad \mathcal{E}_{\mathrm{appr}} \lesssim A^2(NL)^{-4/d} + A^2\exp(-C_3A^2/C_{\mathrm{LSI}})$$

by omitting a polylogarithmic prefactor in N, L, A, n, a prefactor in $\log(1/\underline{t})$, and a prefactor in $d, \kappa, \beta, \sigma, R$. Let $NL \asymp (n\underline{t}^2)^{d/(2d+6)}$ and $A \asymp \log(\log n)$. Then by Lemma 3.29, it holds that

$$\mathbb{E}_{\mathbb{D}_n}\mathbb{E}_{(\mathsf{t},\mathsf{X}_\mathsf{t})}\|\hat{v}_n(\mathsf{t},\mathsf{X}_\mathsf{t}) - v^*(\mathsf{t},\mathsf{X}_\mathsf{t})\|_2^2 \leq \mathcal{E}_{\mathrm{stoc}} + 2\mathcal{E}_{\mathrm{appr}} \lesssim (n\underline{t}^2)^{-2/(d+3)}$$

by omitting a polylogarithmic prefactor in n, a prefactor in $\log(1/\underline{t})$, and a prefactor in $d, \kappa, \beta, \sigma, R$. □

3.8.4 Distribution estimation errors

In the section, we provide proofs for bounding distribution estimation errors. The discretization error is bounded in Lemma 3.19.

Proof of Lemma 3.19. By the definition of Wasserstein-2 distance, it holds

$$\mathcal{W}_2^2(\hat{p}_t, \tilde{p}_t) \le \int_{\mathbb{R}^d} \|\hat{X}_t(x) - \tilde{X}_t(x)\|_2^2 p_0(x)\mathrm{d}x =: E_t.$$

It suffices to consider the propagation of error E_t in time $t \in [0, 1-\underline{t}]$. Recall that $(\hat{X}_t)_{t\in[0,1-\underline{t}]}$ is the linear interpolation of $(\hat{X}_{t_k})_{0\le k\le K}$, thus it is piecewise linear over $[0, 1-\underline{t}]$. To ease the arguments, we consider the dynamics of E_t over each time subinterval $[t_{k-1}, t_k]$ for $1 \le k \le K$. For $t \in [t_{k-1}, t_k]$, it holds that

$$\begin{aligned}
\frac{\mathrm{d}E_t}{\mathrm{d}t} &= \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \tilde{X}_t(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \\
&= \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \hat{X}_{t_{k-1}}(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \qquad (3.54) \\
&\quad + \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \hat{X}_t(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \qquad (3.55) \\
&\quad + \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \hat{X}_t(x)) - \hat{v}_n(t, \tilde{X}_t(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \qquad (3.56)
\end{aligned}$$

For term (3.54), the basic inequality $2\langle a, b\rangle \le \|a\|_2^2 + \|b\|_2^2$ and the fact that $\hat{v}_n$ is L_t-Lipschitz continuous in t imply that

$$\begin{aligned}
&\int_{\mathbb{R}^d} 2\langle \hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \hat{X}_{t_{k-1}}(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \\
\le &\int_{\mathbb{R}^d} \|\hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \hat{X}_{t_{k-1}}(x))\|_2^2 p_0(x)\mathrm{d}x \\
&+ \int_{\mathbb{R}^d} \|\hat{X}_t(x) - \tilde{X}_t(x)\|_2^2 p_0(x)\mathrm{d}x \\
\le &\, d\mathsf{L}_t^2 (t - t_{k-1})^2 + E_t. \qquad (3.57)
\end{aligned}$$

Note that $\hat{X}_t(x) = \hat{X}_{t_{k-1}}(x) + (t - t_{k-1})\hat{v}_n(t_{k-1}, \hat{X}_{t_{k-1}}(x))$. For term (3.55), we use $2\langle a, b\rangle \le$

$\|a\|_2^2 + \|b\|_2^2$ and the fact that $\hat{v}_n$ is L_x-Lipschitz continuous in x to deduce that

$$\begin{aligned}
&\int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \hat{X}_t(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \\
\leq &\int_{\mathbb{R}^d} \|\hat{v}_n(t, \hat{X}_{t_{k-1}}(x)) - \hat{v}_n(t, \hat{X}_t(x))\|_2^2 p_0(x)\mathrm{d}x \\
&+ \int_{\mathbb{R}^d} \|\hat{X}_t(x) - \tilde{X}_t(x)\|_2^2 p_0(x)\mathrm{d}x \\
\leq &\, d\mathsf{L}_x^2 (t - t_{k-1})^2 \|\hat{v}_n\|^2_{L^\infty([0,1-\underline{t}]\times\mathbb{R}^d)} + E_t \\
\leq &\, d\mathsf{L}_x^2 \mathsf{B}^2 (t - t_{k-1})^2 + E_t. \qquad (3.58)
\end{aligned}$$

For term (3.56), by the Cauchy-Schwartz inequality and the fact that $\hat{v}_n$ is L_x-Lipschitz continuous in x, we obtain

$$\int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \hat{X}_t(x)) - \hat{v}_n(t, \tilde{X}_t(x)), \hat{X}_t(x) - \tilde{X}_t(x)\rangle p_0(x)\mathrm{d}x \leq 2\mathsf{L}_x E_t. \qquad (3.59)$$

Combining (3.57), (3.58), and (3.59), we obtain

$$\frac{\mathrm{d}E_t}{\mathrm{d}t} \leq 2(\mathsf{L}_x + 1)E_t + d(\mathsf{L}_x^2\mathsf{B}^2 + \mathsf{L}_t^2)(t - t_{k-1})^2 \quad \text{for } t \in [t_{k-1}, t_k].$$

By Grönwall's inequality, it further yields

$$e^{-2(\mathsf{L}_x+1)t_k} E_{t_k} - e^{-2(\mathsf{L}_x+1)t_{k-1}} E_{t_{k-1}} \leq \frac{1}{3} d(\mathsf{L}_x^2\mathsf{B}^2 + \mathsf{L}_t^2)(t_k - t_{k-1})^3.$$

Taking sum over $k = 1, 2, \cdots, K$ and letting $t_K = 1 - \underline{t}$, we obtain

$$E_{1-\underline{t}} \leq \frac{1}{3} d e^{2(\mathsf{L}_x+1)(1-\underline{t})} (\mathsf{L}_x^2\mathsf{B}^2 + \mathsf{L}_t^2) \sum\nolimits_{k=1}^{K} (t_k - t_{k-1})^3.$$

Let $\Upsilon \equiv t_k - t_{k-1}$ for $k = 1, 2, \cdots, K$. It implies that

$$\mathcal{W}_2(\hat{p}_{1-\underline{t}}, \tilde{p}_{1-\underline{t}}) = \mathcal{O}\Big(\sqrt{d} e^{\mathsf{L}_x} (\mathsf{L}_x \mathsf{B} + \mathsf{L}_t) \Upsilon\Big).$$

This completes the proof. □

The error due to velocity estimation is bounded in Lemma 3.21.

Proof of Lemma 3.21. The proof idea is similar to that of Albergo and Vanden-Eijnden [2023, Proposition 3]. By the definition of the Wasserstein-2 distance, it holds

$$\mathcal{W}_2^2(\tilde{p}_t, p_t) \le \int_{\mathbb{R}^d} \|\tilde{X}_t(x) - X_t(x)\|_2^2 p_0(x)\mathrm{d}x =: R_t,$$

for any $t \in [0, 1-\underline{t}]$. By (3.5) and (3.12), it follows that

$$\begin{aligned}\frac{\mathrm{d}R_t}{\mathrm{d}t} &= \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \tilde{X}_t(x)) - v^*(t, X_t(x)), \tilde{X}_t(x) - X_t(x)\rangle p_0(x)\mathrm{d}x \\ &= \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \tilde{X}_t(x)) - \hat{v}_n(t, X_t(x)), \tilde{X}_t(x) - X_t(x)\rangle p_0(x)\mathrm{d}x \qquad (3.60) \\ &\quad + \int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, X_t(x)) - v^*(t, X_t(x)), \tilde{X}_t(x) - X_t(x)\rangle p_0(x)\mathrm{d}x. \qquad (3.61)\end{aligned}$$

For term (3.60), the fact that $\hat{v}_n$ is L_x-Lipschitz continuous in x imply that

$$\int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, \tilde{X}_t(x)) - \hat{v}_n(t, X_t(x)), \tilde{X}_t(x) - X_t(x)\rangle p_0(x)\mathrm{d}x \le 2\mathsf{L}_x R_t.$$

For term (3.61), the basic inequality $2\langle a, b\rangle \le \|a\|_2^2 + \|b\|_2^2$ imply that

$$\begin{aligned}&\int_{\mathbb{R}^d} 2\langle \hat{v}_n(t, X_t(x)) - v^*(t, X_t(x)), \tilde{X}_t(x) - X_t(x)\rangle p_0(x)\mathrm{d}x \\ &\le R_t + \mathbb{E}_{\mathsf{X}_t \sim p_t}\|\hat{v}_n(t, \mathsf{X}_t) - v^*(t, \mathsf{X}_t)\|_2^2.\end{aligned}$$

Therefore, we have

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} \le (2\mathsf{L}_x + 1)R_t + \mathbb{E}_{\mathsf{X}_t \sim p_t}\|\hat{v}_n(t, \mathsf{X}_t) - v^*(t, \mathsf{X}_t)\|_2^2.$$

By Grönwall's inequality, it further yields

$$R_{1-\underline{t}} \le \exp(2\mathsf{L}_x + 1)\mathbb{E}_{(\mathsf{t}, \mathsf{X}_\mathsf{t})}\|\hat{v}_n(\mathsf{t}, \mathsf{X}_\mathsf{t}) - v^*(\mathsf{t}, \mathsf{X}_\mathsf{t})\|_2^2.$$

We complete the proof by noting that $\mathcal{W}_2^2(\tilde{p}_{1-\underline{t}}, p_{1-\underline{t}}) \le R_{1-\underline{t}}$. □

The early stopping error is bounded in Lemma 3.22.

Proof of Lemma 3.22. The proof is a basic calculation.

$$\begin{aligned}\mathcal{W}_2^2(p_{1-\underline{t}}, p_1) &\le \mathbb{E}[\|\mathsf{X}_{1-\underline{t}} - \mathsf{X}_1\|_2^2] = \mathbb{E}[\|\underline{t}(\mathsf{Z} - \mathsf{X}_1)\|_2^2] \\ &= \underline{t}^2\left(\mathbb{E}[\|\mathsf{Z}\|_2^2] + \mathbb{E}[\|\mathsf{X}_1\|_2^2]\right) \lesssim \underline{t}^2,\end{aligned}$$

where we omit a polynomial prefactor in d, $\mathbb{E}[\|\mathsf{X}_1\|_2^2]$. We complete the proof by taking square roots of both sides. □

Proof of Theorem 3.23. Combining Eq. (3.14), Lemmas 3.19, 3.21, 3.22, and Theorem 3.40, it yields

$$\mathbb{E}_{\mathbb{D}_n}\mathcal{W}_2(\hat{p}_{1-\underline{t}}, p_1) \lesssim (n\underline{t}^2)^{-1/(d+3)} + e^{\mathsf{L}_x}(\mathsf{L}_x\mathsf{B} + \mathsf{L}_t)\Upsilon + \underline{t}.$$

Let $\underline{t} \asymp n^{-1/(d+5)}$, $A \asymp \log(\log n)$, and $\Upsilon = \mathcal{O}(n^{-3/(d+5)})$. Then it implies

$$\mathbb{E}_{\mathbb{D}_n}\mathcal{W}_2(\hat{p}_{1-\underline{t}}, p_1) \lesssim \underline{t} \vee (n\underline{t}^2)^{-1/(d+3)} \lesssim n^{-1/(d+5)},$$

where we omit a prefactor scaling polynomially in $\log n$ and a prefactor with dependence on parameters in $\mathcal{S}_2$. This completes the proof. □

3.8.5 Supporting definitions and lemmas

Sobolev spaces are widely studied in the context of functional analysis and partial differential equations. For ease of reference, we collect several definitions and existing results on Sobolev spaces that assist our proof. For a thorough treatment of Sobolev spaces, the interested reader is referred to Adams and Fournier [2003], Evans [2010]. Moreover, we present some results on polynomial approximation theory in Sobolev spaces that are developed in the classical monograph on the finite element methods [Brenner and Scott, 2008]. In the sequel, let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ denote an open subset of $\mathbb{R}^d$. We denote by $L^\infty(\Omega)$ the standard Lebesgue space on Ω with L^∞ norm.

Sobolev spaces

We list some definitions for defining Sobolev spaces.

Definition 3.63 (Sobolev space)**.** Let $n \in \mathbb{N}_0$. Then the Sobolev space $W^{n,\infty}(\Omega)$ is defined by

$$W^{n,\infty}(\Omega) := \{f \in L^\infty(\Omega) : D^\alpha f \in L^\infty(\Omega) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } \|\alpha\|_1 \le n\}.$$

Moreover, for any $f \in W^{n,\infty}(\Omega)$, we define the Sobolev norm $\|\cdot\|_{W^{n,\infty}(\Omega)}$ by

$$\|f\|_{W^{n,\infty}(\Omega)} := \max_{0 \le \|\alpha\|_1 \le n} \|D^\alpha f\|_{L^\infty(\Omega)}.$$

Definition 3.64 (Sobolev semi-norm)**.** Let $n, k \in \mathbb{N}_0$ with $k \le n$. For any $f \in W^{n,\infty}(\Omega)$, we define the Sobolev semi-norm $|\cdot|_{W^{k,\infty}(\Omega)}$ by

$$|f|_{W^{k,\infty}(\Omega)} := \max_{\|\alpha\|_1 = k} \|D^\alpha f\|_{L^\infty(\Omega)}.$$

Definition 3.65 (Vector-valued Sobolev space)**.** Let $n, k \in \mathbb{N}_0$ with $k \leq n$, and $m \in \mathbb{N}$. Then the vector-valued Sobolev space $W^{n,\infty}(\Omega; \mathbb{R}^m)$ is defined by

$$W^{n,\infty}(\Omega; \mathbb{R}^m) := \{(f_1, f_2, \cdots, f_m) : f_i \in W^{n,\infty}(\Omega), 1 \leq i \leq m\}$$

Moreover, the Sobolev norm $\|\cdot\|_{W^{n,\infty}(\Omega;\mathbb{R}^m)}$ is defined by

$$\|f\|_{W^{n,\infty}(\Omega;\mathbb{R}^m)} := \max_{1\leq i\leq m} \|f_i\|_{W^{n,\infty}(\Omega)},$$

and the Sobolev semi-norm $|\cdot|_{W^{n,\infty}(\Omega;\mathbb{R}^m)}$ is defined by

$$|f|_{W^{k,\infty}(\Omega;\mathbb{R}^m)} := \max_{1\leq i\leq m} |f_i|_{W^{k,\infty}(\Omega)}.$$

Averaged Taylor polynomials

The following definitions and lemmas on averaged Taylor polynomials are collected from Chapter 4 of Brenner and Scott [2008].

Definition 3.66 (Averaged Taylor polynomials)**.** Let $\Omega \subset \mathbb{R}^d$ be a bounded, open subset and $f \in W^{m-1,\infty}(\Omega)$ for some $m \in \mathbb{N}$, and let $x_0 \in \Omega, r > 0, B := \mathbb{B}^d(x_0, r, \|\cdot\|_2)$ with its closure $\bar{B}$ compact in Ω. The Taylor polynomial of order m of f averaged over B is defined as

$$Q^m f(x) := \int_B T_y^m f(x)\phi(y)\mathrm{d}y,$$

where

$$T_y^m f(x) := \sum_{\|\alpha\|_1<m} \frac{1}{\alpha!} D^\alpha f(y)(x-y)^\alpha,$$

and ϕ is an arbitrary cut-off function supported in $\bar{B}$ being infinitely differentiable, that is, $\phi \in C^\infty(\mathbb{R}^d)$ with $\mathrm{supp}(\phi) = \bar{B}$ and $\int_{\mathbb{R}^d} \phi(x)\mathrm{d}x = 1$.

Example 3.67. Let $\psi(x)$ be defined by

$$\psi(x) := \begin{cases} \exp\{-1/(1-(\|x-x_0\|_2/r)^2)\}, & \text{if } \|x-x_0\|_2 < r, \\ 0, & \text{if } \|x-x_0\|_2 \geq r, \end{cases}$$

and let $c = \int_{\mathbb{R}^d} \psi(x)\mathrm{d}x$ with $c > 0$, then $\phi(x) = \psi(x)/c$ is an example of the cut-off function on the ball $B := \mathbb{B}^d(x_0, r, \|\cdot\|_2)$. Moreover, it holds that $\|\phi\|_{L^\infty(B)} \leq C(d)r^{-d}$ where $C(d) > 0$ is a constant in d.

Example 3.68. Let $\phi(x)$ be defined by

$$\phi(x) = \begin{cases} \pi^{-d/2}\Gamma(d/2+1)r^{-d}, & \text{if } \|x - x_0\|_2 < r, \\ 0, & \text{if } \|x - x_0\|_2 \geq r, \end{cases}$$

then $\phi(x)$ is another example of the cut-off function where ϕ puts constant weight over the ball $B := \mathbb{B}^d(x_0, r, \|\cdot\|_2)$.

Lemma 3.69 (Lemma B.9 in Gühring et al. [2020]). *Let $\Omega \subset \mathbb{R}^d$ be a bounded, open subset and $f \in W^{m-1,\infty}(\Omega)$ for some $m \in \mathbb{N}$, and let $x_0 \in \Omega, r > 0, B := \mathbb{B}^d(x_0, r, \|\cdot\|_2)$ with its closure $\bar{B}$ compact in Ω. The Taylor polynomial of order m of f averaged over B denoted by $Q^m f(x)$ is a polynomial of degree less than m in x.*

Definition 3.70 (Star-shaped set)**.** Let $\Omega, B \subset \mathbb{R}^d$. We say Ω is star-shaped with respect to B if for all $x \in \Omega$, the closed convex hull of $\{x\} \cup B$ is a subset of Ω.

Definition 3.71 (Chunkiness parameter)**.** Suppose that $\Omega \subset \mathbb{R}^d$ has diameter d_Ω and is star-shaped with respect to a ball B. Let

$$r_{\max} := \sup\{r > 0 : \Omega \text{ is star-shaped with respect to a ball of radius } r\}.$$

Then the chunkiness parameter of Ω is defined by $\gamma := d_\Omega / r_{\max}$.

Lemma 3.72 (Bramble-Hilbert, Lemma 4.3.8 in Brenner and Scott [2008])**.** *Let B be a ball in $\Omega \subset \mathbb{R}^d$ such that Ω is star-shaped with respect to B and such that its radius $r > r_{\max}/2$, where $r_{\max}$ is defined in Definition 3.71. Moreover, let d_Ω be the diameter of Ω, γ be the chunkiness parameter of Ω, and $Q^m f$ be the Taylor polynomial of order m of f averaged over B for any $f \in W^{m,\infty}(\Omega)$. Then there exists a constant $C(d, m, \gamma) > 0$ such that*

$$|f - Q^m f|_{W^{k,\infty}(\Omega)} \leq C(d, m, \gamma) d_\Omega^{m-k} |f|_{W^{m,\infty}(\Omega)}, \quad k = 0, 1, \cdots, m.$$

3.8.6 Additional lemmas on approximation

Lemma 3.73 (Corollary B.5 in Gühring et al. [2020])**.** *Let $d, m \in \mathbb{N}$ and $\Omega_1 \subset \mathbb{R}^d, \Omega_2 \subset \mathbb{R}^m$ both be open, bounded, and convex. If $f \in W^{1,\infty}(\Omega_1; \mathbb{R}^m)$ and $g \in W^{1,\infty}(\Omega_2)$ with $\operatorname{rad}(f) \subset \Omega_2$, then for the composition $g \circ f$, it holds that $g \circ f \in W^{1,\infty}(\Omega_1)$ and we have*

$$|g \circ f|_{W^{1,\infty}(\Omega_1)} \leq \sqrt{d} m |g|_{W^{1,\infty}(\Omega_2)} |f|_{W^{1,\infty}(\Omega_1; \mathbb{R}^m)}$$

and

$$\|g \circ f\|_{W^{1,\infty}(\Omega_1)} \leq \sqrt{d} m \max\{\|g\|_{L^\infty(\Omega_2)}, |g|_{W^{1,\infty}(\Omega_2)} |f|_{W^{1,\infty}(\Omega_1; \mathbb{R}^m)}\}.$$

Lemma 3.74 (Corollary B.6 in Gühring et al. [2020]). *Let $f \in W^{1,\infty}(\Omega)$ and $g \in W^{1,\infty}(\Omega)$. Then $fg \in W^{1,\infty}(\Omega)$ and we have*

$$|fg|_{W^{1,\infty}(\Omega)} \leq |f|_{W^{1,\infty}(\Omega)}\|g\|_{L^\infty(\Omega)} + \|f\|_{L^\infty(\Omega)}|g|_{W^{1,\infty}(\Omega)}$$

and

$$\|fg\|_{W^{1,\infty}(\Omega)} \leq \|f\|_{W^{1,\infty}(\Omega)}\|g\|_{L^\infty(\Omega)} + \|f\|_{L^\infty(\Omega)}\|g\|_{W^{1,\infty}(\Omega)}.$$

Lemma 3.75 (Proposition 4 in Yang et al. [2023]). *For any $N, L \in \mathbb{N}$ and $a > 0$, there exists a deep ReLU network $\phi_{\times,a}$ with width $15N$ and depth $2L$ such that $\|\phi_{\times,a}\|_{W^{1,\infty}((-a,a)^2)} \leq 12a^2$ and*

$$\|\phi_{\times,a}(x,y) - xy\|_{W^{1,\infty}((-a,a)^2)} \leq 6a^2N^{-L}.$$

Furthermore, it holds that

$$\phi_{\times,a}(0,y) = \frac{\partial \phi_{\times,a}(0,y)}{\partial y} = 0 \text{ for } y \in (-a,a).$$

Lemma 3.76. *For any $N, L \in \mathbb{N}$ and $a, b > 0$, there exists a deep ReLU network $\phi_{\times,(a,b)}$ with width $15N$ and depth $2L$ such that $\|\phi_{\times,(a,b)}\|_{W^{1,\infty}((-a,a)\times(-b,b))} \leq 12ab$ and*

$$\|\phi_{\times,(a,b)}(x,y) - xy\|_{W^{1,\infty}((-a,a)\times(-b,b))} \leq 6abN^{-L}.$$

Furthermore, it holds that

$$\phi_{\times,(a,b)}(x,0) = \frac{\partial \phi_{\times,(a,b)}(x,0)}{\partial x} = 0 \text{ for } x \in (-a,a).$$

Proof. The proof idea is similar to that of Proposition 4 in Yang et al. [2023]. □

Lemma 3.77 (Proposition 5 in Yang et al. [2023]). *For any $N, L, s \in \mathbb{N}$ with $s \geq 2$, there exists a deep ReLU network $\phi_\times$ with width $\mathcal{O}(s \vee N)$ and depth $\mathcal{O}(s^2L)$ such that*

$$\|\phi_\times(x) - x_1x_2\cdots x_s\|_{W^{1,\infty}((0,1)^s)} \lesssim s(N+1)^{-7sL}.$$

Furthermore, for any $i = 1, 2, \cdots, s$, if $x_i = 0$, then we have

$$\phi_\times(x_1, x_2, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_s) = \frac{\partial \phi_\times(x_1, x_2, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_s)}{\partial x_j} = 0, \ i \neq j.$$

Lemma 3.78 (Lemma 6 in Yang et al. [2023]). *Let $\{g_m\}_{m\in\{1,2\}^d}$ be the partition of unity given in Definition 3.53. Then it satisfies:*

(1) $\sum_{m\in\{1,2\}^d} g_m(x) = 1$ *for every* $x \in [0,1]^d$;

(2) $\mathrm{supp}(g_m) \cap [0,1]^d \subset \Omega_m$ *where* Ω_m *is given in Definition 3.52;*

(3) *For any* $m = [m_1, m_2, \cdots, m_d]^\top \in \{1,2\}^d$ *and* $x = [x_1, x_2, \cdots, x_d]^\top \in [0,1]^d \setminus \Omega_m$, *there exists an index* $j \in \{1, 2, \cdots, d\}$ *such that* $g_{m_j} = 0$ *and* $\frac{\mathrm{d}g_{m_j}(x_j)}{\mathrm{d}x_j} = 0$.

Lemma 3.79 (Lemma 7 in Yang et al. [2023]). *For any* $\chi(x) \in W^{1,\infty}((0,1)^d)$, *let*

$$B_5 := \max\{\|\chi\|_{W^{1,\infty}((0,1)^d)}, \|\phi_m\|_{W^{1,\infty}((0,1)^d)}\}.$$

Then for any $m \in \{1,2\}^d$, *it holds that*

$$\begin{aligned}
&\|\phi_m(x)\cdot\chi(x)\|_{W^{1,\infty}((0,1)^d)} = \|\phi_m(x)\cdot\chi(x)\|_{W^{1,\infty}(\Omega_m)},\\
&\|\phi_m(x)\cdot\chi(x) - \phi_{\times,B_5}(\phi_m(x),\chi(x))\|_{W^{1,\infty}((0,1)^d)}\\
&= \|\phi_m(x)\cdot\chi(x) - \phi_{\times,B_5}(\phi_m(x),\chi(x))\|_{W^{1,\infty}(\Omega_m)}.
\end{aligned}$$

Lemma 3.80. *For any* $\chi(x) \in L^\infty((0,1)^d)$, *let* $B_6 := \max\{\|\chi\|_{L^\infty((0,1)^d)}, \|\phi_m\|_{L^\infty((0,1)^d)}\}$. *Then for any* $m \in \{1,2\}^d$, *it holds that*

$$\begin{aligned}
&\|\phi_m(x)\cdot\chi(x)\|_{L^\infty((0,1)^d)} = \|\phi_m(x)\cdot\chi(x)\|_{L^\infty(\Omega_m)},\\
&\|\phi_m(x)\cdot\chi(x) - \phi_{\times,B_6}(\phi_m(x),\chi(x))\|_{L^\infty((0,1)^d)}\\
&= \|\phi_m(x)\cdot\chi(x) - \phi_{\times,B_6}(\phi_m(x),\chi(x))\|_{L^\infty(\Omega_m)}.
\end{aligned}$$

Proof. The proof is similar to that of [Yang et al., 2023, Lemma 7]. To prove the equalities, we need to show that

$$\|\phi_m(x)\cdot\chi(x)\|_{L^\infty((0,1)^d\setminus\Omega_m)} = 0 \quad \text{and} \quad \|\phi_{\times,B_6}(\phi_m(x),\chi(x))\|_{L^\infty((0,1)^d\setminus\Omega_m)} = 0.$$

In Lemma 3.54, it is shown that

$$\phi_m(x) := \phi_\times(g_{m_1}(x_1), g_{m_2}(x_2), \cdots, g_{m_d}(x_d)) = 0$$

where $g_m(x) = [g_{m_1}(x_1), g_{m_2}(x_2), \cdots, g_{m_d}(x_d)]^\top$ is defined in Definition 3.53. Then by Lemma 3.78, for any $x = [x_1, x_2, \cdots, x_d]^\top \in (0,1)^d \setminus \Omega_m$, there exists m_j such that $g_{m_j}(x_j) = 0$. By the definition of $\phi_m(x)$ and Lemma 3.77, it yields that

$$\phi_m(x) = 0, \quad \forall x \in (0,1)^d \setminus \Omega_m.$$

Therefore, for any $x \in (0,1)^d \setminus \Omega_m$, it holds that $|\phi_m(x) \cdot \chi(x)| = 0$.

Similarly, for any $x \in (0,1)^d \setminus \Omega_m$, it holds that

$$\phi_{\times,B_6}(\phi_m(x), \chi(x)) = \phi_{\times,B_6}(0, \chi(x)) = 0.$$

This completes the proof. □

Lemma 3.81 (Proposition 4.3 in Lu et al. [2021]). *Given any $N, L \in \mathbb{N}$ and $\delta \in (0, 1/(3K)]$ for $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, there exists a deep ReLU network ϕ with width $4N+5$ and depth $4L+4$ such that*

$$\phi(x) = k, \quad x \in \left[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot \mathrm{Id}_{\{k<K-1\}} \right], \quad k = 0, 1, \cdots, K-1.$$

Lemma 3.82 (Proposition 4.4 in Lu et al. [2021]). *Given any $N, L, s \in \mathbb{N}$ and $\xi_i \in [0,1]$ for $i = 0, 1, \cdots, N^2L^2 - 1$, there exists a deep ReLU network ϕ with width $16s(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that*

$$|\phi(i) - \xi_i| \leq (NL)^{-2s} \text{ for } i = 0, 1, \cdots, N^2L^2 - 1$$

and that

$$\phi(x) \in [0,1], \ x \in \mathbb{R}.$$

3.8.7 Hatsell-Nolte identity

Here, we exhibit Hatsell-Nolte identity [Hatsell and Nolte, 1971, Dytso et al., 2023b].

Lemma 3.83. *Suppose that $\mathsf{X} \sim \mu$ and $\epsilon \sim \gamma_{d,\sigma^2}$. Let $\mathsf{Y} = \mathsf{X} + \epsilon$ and $p(y)$ be the marginal density of Y. Then it holds that*

$$\mathrm{Cov}(\mathsf{X}|\mathsf{Y} = y) = \sigma^2 \nabla_y \mathbb{E}[\mathsf{X}|\mathsf{Y} = y] = \sigma^2 \mathrm{I}_d + \sigma^4 \nabla_y^2 \log p(y).$$

Chapter 4

Statistical Analysis of a Bayesian Fine-tuning Approach

Diffusion models are a class of continuous generative models built with SDEs whose generation ability has been largely reinforced by various fine-tuning procedures. However, the mystery of fine-tuning has seldom been uncovered from a statistical perspective. In this chapter, we address the gap in the systematic understanding of the advantages of fine-tuning mechanisms from a statistical perspective. We prove that a pre-trained large diffusion model can gain a faster convergence rate from the Bayesian fine-tuning procedure when adapted to perform conditional generation tasks. This improvement in the convergence rate justifies that a pre-trained large diffusion model would perform better on a downstream conditional generation task than a standard conditional diffusion model, whenever an appropriate fine-tuning procedure is implemented.

4.1 Introduction

Recently, fine-tuning of large models has been successfully applied across various domains, including natural language processing, computer vision, and speech recognition. This process involves taking a model that has already been trained on a broad dataset and refining it using a smaller and task-specific dataset. The goal is to use the general knowledge embedded in the large model while tailoring its capabilities to meet particular needs. Fine-tuning not only enhances performance on specialized tasks but also reduces the computational resources and time required compared to training a model from scratch. This approach has become increasingly important as models grow in size and complexity, offering a practical pathway to harness their full potential across diverse applications.

Current work, including T2I-Adapter [Mou et al., 2024], ControlNet [Zhang et al., 2023], and Bayesian Power Steering (BPS, Huang et al. [2024]), uses fine-tuning tech-

niques for Stable Diffusion (SD, Rombach et al. [2022]), facilitating precise spatial control over image generation. While extensive experimental results [Huang et al., 2024, Cheng et al., 2023a] have demonstrated that fine-tuning large-scale pre-trained models can yield exceptional generative outcomes, even with limited training datasets, the mechanisms underlying the efficacy of using pre-trained models remain ambiguous.

In this chapter, we aim to fill the gap in the systematic understanding of the advantages of pre-training mechanisms from a statistical perspective. We employ Stable Diffusion, a cutting-edge open-source large image model developed on the LAION-5B dataset, which comprises 585 billion images and effectively captures the probability space of natural images. Based on this large model, we consider the generic setting where the support of the fine-tuning target Z_0' is a subset of the support of the pre-training target Z_0. Then we adopt the Bayesian fine-tuning approach [Ho and Salimans, 2022, Huang et al., 2024] that is widely used in diffusion models and has demonstrated effectiveness in numerous experiments [Ho and Salimans, 2022, Mou et al., 2024, Zhang et al., 2023, Huang et al., 2024].

We prove that, under some regularity conditions, the Bayesian fine-tuning approach achieves the convergence rate $m^{-\frac{2\beta}{d+2\beta}} \vee n^{-\frac{2\alpha}{d+k+2\alpha}}$, where m is the sample size of pre-training, n is the labeled data size for fine-tuning, and β, α are smoothness indices. Meanwhile, if we train a conditional diffusion model from scratch using only the labeled data, the convergence rate is $n^{-\frac{2\delta}{d+k+2\delta}}$ with $\delta \leq \min(\alpha, \beta)$. By comparing these two rates, we can justify the benefit of pre-training when we have abundant data ($m >> n$) from the prior data space.

The rest of this chapter is organized as follows. We introduce the problem setting in Section 4.2. Then we revisit the background of diffusion models and the Bayesian fine-tuning approach in Sections 4.3 and 4.4. The main results are presented in Section 4.5 which provides an overview of the statistical error analysis of the pre-training stage and the fine-tuning stage in Sections 4.6 and 4.7. We defer the proofs of the theoretical results to Section 4.9.

4.2 Problem setting

This section formally formulates the problem and provides the background in the context of fine-tuning generation tasks. Let the generative target of the pre-trained model be denoted as $\mathsf{Z}_0 \in \mathcal{Z} \subseteq \mathbb{R}^d$, where $\mathcal{Z} := (\Omega, \mathsf{F}, \mathcal{P})$ is the prior probability space. The goal of fine-tuning is to generate samples from the probability space defined on a non-zero measurable subset $\Delta \in \mathsf{F}$, termed the fine-tuning probability space, i.e., $\mathcal{Z}_{\Delta,\mathfrak{g}} := (\Delta, \Delta \cap \mathsf{F}, \mathcal{P}(\cdot|\mathfrak{g}))$. Here, F is the Borel field generated by the random variable Z_0, and $\Delta \cap \mathsf{F} := \{\mathsf{E} \cap \Delta | \mathsf{E} \in \mathsf{F}\}$. The subfield $\mathfrak{g}$ represents an area of interest, with $\Delta \in \mathfrak{g} \subseteq \Delta \cap \mathsf{F}$. Specifically,

$$\forall \mathsf{E} \in \Delta \cap \mathsf{F}, \Lambda \in \mathfrak{g}: \quad \mathcal{P}(\mathsf{E}|\Lambda) := \frac{\mathcal{P}(\mathsf{E} \cap \Lambda)}{\mathcal{P}(\Lambda)} \text{ and } \mathcal{P}(\cdot|\mathfrak{g}) = \{\mathcal{P}(\cdot|\Lambda) | \Lambda \in \mathfrak{g}\}.$$

Note that $\mathcal{P}(\mathsf{E}|\mathfrak{g})$ is any one of the equivalence classes of random variables belonging to $\mathfrak{g}$. The following lemma further illustrates and instantiates this description to facilitate its application.

Lemma 4.1 (Chung [2001]). *For $k \in \mathbb{N}^+$, there exists some extended-value measurable function $\psi : \mathbb{R}^d \to \mathbb{R}^k$ such that $\mathsf{C} := \psi(\mathsf{Z}_0)$ and $\mathsf{C} : \mathfrak{g} \to \mathbb{R}^k$ is a random variable.*

Lemma 4.1 states that the target probability space can be identified by conditional sampling in applications. For convenience, we define $\mathsf{Z}_0|\mathsf{C} \in \mathcal{Z}_{\Delta,\mathfrak{g}} = \mathcal{Z}_{\Delta,\mathsf{C}} := (\Delta, \Delta \cap \mathsf{F}, \bar{\mathcal{P}}(\cdot|\mathsf{C}))$, where the target probability measure can be defined as the conditional probability measure $\bar{\mathcal{P}}(\mathsf{E}|\mathsf{C}) := \mathcal{P}(\mathsf{E} \cap \psi^{-1}(\mathsf{C}))/\mathcal{P}(\psi^{-1}(\mathsf{C}))$ for $\mathsf{E} \in \Delta \cap \mathsf{F}$. In practice, the condition C often represents the properties of interest, while elements in $\mathfrak{g}$ are sets that satisfy certain properties. Specifically, C often corresponds to attributes such as sketches, poses [Mou et al., 2024, Zhang et al., 2023], layouts [Huang et al., 2024, Cheng et al., 2023a], and class labels [Ho and Salimans, 2022] that are of particular interest in the field of image generation.

From an empirical perspective, large models are trained using finite samples $\{\mathsf{Z}_{0,i}\}_{i=1}^m$ drawn from the prior probability distribution $\mathcal{P}$ during the pre-training phase, where m is the sample size. We consider a fine-tuning architecture that incorporates a relatively small, trainable neural network on top of a fixed large model structure and parameters. During fine-tuning, limited labeled paired samples $\{(\mathsf{C}_i, \mathsf{Z}'_{0,i})\}_{i=1}^n$ that follow the probability distribution $\bar{\mathcal{P}}(\mathsf{Z}|\mathsf{C})$ to train this neural network, with n representing the sample

size. Generally, the fine-tuning task assumes that the labeled data are significantly less than the unlabeled data used for pre-training, that is $n << m$. In this chapter, we consider Stable Diffusion as the pre-trained model and a Bayesian fine-tuning approach as the fine-tuning architecture, with further details provided in the following section.

4.3 Diffusion models

Diffusion models [Ho et al., 2020] and their extensions [Rombach et al., 2022] have demonstrated significant success in generating images, videos, and text. These models employ a pre-defined forward process to transform the target random variable into Gaussian noise. Subsequently, a corresponding backward process is modeled to convert Gaussian noise back into the target random variable for sampling. This section introduces some preliminaries and key ingredients for Stable Diffusion, a widely recognized extension for diffusion models.

4.3.1 Stable diffusion

The key idea of Stable Diffusion is grounded in the manifold hypothesis, which suggests that image data is supported on a lower-dimensional substructure. It first maps the image data into latent probability space and subsequently employs diffusion models within this space to generate image representations. The SD model comprises three key modules: an auto-encoder, a language encoder (e.g., contrastive language-image pre-training [Radford et al., 2021]), and the diffusion model [Ho et al., 2020].

The auto-encoder consists of an encoder and a decoder, responsible for data compression and reconstruction. The encoder maps the standardized image data $\mathsf{X}_0 \in \mathcal{X} \subset \prod_{i=1}^{d_{\text{img}}}[-1,1]$ to a lower-dimensional probability space $\mathcal{Z} := (\Omega, \mathsf{F}, \mathcal{P}) \subset \mathbb{R}^d$ (with $d = 64^2 \times 4 < d_{\text{img}} = 512^2 \times 3$) via the transformation:

$$\mathsf{Z}_0 = E_{\theta_\mu}(\mathsf{X}_0) + E_{\theta_\sigma}(\mathsf{X}_0) \circ \epsilon, \text{ where } \epsilon \sim N(0, \mathrm{I}_d), \tag{4.1}$$

where $E_\theta = (E_{\theta_\mu}^\top, E_{\theta_\sigma}^\top)^\top$ represents the encoder and $\circ$ denotes the Hadamard product. The functions $E_{\theta_\mu}, E_{\theta_\sigma} : \mathcal{X} \to \mathbb{R}^d$ parameterize the conditional mean and variance of the

latent embedding $\mathsf{Z}_0|\mathsf{X}_0$. The decoder $D : \mathcal{Z} \to \mathcal{X}$ reconstructs the image as $D(\mathsf{Z}_0) = \mathsf{X}_0$.

The language encoder embeds the text prompt in Euclidean space $\mathbb{R}^{k_{\text{text}}}$. Without loss of generality, we treat the text condition as the constant phrase "a high-quality, detailed, and professional image," thereby considering Stable Diffusion as an unconditional generative model and omitting a detailed discussion of the language encoder.

Given that the target dataset is supported on a subset of $\mathcal{X}$, we use the pre-trained auto-encoder to map the target data into the latent space for fine-tuning, such that the corresponding support Δ is also the subset of Ω, consistent with our generic setting. Our primary interest lies in the probability space of the image representation $\mathcal{Z}$ resulting from the auto-encoder's transformation. The diffusion model is subsequently used to create representations of the image in the latent space $\mathbb{R}^d$.

We now define the forward process of the diffusion model.

Definition 4.2 (forward process)**.** The forward process of the diffusion model is denoted as $\{\mathsf{Z}_t\}_{t=0}^T$ for $t \in [T] := \{0, 1, ..., T\}$, and follows the iterative form:

$$\mathsf{Z}_t = \sqrt{\alpha_t}\mathsf{Z}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1},\ \epsilon_0, ..., \epsilon_{T-1} \overset{iid}{\sim} N(0, \mathrm{I}_d), \tag{4.2}$$

where the sequence $\{\alpha_t\}_{t=0}^\infty \subsetneq \mathbb{R}$ satisfy the following conditions:

(1) $\alpha_t \in (0, 1)$,

(2) let $\bar{\alpha}_t := \Pi_{i=1}^t \alpha_i$, then $\lim_{t\to\infty} \bar{\alpha}_t = 0$.

In Definition 4.2, the parameter α_t is predefined and must satisfy two conditions. The first condition ensures the well-defined variance of Z_t, and the second guarantees that as t approaches infinity, Z_t converges to a multivariate standard Gaussian distribution. Additionally, sampling with the iterative format in Definition 4.2 at any arbitrary time t can induce high computational costs. An equivalent closed form of the forward process addresses this issue, as presented formally in the following lemma.

Lemma 4.3. *The sequence of random variables $\{\mathsf{Z}_t\}_{t=0}^T$ defined in (4.2) satisfy*

$$\mathsf{Z}_t = \sqrt{\bar{\alpha}_t}\mathsf{Z}_0 + \sqrt{1-\bar{\alpha}_t}\eta,\ \eta \sim N(0, \mathrm{I}_d). \tag{4.3}$$

If assumption 4.2 holds, $\{\bar{\alpha}_t\}_{t=1}^\infty$ is a strictly decreasing sequence within the interval $(0, 1)$ and we have $\lim_{t\to\infty} \mathsf{Z}_t = \eta$.

Lemma 4.3 implies the role of the parameters $\bar{\alpha}_t$ is to show the dynamic signal-to-noise ratio. The forward process be conceptualized as a weighted average between the signal and noise components, progressively perturbing the target data Z_0 toward Gaussian noise as time approaches infinity. In practice, T is chosen to be sufficiently large to ensure that the data distribution is mapped to a multivariate standard Gaussian distribution through this dynamical system with controllable bias.

The backward process, known as Denoising Diffusion Probabilistic Model (DDPM, Ho et al. [2020]) sampler, is used for sampling, which initiates with Gaussian noise and proceeds through the following iterations.

Definition 4.4 (DDPM sampler)**.** The backward process of the diffusion model, that is the DDPM sampler, is denoted as $\{\mathsf{Z}_t\}_{t=0}^{T}$ for $t \in [T]$, and follows the iteration rule:

$$\mathsf{Z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathsf{Z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} f^*(\mathsf{Z}_t, t)\right) + \sigma_t \eta, \text{ with } \mathsf{Z}_T, \eta \sim N(0, \mathrm{I}_d), \tag{4.4}$$

where η is independent of Z_t and η, $\sigma_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\alpha_t$, the denoising function $f^* : \mathbb{R}^d \times [T] \to \mathbb{R}^d$ is defined as $f^*(t,z) := \mathbb{E}[\eta | \mathsf{Z}_t = z]$.

Here, the denoising function $f^*(t,z,c)$ is the target model during pre-training. Let $p(t,z) : \mathbb{R}^d \times [T] \to \mathbb{R}$ denote the time-dependent density function of Z_t, then the score function $\nabla \log p(t,z) : \mathbb{R}^d \times [T] \to \mathbb{R}^d$ takes the following form:

$$\nabla \log p(t,z) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}} f^*(t,z).$$

Given a fixed auto-encoder, the target of pre-trained modeling is to estimate the denoising function f^* from extensive observations of Z_0, followed by sampling through the DDPM sampler (4.4).

4.3.2 Conditional DDPM sampler

Conditional generation is an effective method for transferring the support of the pre-trained model Ω to the support of target data Δ, as outlined in Lemma 4.1. This process is realized by employing the conditional score function within the backward process. We introduce the definition of conditional DDPM sampler as follows.

Definition 4.5 (Conditional DDPM sampler)**.** With a condition variable $\mathsf{C} \in \mathcal{C}$, the conditional DDPM sampler is defined as:

$$\mathsf{Z}_{t-1}^{\mathsf{C}} = \frac{1}{\sqrt{\alpha_t}}\left(\mathsf{Z}_t^{C} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} F^*(\mathsf{Z}_t^{C}, t, \mathsf{C})\right) + \sigma_t \eta, \text{ with } \mathsf{Z}_T^{\mathsf{C}} = \mathsf{Z}_T. \tag{4.5}$$

Here, $\mathsf{Z}_t^{\mathsf{C}} := \mathsf{Z}_t|\mathsf{C}$, and the conditional denoising function $F^*(t,z,c) : \mathbb{R}^{d+k} \times [\tau, T] \to \mathbb{R}^d$ is defined by $F^*(t,z,c) := \mathbb{E}[\eta|\mathsf{Z}_t = z, \mathsf{C} = c]$.

In this context, the conditional denoising function $F^*(t,z,c)$ serves as the target model in the fine-tuning task. Let $\bar{p}(t,z|c) : \mathbb{R}^{d+k} \times [T] \to \mathbb{R}$ be the corresponding conditional density function, and let $\nabla \log \bar{p}(t,z|c) : \mathbb{R}^{d+k} \times [T] \to \mathbb{R}^d$ represent the conditional score function. Similarly, we have the closed form:

$$\nabla \log \bar{p}(t,z|c) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}} F^*(t,z,c).$$

4.4 A Bayesian fine-tuning approach

The Bayesian fine-tuning approach [Huang et al., 2024] aims to estimate the conditional denoising function F^* given a pre-trained denoising function f^*. This section introduces the principle and implementation of this method.

4.4.1 Basic principle of Bayesian fine-tuning

The core of the Bayesian fine-tuning approach is to establish the relationship between the denoising function f^* and its conditional counterpart F^* using Bayes' rule. We revisit the idea in the following lemma.

Lemma 4.6 (Huang et al. [2024])**.** *Let $M^* : \mathbb{R}^{d+k} \times [T] \to \mathbb{R}^d$ be defined as*

$$M^*(t,z,c) := -\sqrt{1-\bar{\alpha}_t} \nabla \log p(\mathsf{C} = c|\mathsf{Z}_t = z).$$

Then, it holds that

$$F^*(t,z,c) = f^*(t,z) + M^*(t,z,c). \tag{4.6}$$

Lemma 4.6 explores the effective integration of the denoising function f^* with a gradient of the time-dependent classifier M^* to obtain the target function F^*.

4.4.2 Estimation for Bayesian fine-tuning

The Bayesian fine-tuning approach is a two-stage approach. In the pre-training stage, we estimate the denoising function f^* while we estimate the difference function M^* with the aid of an estimate of f^* during the fine-tuning stage. We use deep ReLU networks in both stages for the purpose of nonparametric estimation. Recall that we have defined the class of deep ReLU networks in Definition 3.5. We further restrict the function class of deep ReLU networks to be a class of Lipschitz functions. Let $0 < \tau \ll 1$ be an early stopping time for the estimation of denoising functions. Then, for any $t \in [\tau, T]$, we consider estimating the denoising function $f^*(t,\cdot)$ using a deep ReLU network $f \in \mathcal{F}_m := \mathcal{NN}(\mathsf{S}_m, \mathsf{W}_m, \mathsf{D}_m, \mathsf{B}_m, d, d) \cap W^{1,\infty}(\mathbb{R}^d; \mathbb{R}^d)$, where $W^{1,\infty}(\mathbb{R}^d; \mathbb{R}^d)$ stands for the class of Lipschitz functions $f : \mathbb{R}^d \to \mathbb{R}^d$.

In the procedure of pre-training, the DDPM methods [Ho et al., 2020, Rombach et al., 2022] solve a nonparametric regression problem to estimate the denoising function $f^*(t,\cdot)$ on the domain $\mathbb{R}^d$,

$$f^* \in \arg\min_f \left\{ \mathcal{L}_t(f) := \mathbb{E}\|\eta - f(t, \mathsf{Z}_t)\|^2 \right\}. \tag{4.7}$$

Let $\mathbb{D}_m^p := \{(\mathsf{Z}_{0,i}, \eta_i)\}_{i=1}^m$ be empirical observations of (Z_0, η) with sample size m. Then the empirical risk $\mathcal{L}_{t,m}$ is defined by

$$\mathcal{L}_{t,m}(f) := \frac{1}{m} \sum_{i=1}^m \|\eta_i - f(\mathsf{Z}_{t,i}, t)\|^2, \tag{4.8}$$

where $\mathsf{Z}_{t,i} := \sqrt{\bar{\alpha}_t}\mathsf{Z}_{0,i} + \sqrt{1-\bar{\alpha}_t}\eta_i$.

In practice, we approximate the denoising function f^* using the class of deep ReLU networks $\mathcal{F}_m$. Therefore, the pre-trained model is an empirical risk minimizer over the function class $\mathcal{F}_m$, defined by

$$\hat{f}_m = \arg\min_{f \in \mathcal{F}_m} \mathcal{L}_{t,m}(f). \tag{4.9}$$

In the procedure of fine-tuning, the approach estimates the gradient of the time-dependent classifier M^* by solving the following nonparametric regression problem over the domain $\mathbb{R}^{d+k}$:

$$(F^* - \hat{f}_m) \in \arg\min_M \left\{ \mathcal{J}_t(M) := \mathbb{E}\|\eta - \hat{f}_m(t, \mathsf{Z}_t) - M(t, \mathsf{Z}_t, \mathsf{C})\|^2 \right\}. \tag{4.10}$$

Let $\mathbb{D}_n^f := \{(\mathsf{Z}_{0,j}, \mathsf{C}_j, \eta_j)\}_{j=m+1}^{m+n} \subseteq \mathcal{Z} \times \mathcal{C} \times \mathbb{R}^d$ be the data for fine-tuning that are independent of the data $\mathbb{D}_m^p$ for pre-training. Given a pre-trained denoising model $\hat{f}_m$, the population risk defined in (4.10) leads to the following empirical risk:

$$\mathcal{J}_{t,n}(M) := \frac{1}{n} \sum_{j=m+1}^{m+n} \|\eta_j - \hat{f}_m(t, \mathsf{Z}_{t,j}) - M(t, \mathsf{Z}_{t,j}, \mathsf{C}_j)\|^2, \tag{4.11}$$

where $\mathsf{Z}_{t,i} := \sqrt{\bar{\alpha}_t}\mathsf{Z}_{0,i} + \sqrt{1-\bar{\alpha}_t}\eta_i$.

Next, we introduce a deep ReLU network with Lipschitz regularity as $M(t,\cdot,\cdot) \in \mathcal{G}_n := \mathcal{NN}(\mathsf{S}_n, \mathsf{W}_n, \mathsf{D}_n, \mathsf{B}_n, d+k, d) \cap W^{1,\infty}(\mathbb{R}^{d+k}; \mathbb{R}^d)$ to approximate the gradient of the time-dependent classifier $M^*(t,\cdot,\cdot)$. Based on a pre-trained model $\hat{f}_m$, we construct an estimator for M^* within the function class $\mathcal{G}_n$ as follows:

$$\widehat{M}_n := \arg\min_{M \in \mathcal{G}_n} \mathcal{J}_{t,n}(M). \tag{4.12}$$

Combining the two stages, we obtain an estimator of the conditional denoising function F^*, defined by

$$\hat{F}_{m,n}(t,\cdot,\cdot) = \hat{f}_m(t,\cdot,\cdot) + \widehat{M}_n(t,\cdot,\cdot) \in \mathcal{H}_{m,n} := \{f + M : \forall f \in \mathcal{F}_m, \forall M \in \mathcal{G}_n\}. \tag{4.13}$$

When an estimate of the conditional denoising function is available, we are ready to generate samples from the conditional diffusion model using the conditional DDPM sampler presented in Definition 4.5.

4.5 Main results

In this section, we first present the convergence rates for estimating denoising functions, which provide a statistical guarantee for models that adopt a heavy training-from-scratch strategy, facilitating future comparisons. It is worth emphasizing upfront that our analysis directly quantifies the effect of imperfect score estimation, which is filling the gaps in existing theories. We then present our error bound for the Bayesian fine-tuning approach, which offers a rigorous theoretical understanding of why a pre-trained large model generally helps.

4.5.1 Assumptions

Before proceeding, we impose a few mild assumptions on the regularity properties of the denoising function f^*, the difference M^*, the conditional denoising function $F^*(t,z,c)$, and the target data distribution.

Assumption 4.7. Let $\beta \geq 1$. For any $A > 0$ and any $t \in [\tau, T]$, the denoising function $f^*(t,\cdot)$ satisfies the regularity properties:

$$f^*(t,\cdot) \in W^{\beta,\infty}([-A,A]^d;\mathbb{R}^d), \quad \|f^*(t,\cdot)\|_{W^{\beta,\infty}([-A,A]^d;\mathbb{R}^d)} \lesssim A.$$

Assumption 4.8. Let $\alpha \geq 1$. For any $A > 0$ and any $t \in [\tau, T]$, the difference $M^*(t,\cdot,\cdot)$ satisfies the regularity properties:

$$M^*(t,\cdot,\cdot) \in W^{\alpha,\infty}([-A,A]^{d+k};\mathbb{R}^d), \quad \|M^*(t,\cdot,\cdot)\|_{W^{\alpha,\infty}([-A,A]^{d+k};\mathbb{R}^d)} \lesssim A.$$

Assumption 4.9. Let $\delta \geq 1$. For any $A > 0$ and any $t \in [\tau, T]$, the conditional denoising function $F^*(t,\cdot,\cdot)$ satisfies the regularity properties:

$$F^*(t,\cdot,\cdot) \in W^{\delta,\infty}([-A,A]^{d+k};\mathbb{R}^d), \quad \|F^*(t,\cdot,\cdot)\|_{W^{\delta,\infty}([-A,A]^{d+k};\mathbb{R}^d)} \lesssim A.$$

Remark 4.10. We impose regularity assumptions on the denoising functions f^*, F^*, and their difference M^* in Assumptions 4.7, 4.8, and 4.9. These smoothness assumptions are common in the literature on nonparametric statistics [Györfi et al., 2002, Tsybakov, 2009] and deep nonparametric regression [Suzuki, 2019, Schmidt-Hieber, 2020, Kohler and Langer, 2021, Jiao et al., 2023a]. Given these smoothness properties, we can quantify the approximation errors and the stochastic errors in the estimation of the denoising functions which lead to an overall error bound for distribution learning.

Remark 4.11. According to Lemma 4.6, it holds $F^* = f^* + M^*$. Then it can be supposed that the smoothness indices in Assumptions 4.7, 4.8, and 4.9 satisfy $\delta \leq \min(\alpha, \beta)$. This inequality is motivated by the relative regularity of the denoising function and its conditional counterpart. The low regularity of the conditional denoising function F^* may be sourced from either the unconditional denoising function f^* or the derivative of the classifier-guidance M^*.

The definition of the forward process $\{Z_t\}_{t=0}^T$ implies that the denoising functions and their difference are defined over unbounded domains. We consider the sub-Gaussian

property of the random vector $(\mathsf{Z}_0, \mathsf{C})$ to facilitate bounding the estimation errors on these unbounded domains. Specifically, the property of Z_0 can be derived by imposing mild constraints on the encoder function within the auto-encoder. We formalize this as follows.

Assumption 4.12. The codomain of encoder function $E_\theta = (E_{\theta_\mu}^\top, E_{\theta_\sigma}^\top)^\top$ is bounded, and the probability distribution of the random variable C is sub-Gaussian, that is, there exists a constant K such that $\mathbb{E}[\exp(K\|\mathsf{C}\|_2^2)] < \infty$.

Remark 4.13. The first constraint outlined in Assumption 4.12 is readily satisfied, particularly in the context of Stable Diffusion, where the image data X_0 is bounded. If the encoder function E_θ is continuous, the first constraint is met, representing a mild condition, as E_θ is implemented via a neural network.

4.5.2 Error bounds of drift estimation

The analysis of estimation errors in denoising functions is notably challenging and has been largely overlooked in existing literature. Typically, the estimation error is treated as a constant to simplify the error analysis of the sampling distribution. This chapter enhances the understanding of estimation errors in denoising functions, providing a comprehensive examination presented in the following sections.

Theorem 4.14. *Suppose that Assumptions 4.2-4.12 hold. Let the parameters of the deep ReLU network classes be properly specified. For any $t \in [\tau, T]$, the excess risk of drift estimation satisfies*

$$\mathbb{E}_{\mathbb{D}_m^p} \mathbb{E}_{\mathsf{Z}_t} \|\hat{f}_m - f^*\|_2^2 \lesssim m^{-\frac{2\beta}{d+2\beta}}$$

by omitting a polylogarithmic factor in m and a prefactor in d, k, β.

To quantitatively analyze the advantages of introducing pre-trained large models in fine-tuning, we present an assessment of the error associated with training the conditional denoising function from scratch under the premise of n training samples $\mathbb{D}_n^f$. Let the model training from scratch $\tilde{F}_n(t,\cdot,\cdot) : \mathbb{R}^{d+k} \to \mathbb{R}^d$ be defined as

$$\tilde{F}_n(t,\cdot,\cdot) = \arg\min_{F \in \mathcal{G}_n} \tilde{\mathcal{L}}_{t,n}(F),$$

where the empirical risk is defined by

$$\tilde{\mathcal{L}}_{t,n}(F) := \frac{1}{n}\sum_{j=m+1}^{m+n} \|\eta_j - F(t, \mathsf{Z}_{t,j}, \mathsf{C}_j)\|^2.$$

We then formulate the non-asymptotic error bound for the model trained from scratch in the following theorem.

Theorem 4.15. *Suppose that Assumptions 4.2-4.12 hold. Let the parameters of the deep ReLU network classes be properly specified. For any $t \in [\tau, T]$, the excess risk of drift estimation satisfies*

$$\mathbb{E}_{\mathbb{D}_n^f}\mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|\tilde{F}_n - F^*\|_2^2 \lesssim n^{-\frac{2\delta}{d+k+2\delta}}$$

by omitting a polylogarithmic factor in n and a prefactor in d, k, δ.

We are now positioned to state our theoretical guarantees for the Bayesian fine-tuning approach.

Theorem 4.16. *Suppose that Assumptions 4.2-4.12 hold. Let the parameters of the deep ReLU network classes be properly specified. For any $t \in [\tau, T]$, the excess risk of drift estimation satisfies*

$$\mathbb{E}_{\mathbb{D}_m^p \cup \mathbb{D}_n^f}\mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|\hat{F}_{m,n} - F^*\|_2^2 \lesssim m^{-\frac{2\beta}{d+2\beta}} \vee n^{-\frac{2\alpha}{d+k+2\alpha}}$$

by omitting a polylogarithmic factor in n and a prefactor in d, k, α, β.

Theorem 4.15 and Theorem 4.16 demonstrate the superiority of the fine-tuning framework. In practical scenarios, the sample size for pre-training is typically much larger than that for fine-tuning, which means $m \gg n$. Given the observation $\delta \leq \min(\alpha, \beta)$, the convergence rate of the fine-tuning approach is faster than that of learning the conditional diffusion model without pre-training by comparing the rates in Theorem 4.16 and 4.15.

4.6 Statistical analysis for the denoising models

In this section, we prove Theorems 4.14 and 4.15 to provide theoretical guarantees for learning the denoising models.

The estimators of the denoising function $\hat{f}_m$ and $\tilde{F}_n$ exhibit similar nonparametric convergence properties, with their excess risk upper bounds being exponential in relation to the sample size, as demonstrated in Theorem 4.14 and Theorem 4.15. This section outlines our approach to analyzing these risks.

To begin with, we present a basic inequality for the excess risk in terms of the stochastic and approximation errors.

Lemma 4.17. *For any $t \in [\tau, T]$ and any random sample $\mathbb{D}_m^p$, the excess risk of the denoising function estimator $\hat{f}_m$ satisfies*

$$\begin{aligned} \mathcal{E}_p &:= \mathbb{E}_{\mathbb{D}_m^p} \mathbb{E}_{Z_t} \|\hat{f}_m(t, Z_t) - f^*(t, Z_t)\|_2^2 \\ &\leq 2\mathcal{E}_{\text{appr}}^p + \mathcal{E}_{\text{stoc}}^p, \end{aligned} \tag{4.14}$$

where the approximation error

$$\mathcal{E}_{\text{appr}}^p := \inf_{f \in \mathcal{F}_m} \mathbb{E}_{Z_t} \|f(t, Z_t) - f^*(t, Z_t)\|_2^2,$$

and the stochastic error

$$\mathcal{E}_{\text{stoc}}^p := \mathbb{E}_{\mathbb{D}_m^p} [\mathcal{L}_t(f^*) - 2\mathcal{L}_{t,m}(\hat{f}_m) + \mathcal{L}_t(\hat{f}_m)].$$

The decomposition (3.18) of the excess risk $\mathcal{E}_p$ is common in the literature on nonparametric regression. We refer the readers to Theorem 11.5 in [Györfi et al., 2002], Lemma 3.1 in [Jiao et al., 2023a], and Theorem 3.29.

4.6.1 Approximation error of pre-training

The approximation capacity of deep neural networks on bounded domains has been well studied in the literature (c.f. Yarotsky [2017, 2018], Shen et al. [2020, 2021], Lu et al. [2021]). In our analysis, we need to bound the approximation error on an unbounded domain $\mathbb{R}^d$. To tackle the challenges posed by the unbounded support in the space variable Z_t for $t \in [T]$, we employ a truncated approximation that decomposes the approximation error into two parts: the truncation error and the truncated approximation error, as detailed in the following lemma.

Lemma 4.18. *Let $\Omega_A := [-A, A]^d$. For $\bar{f} \in \mathcal{F}_m$ and any $A > 0$, $t \in [\tau, T]$, the approximation error of the pre-training process satisfies a basic inequality as follows:*

$$\mathcal{E}_{\text{appr}}^p = \inf_{f \in \mathcal{F}_m} \mathbb{E}_{Z_t} \|f - f^*\|_2^2 \lesssim \mathcal{E}_{\text{trunc}}^p + \mathcal{E}_{\text{appr}}^{p,\text{trunc}}, \tag{4.15}$$

where the truncation error

$$\mathcal{E}^{p}_{\text{trunc}} := \mathbb{E}_{\mathsf{Z}_t} \|(\bar{f} - f^*)\,\mathrm{Id}_{\Omega_A^c}(\mathsf{Z}_t)\|_2^2,$$

and the truncated approximation error

$$\mathcal{E}^{p,\text{trunc}}_{\text{appr}} := \mathbb{E}_{\mathsf{Z}_t} \|(\bar{f} - f^*)\,\mathrm{Id}_{\Omega_A}(\mathsf{Z}_t)\|_2^2.$$

The proof of Lemma 4.18 is similar to that of Lemma 3.30. In what follows, we focus on the truncation error $\mathcal{E}^{p}_{\text{trunc}}$. We prove that the truncation error $\mathcal{E}^{p}_{\text{trunc}}$ decays very fast in the parameter A, as a result of the sub-Gaussian property of $(\mathsf{Z}_0, \mathsf{C})$.

Lemma 4.19. *Suppose that Assumption 4.12 is satisfied. For any $A > 0$ and $t \in [\tau, T]$, the truncation error satisfies*

$$\mathcal{E}^{p}_{\text{trunc}} \lesssim A^2 \exp(-C_1 A^2),$$

where C_1 is a constant, and we omit a constant in d and the fourth moment of the random variable Z_0.

The proof of Lemma 4.19 is similar to that of Lemma 3.36. Moreover, the truncated approximation error $\mathcal{E}^{p,\text{trunc}}_{\text{appr}}$ can be bounded by constructing an approximation function with deep neural networks on the hypercube $[-A, A]^d$.

Lemma 4.20 (Truncated approximation error)**.** *Suppose that Assumption 4.7 is satisfied. For any $N, L \in \mathbb{N}$ and any $t \in [\tau, T]$, there exists a function $\bar{f}(z)$ implemented by a deep ReLU network with width $\mathcal{O}(2^d \beta^2 d^{\beta-1} N \log N)$, depth $\mathcal{O}(d^2 \beta^2 L \log L)$ such that the following properties hold simultaneously:*

(i) Boundedness and Lipschitz regularity:

$$\sup_{z \in \mathbb{R}^d} \|\bar{f}(z)\|_\infty \lesssim A,$$

$$\sup_{z, y \in \mathbb{R}^d} \|\bar{f}(z) - \bar{f}(y)\|_\infty \lesssim A \|y - z\|_\infty.$$

(ii) Approximation error bound:

$$\sup_{z \in [-A, A]^d} \|\bar{f}(z) - f^*(t, z)\|_\infty \lesssim A^2 (NL)^{-2\beta/d}.$$

Note that we omit some prefactors in d and β.

The proof of Lemma 4.20 is given in Section 4.9.3. Combining the results of Lemmas 4.19 and 4.20, we can derive the upper bound for the approximation error based on the inequality (4.15), with a properly selected parameter A.

4.6.2 Stochastic error of pre-training

The stochastic error $\mathcal{E}^p_{\text{stoc}}$ can be bounded by the complexity of $\mathcal{F}_m$ using the empirical process theory [Anthony and Bartlett, 1999, Bartlett et al., 2019, Jiao et al., 2023a]. We present our stochastic error analysis based on recent advancements in deep nonparametric regression [Jiao et al., 2023a]. Following Lemma 3.2 in Jiao et al. [2023a], we show that the stochastic errors are bounded in terms of the parameters of the deep ReLU networks classes $\mathcal{F}_m$.

Lemma 4.21. *Consider the pre-training model and the hypothesis class $\mathcal{F}_m \subseteq \mathcal{NN}(\mathsf{S}_m, \mathsf{W}_m, \mathsf{D}_m, \mathsf{B}_m, d, d)$. For any $m \in \mathbb{N}$ satisfying $m \geq \text{Pdim}(\mathcal{F}_m)$, the stochastic error satisfies*

$$\begin{aligned}\mathcal{E}^p_{\text{stoc}} &= \mathbb{E}_{\mathbb{D}^p_m}[\mathcal{L}_t(f^*) - 2\mathcal{L}_{t,m}(\hat{f}_m) + \mathcal{L}_t(\hat{f}_m)] \\ &\lesssim \frac{1}{m}(\log m)^4 d\mathsf{B}_m^4 \mathsf{S}_m \mathsf{D}_m \log(\mathsf{S}_m)\log(\mathsf{B}_m m^2).\end{aligned}$$

The proof of Lemma 4.21 is given in Section 4.9.4. We evaluate each error term in Lemmas 4.17, 4.18, 4.19, 4.20, and 4.21. Our main results in Theorems 4.14 and 4.15 are derived by balancing the error terms on the right-hand side of (3.18) with respect to the corresponding sample size and function class.

4.7 Statistical analysis of the fine-tuning approach

This section is devoted to establishing Theorem 4.16. Given the pre-trained denoising function estimator $\hat{f}_m$, we further consider the error decomposition for the fine-tuning stage. We derive that the overall error for estimating the conditional denoising function is upper bounded by the summation of the approximation errors and the stochastic errors induced in both the pre-training stage and the fine-tuning stage.

Lemma 4.22. *For any $t \in [\tau, T]$ and any random sample $\mathbb{D}_m^p \cup \mathbb{D}_n^f$, the excess risk of the conditional denoising function estimator $\hat{F}_{m,n}$ satisfies*

$$\mathbb{E}_{\mathbb{D}_m^p \cup \mathbb{D}_n^f} \mathbb{E}_{(\mathsf{Z}_t, \mathsf{C})} \|\hat{F}_{m,n} - F^*\|_2^2 \leq 4\mathcal{E}_{\text{appr}}^f + \mathcal{E}_{\text{stoc}}^f + 4\mathcal{E}_p, \tag{4.16}$$

where the approximation error

$$\mathcal{E}_{\text{appr}}^f := \inf_{M \in \mathcal{G}_n} \mathbb{E}_{(\mathsf{Z}_t, \mathsf{C})} \|M - M^*\|_2^2,$$

and the stochastic error

$$\mathcal{E}_{\text{stoc}}^f := \mathbb{E}_{\mathbb{D}_m^p \cup \mathbb{D}_n^f} [\mathcal{J}_t(F^* - \hat{f}_m) - 2\mathcal{J}_{t,n}(\widehat{M}_n) + \mathcal{J}_t(\widehat{M}_n)].$$

The proof of Lemma 4.22 is given in Section 4.9.2. In addition to the bias-variance trade-off of the denoising estimator in fine-tuning, the decomposition in (4.16) of the excess risk also encompasses errors from the pre-trained model. However, as established in Lemma 4.22, the significant discrepancy in sample sizes between the pre-training and fine-tuning datasets results in excess risk that is primarily dominated by errors introduced by the fine-tuning structure.

In what follows, we focus on the analysis of the approximation error and the stochastic error of the fine-tuning stage.

4.7.1 Approximation error of fine-tuning

We derive similar error bounds to quantify the approximation error of the fine-tuning stage.

Lemma 4.23. *Let $\Omega_B := [-B, B]^{d+k}$. For $\bar{M} \in \mathcal{F}_n$, $t \in [\tau, T]$, and any $B > 0$, the approximation error of the supervised part satisfies a basic inequality as follows:*

$$\mathcal{E}_{\text{appr}}^f = \inf_{M \in \mathcal{F}_n} \mathbb{E}_{(\mathsf{Z}_t, \mathsf{C})} \|M - M^*\|_2^2 \lesssim \mathcal{E}_{\text{appr}}^{f,\text{trunc}} + \mathcal{E}_{\text{trunc}}^f, \tag{4.17}$$

where the truncated approximation error

$$\mathcal{E}_{\text{appr}}^{f,\text{trunc}} := \mathbb{E}_{(\mathsf{Z}_t, \mathsf{C})} \|(\bar{M} - M^*) \operatorname{Id}_{\Omega_B}(\mathsf{Z}_t, \mathsf{C})\|_2^2,$$

and the truncation error

$$\mathcal{E}_{\text{trunc}}^f := \mathbb{E}_{(\mathsf{Z}_t, \mathsf{C})} \|(\bar{M} - M^*) \operatorname{Id}_{\Omega_B^c}(\mathsf{Z}_t, \mathsf{C})\|_2^2.$$

The error decomposition (4.17) is similar to that in Lemma 3.30.

Lemma 4.24 (Truncation error). *Suppose that Assumption 4.12 is satisfied. For any $B > 0$, the truncation error satisfies*

$$\mathcal{E}^f_{\text{trunc}} = \mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|(\bar{M} - M^*)\,\mathrm{Id}\,_{\Omega_B^c}(\mathsf{Z}_t,\mathsf{C})\|_2^2 \lesssim B^2 \exp(-C_2 B^2),$$

where C_2 is a constant, and we omit a constant in d, k, and the fourth moment of Z_0.

In Lemma 4.24, we still bound the truncation error using the sub-Gaussian property of the random vector $(\mathsf{Z}_0, \mathsf{C})$.

Lemma 4.25 (Truncated approximation error). *Suppose that Assumption 4.8 is satisfied. For any $N, L \in \mathbb{N}$ and any $t \in [\tau, T]$, there exists a function $\bar{M}(z,c)$ implemented by a deep ReLU network with width $\mathcal{O}(2^{d+k}\alpha^2(d+k)^{\alpha-1}N\log N)$, depth $\mathcal{O}((d+k)^2\alpha^2 L\log L)$ such that the following properties hold simultaneously:*

(i) Boundedness and Lipschitz regularity: for any $y, z \in \mathbb{R}^d$ and any $b, c \in \mathbb{R}^k$,

$$\sup_{(z,c)\in\mathbb{R}^{d+k}} \|\bar{M}(z,c)\|_\infty \lesssim B^2,$$

$$\sup_{c\in\mathbb{R}^k} \|\bar{M}(z,c) - \bar{M}(y,c)\|_\infty \lesssim B\|y - z\|_\infty,$$

$$\sup_{z\in\mathbb{R}^d} \|\bar{M}(z,c) - \bar{M}(z,b)\|_\infty \lesssim B\|b - c\|_\infty.$$

(ii) Approximation error bound:

$$\sup_{(z,c)\in[-A,A]^{d+k}} \|\bar{M} - M^*\|_\infty \lesssim B^2(NL)^{-\frac{2\alpha}{d+k}}.$$

Note that we omit some prefactors in d, k, and α.

The proof of Lemma 4.25 is given in Section 4.9.3. We not only consider the approximation rate in Lemma 4.25 but also show that the constructed approximation function has a few regularity properties, such as the boundedness and the Lipschitz continuity. Then we can restrict the hypothesis space $\mathcal{F}$ to be a subset of the Lipschitz function class. The Lipschitzness is useful for controlling the approximation error over the unbounded support when combined with a sub-Gaussian tail of $(\mathsf{Z}_0, \mathsf{C})$.

4.7.2 Stochastic error of fine-tuning

Lemma 4.26. *Consider the fine-tuning model and the hypothesis class* $\mathcal{G}_n \subseteq \mathcal{NN}(\mathsf{S}_n, \mathsf{W}_n, \mathsf{D}_n, \mathsf{B}_n, d+k, d)$. *For any* $n \in \mathbb{N}$ *satisfying* $n \geq \mathrm{Pdim}(\mathcal{G}_n)$, *the stochastic error satisfies*

$$\begin{aligned}\mathcal{E}^f_{\mathrm{stoc}} &= \mathbb{E}_{\mathbb{D}^p_m \cup \mathbb{D}^f_n}[\mathcal{J}_t(F^* - \hat{f}_m) - 2\mathcal{J}_{t,n}(\widehat{M}_n) + \mathcal{J}_t(\widehat{M}_n)] \\ &\lesssim \frac{1}{n}(\log n)^4 (d+k)\mathsf{B}_n^4 \mathsf{S}_n \mathsf{D}_n \log(\mathsf{S}_n)\log(\mathsf{B}_n n^2).\end{aligned}$$

The proof of Lemma 4.26 is given in Section 4.9.4. By balancing the error terms in Lemmas 4.22, 4.23, 4.24, 4.25, and 4.26, and by incorporating the excess risk from pre-training, we derive our main results in Theorem 4.16.

4.8 Conclusion

We have conducted a systematic analysis of the pre-training and fine-tuning approach for diffusion models. We prove that, under some regularity conditions, the Bayesian fine-tuning approach achieves a faster convergence rate than the rate yielded by training from scratch using only the labeled data. This result provides a theoretical justification for the benefit of pre-training with abundant unlabelled data.

4.9 Proofs

In the section, we present proofs of the lemmas and theorems for establishing the convergence rate of the Bayesian pre-tuning and fine-tuning approach.

4.9.1 Proof of Lemma 4.3

Proof. Standard induction arguments can yield the result of the lemma. □

4.9.2 Proof of Lemma 4.22

Proof. It can be shown that the L^2 error of $\hat{F}_{m,n}$ satisfies

$$\begin{aligned}
&\mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|\hat{F}_{m,n} - F^*\|_2^2 \\
&= \mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|\widehat{M}_n - (F^* - \hat{f}_m)\|_2^2 \\
&= \mathcal{J}_t(\widehat{M}_n) - \mathcal{J}_t(F^* - \hat{f}_m),
\end{aligned} \tag{4.18}$$

where the second equality is due to $F^* - \hat{f}_m$ is the minimizer of the L^2 risk $\mathcal{J}_t$. Let us denote

$$\mathcal{E}^f_{\text{stoc}} := \mathbb{E}_{\mathbb{D}^p_m \cup \mathbb{D}^f_n}[\mathcal{J}_t(F^* - \hat{f}_m) - 2\mathcal{J}_{t,n}(\widehat{M}_n) + \mathcal{J}_t(\widehat{M}_n)],$$

and

$$\mathcal{E}_{\text{appr}} := \mathbb{E}_{\mathbb{D}^p_m} \inf_{M \in \mathcal{G}_n} \mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|M - (F^* - \hat{f}_m)\|_2^2.$$

Similar to the proofs of Jiao et al. [2023a, Lemma 3.1], it can be shown that

$$\mathbb{E}_{\mathbb{D}^p_m \cup \mathbb{D}^f_n}[\mathcal{J}_t(\widehat{M}_n) - \mathcal{J}_t(F^* - \hat{f}_m)] \le \mathcal{E}^f_{\text{stoc}} + 2\mathcal{E}_{\text{appr}}. \tag{4.19}$$

Let us also denote

$$\mathcal{E}^f_{\text{appr}} := \inf_{M \in \mathcal{G}_n} \mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|M - M^*\|_2^2,$$

and

$$\mathcal{E}_p := \mathbb{E}_{\mathbb{D}^p_m} \mathbb{E}_{\mathsf{Z}_t}\|f^* - \hat{f}_m\|_2^2.$$

It further holds that

$$\mathcal{E}_{\text{appr}} \le 2\mathcal{E}^f_{\text{appr}} + 2\mathcal{E}_p. \tag{4.20}$$

Combining (4.18), (4.19), and (4.20), we complete the proof by showing that

$$\mathbb{E}_{\mathbb{D}^p_m \cup \mathbb{D}^f_n} \mathbb{E}_{(\mathsf{Z}_t,\mathsf{C})}\|\hat{F}_{m,n} - F^*\|_2^2 \le \mathcal{E}^f_{\text{stoc}} + 4\mathcal{E}^f_{\text{appr}} + 4\mathcal{E}_p.$$

□

4.9.3 Proofs of approximation error bounds

In this section, we present proofs of lemmas for bounding the approximation errors of pre-training and fine-tuning. To begin with, we show an approximation bounds that is useful for further approximation error analysis.

Polynomial approximation in Hölder classes

The polynomial approximation theory in Hölder or general Sobolev spaces plays a central role in studying approximation rates of finite element methods and deep neural networks.

We construct polynomial approximations of functions in Hölder classes, and generalize the classical Bramble-Hilbert lemma to functions of Hölder smoothness.

Building on Lemma A.8 in Petersen and Voigtlaender [2018], we construct a polynomial approximation in L^∞-norm for functions in Hölder classes, which extends [Dupont and Scott, 1980, Proposition 6.1].

Lemma 4.27 (Lemma A.8 in Petersen and Voigtlaender [2018]). *Let $\beta > 0$ with $\beta = s + r$ where $s \in \mathbb{N}_0$ and $r \in (0,1]$, and let $d \in \mathbb{N}$. Then there exists a constant $C(\beta, d) > 0$ with the following property:*
For each $f \in W^{\beta,\infty}((0,1)^d)$ with $B := \|f\|_{W^{\beta,\infty}((0,1)^d)} \leq \infty$ and any $x_0 \in (0,1)^d$, there is a polynomial $p(x) = \sum_{\|\alpha\|_1 \leq s} c_\alpha (x - x_0)^\alpha$ with $c_\alpha \in [-CB, CB]$ for all $\alpha \in \mathbb{N}_0^d$ with $\|\alpha\|_1 \leq s$ and such that

$$\|p - f\|_{L^\infty((0,1)^d)} \leq CB\|x - x_0\|_2^\beta.$$

In fact, $p = p_{f,x_0}$ is the Taylor polynomial of f of degree s.

Lemma 4.28 (Proposition 6.1 in Dupont and Scott [1980]). *Suppose that $p \in [1, \infty]$, that $m = \bar{m} + \theta$ where $\bar{m} \in \mathbb{N}_0$ and $\theta \in (0,1)$, and that $l = \bar{m} + 1$. Then there is a constant $C = C(n, \phi, d, m)$ such that for $f \in W^{m,p}(D)$*

$$\|f - Q^l f\|_{L^p(D)} \leq C|f|_{W^{m,p}(D)}.$$

Lemma 4.29 (Theorem 6.1 in Dupont and Scott [1980]). *Suppose that $m = \bar{m} + \theta$ where $\bar{m} \in \mathbb{N}_0$ and $\theta \in (0,1)$. Let $l = \bar{m} + 1$. Then there exists a constant $C = C(n, \phi, d, m)$ such that, for $p \in [0, \infty]$ and $f \in W^{m,p}(D)$,*

$$\|f - Q^l f\|_{W^{m,p}(D)} \leq C|f|_{W^{m,p}(D)}.$$

Lemma 4.30. *Let $\beta \geq 1$ and $m := \lfloor \beta \rfloor$. Let B be a ball in $\Omega \subset \mathbb{R}^d$ such that Ω is star-shaped with respect to B and such that its radius $\mathrm{rad} > r_{\max}/2$, where $r_{\max}$ is defined in Definition 3.71. Moreover, let d_Ω be the diameter of Ω, γ be the chunkiness parameter of Ω,*

and $Q^m f$ be the Taylor polynomial of degree m of f averaged over B for any $f \in W^{\beta,\infty}(\Omega)$. Then there exists a constant $C(d,m,\gamma) > 0$ such that

$$|f - Q^m f|_{W^{k,\infty}(\Omega)} \leq C(d,m,\gamma) d_\Omega^{\beta-k} |f|_{W^{\beta,\infty}(\Omega)}, \quad k = 0, 1, \cdots, m.$$

Auxiliary lemmas

Lemma 4.31 (Corollary B.5 in Gühring et al. [2020]). *Let $d, m \in \mathbb{N}$ and $\Omega_1 \subset \mathbb{R}^d, \Omega_2 \subset \mathbb{R}^m$ both be open, bounded, and convex. If $f \in W^{1,\infty}(\Omega_1; \mathbb{R}^m)$ and $g \in W^{1,\infty}(\Omega_2)$ with $\mathrm{rad}(f) \subset \Omega_2$, then for the composition $g \circ f$, it holds that $g \circ f \in W^{1,\infty}(\Omega_1)$ and we have*

$$|g \circ f|_{W^{1,\infty}(\Omega_1)} \leq \sqrt{d} m |g|_{W^{1,\infty}(\Omega_2)} |f|_{W^{1,\infty}(\Omega_1;\mathbb{R}^m)}$$

and

$$\|g \circ f\|_{W^{1,\infty}(\Omega_1)} \leq \sqrt{d} m \max\{\|g\|_{L^\infty(\Omega_2)}, |g|_{W^{1,\infty}(\Omega_2)} |f|_{W^{1,\infty}(\Omega_1;\mathbb{R}^m)}\}.$$

Lemma 4.32 (Corollary B.6 in Gühring et al. [2020]). *Let $f \in W^{1,\infty}(\Omega)$ and $g \in W^{1,\infty}(\Omega)$. Then $fg \in W^{1,\infty}(\Omega)$ and we have*

$$|fg|_{W^{1,\infty}(\Omega)} \leq |f|_{W^{1,\infty}(\Omega)} \|g\|_{L^\infty(\Omega)} + \|f\|_{L^\infty(\Omega)} |g|_{W^{1,\infty}(\Omega)}$$

and

$$\|fg\|_{W^{1,\infty}(\Omega)} \leq \|f\|_{W^{1,\infty}(\Omega)} \|g\|_{L^\infty(\Omega)} + \|f\|_{L^\infty(\Omega)} \|g\|_{W^{1,\infty}(\Omega)}.$$

Lemma 4.33 (Proposition 3.6 in Hon and Yang [2022]). *For any $N, L, s \in \mathbb{N}$ and $\|\alpha\|_1 \leq s$, there exists a deep ReLU network ϕ with the width $9(N+1)+s-1$ and depth $14s^2L$ such that $\|\phi\|_{W^{1,\infty}((0,1)^d)} \leq 18$ and that*

$$\|\phi(x) - x^\alpha\|_{W^{1,\infty}((0,1)^d)} \leq 10s(N+1)^{-7sL}.$$

Approximation with Lipschitz regularity control

In this part, we study the approximation capacity of deep ReLU networks joint with an estimate of the Lipschitz regularity. The strong expressive power of deep ReLU networks has been established in the literature by a localized approximation approach [Yarotsky, 2017, Petersen and Voigtlaender, 2018, Suzuki, 2019, Gühring et al., 2020, Shen et al., 2020, Lu et al., 2021, Shen et al., 2022b]. A recent progress is that Yang et al. [2023] provide a

nearly optimal approximation estimate for functions in the Sobolev space measured by the Sobolev norm. We follow the localized approximation approach, and establish the global Lipschitz continuity and nonasymptotic approximation estimate of deep ReLU networks.

Lemma 4.34. *Let $\beta \geq 1$ with $\beta = s + r$ where $s \in \mathbb{N}_0$ and $r \in (0, 1]$. Given any $f \in W^{\beta,\infty}((0,1)^d)$ with $\|f\|_{W^{\beta,\infty}((0,1)^d)} \leq 1$, for any $N, L \in \mathbb{N}$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(2^d \beta^2 d^{\beta-1} N \log N)$ and depth $\mathcal{O}(d^2 \beta^2 L \log L)$ such that $\|\phi\|_{W^{1,\infty}((0,1)^d)} \lesssim 1$ and*

$$\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim (NL)^{-2\beta/d},$$

where we hide constants depending only on β and d.

Lemma 4.34 is a direct extension of Lemma 3.49 from the class of Lipschitz functions to the class of Sobolev functions with fractional smoothness indices.

Corollary 4.35. *Let $\beta \geq 1$ with $\beta = s + r$ where $s \in \mathbb{N}_0$ and $r \in (0, 1]$. Given any $f \in W^{\beta,\infty}((0,1)^d)$ with $\|f\|_{W^{\beta,\infty}((0,1)^d)} \leq \infty$, for any $N, L \in \mathbb{N}$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(2^d \beta^2 d^{\beta-1} N \log N)$ and depth $\mathcal{O}(d^2 \beta^2 L \log L)$ such that $\|\phi\|_{W^{1,\infty}((0,1)^d)} \lesssim \|f\|_{W^{\beta,\infty}((0,1)^d)}$ and*

$$\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim \|f\|_{W^{\beta,\infty}((0,1)^d)} (NL)^{-2\beta/d},$$

where we hide constants depending only on β and d.

Remark 4.36. The approximation rate in Lemma 4.34 and Corollary 4.35 is nearly optimal for the unit ball of functions in $W^{\beta,\infty}((0,1)^d)$ due to the lower bounds of approximation errors proved in Shen et al. [2020, 2022b] and Lu et al. [2021].

Proof sketch of Lemma 3.49 The proof idea is similar to that of [Yang et al., 2023, Theorem 3], and we divide the proof into three steps.

Step 1. Discretization. We use a partition of unity to discretize the set $(0,1)^d$. As in Definitions 3.52 and 3.53, we construct a partition of unity $\{g_m\}_{m \in \{1,2\}^d}$ on $(0,1)^d$ with $\operatorname{supp}(g_m) \cap (0,1)^d \subset \Omega_m$ for any $m \in \{1,2\}^d$. Then we approximate the partition of unity $\{g_m\}_{m \in \{1,2\}^d}$ by a collection of deep ReLU networks $\{\phi_m\}_{m \in \{1,2\}^d}$ as in Lemma 3.54.

Step 2. Approximation on Ω_m. Given any $m \in \{1,2\}^d$, for each subset $\Omega_m \subset [0,1]^d$, we find a piecewise polynomial function $f_{K,m}$ satisfying

$$\|f_{K,m} - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|f_{K,m} - f\|_{L^\infty(\Omega_m)} \lesssim K^{-\beta},$$

where we omit constants in d. Piecewise polynomial functions can be approximated by deep ReLU networks. Then, following Lu et al. [2021], Yang et al. [2023], we construct a deep ReLU network ψ_m with width $\mathcal{O}(2^d dN \log N)$ and depth $\mathcal{O}(d^2 L \log L)$ such that

$$\|\psi_m - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim (NL)^{-2\beta/d},$$

where we omit constants in β and d.

Step 3. Approximation on $[0,1]^d$. Combining the approximations on each subset Ω_m properly, we construct an approximation of the target function f on the domain $[0,1]^d$. That is, for any $N, L \in \mathbb{N}$, there exists a function ϕ implemented by a deep ReLU network with width $\mathcal{O}(N \log N)$ and depth $\mathcal{O}(L \log L)$ such that

$$\|\phi - f\|_{L^\infty([0,1]^d)} \lesssim (NL)^{-2\beta/d} \quad \text{with} \quad \|\phi\|_{W^{1,\infty}((0,1)^d)} \lesssim 1,$$

where we omit constants in β and d.

Lemma 4.37. *Let $K \in \mathbb{N}$. For any $f \in W^{\beta,\infty}((0,1)^d)$ with $\|f\|_{W^{\beta,\infty}((0,1)^d)} \le 1$ and $m \in \{1,2\}^d$, there exists a piecewise polynomial functions $f_{K,m}$ on Ω_m satisfying*

$$\|f_{K,m} - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|f_{K,m} - f\|_{L^\infty(\Omega_m)} \lesssim K^{-\beta}.$$

with constants in β and d omitted.

Proof of Lemma 4.37. The proof idea follows those of [Gühring et al., 2020, Lemma C.4] and [Yang et al., 2023, Theorem 6]. We leverage approximation properties of averaged Taylor polynomials [Brenner and Scott, 2008, Definition 4.1.3] and the fractional Bramble-Hilbert lemma 4.30 to deduce local estimates and then combine them through a partition of unity to obtain a global estimate. The key observation is that the L^∞ approximation bound can be established while uniformly controlling the Lipschitz constant of the piecewise constant function with a mild regularity assumption on the target function such as $f \in W^{1,\infty}((0,1)^d)$.

Without loss of generality, let us assume $m = m_* := [1,1,\cdots,1]^\top$. Following the proofs of [Gühring et al., 2020, Lemma C.4] and [Yang et al., 2023, Theorem 6], we first

define an extension operator $E : W^{1,\infty}((0,1)^d) \to W^{1,\infty}(\mathbb{R}^d)$ to handle the boundary. Accordingly, let $\tilde{f} := Ef$ and C_E be the norm of the extension operator. Then for any $\Omega \subset \mathbb{R}^d$, it holds that

$$|\tilde{f}|_{W^{1,\infty}(\Omega)} \le \|\tilde{f}\|_{W^{1,\infty}(\Omega)} \le C_E \|f\|_{W^{1,\infty}((0,1)^d)} \le C_E.$$

Next, we define an averaged Taylor polynomial of degree $\lfloor \beta \rfloor$ over $B_{i,K} := \mathbb{B}^d(\frac{8i+3}{8K}, \frac{1}{4K}, \|\cdot\|_2)$ by

$$p_{f,i}(x) := \int_{B_{i,K}} T_y^\beta \tilde{f}(x) \phi_K(y) \mathrm{d}y$$

where ϕ_K is a cut-off function supported on $\bar{B}_{i,K}$ as given in Example 3.67. By Lemma 3.69, it holds that

$$p_{f,i}(x) = \sum_{\|\alpha\|_1 < \lfloor \beta \rfloor} c_{f,i,\alpha} x^\alpha \quad \text{with } |c_{f,i,\alpha}| \le C_2(\beta, d).$$

Step 1. Get local estimates. For any $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K\}^d$, we would like to employ the factional Bramble-Hilbert lemma 3.72 on the subset

$$\Omega_{m_*,i} = \bar{\mathbb{B}}^d\left(\frac{8i+3}{8K}, \frac{3}{8K}, \|\cdot\|_\infty\right) = \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right].$$

It is easy to check the conditions of the Bramble-Hilbert lemma are fulfilled as

$$\frac{1}{4K} \ge \frac{1}{2} \times \frac{3}{8K} = \frac{1}{2} r_{\max}(\Omega_{m_*,i}), \quad \gamma(\Omega_{m_*,i}) = \frac{d_{\Omega_{m_*,i}}}{r_{\max}(\Omega_{m_*,i})} = 2\sqrt{d}.$$

Hence, by the Bramble-Hilbert Lemma 3.72, it yields that

$$\|\tilde{f} - p_{f,i}\|_{L^\infty(\Omega_{m_*,i})} \le C_1(\beta, d) |\tilde{f}|_{W^{1,\infty}(\Omega_{m_*,i})} K^{-\beta}, \quad |\tilde{f} - p_{f,i}|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(\beta, d) |\tilde{f}|_{W^{1,\infty}(\Omega_{m_*,i})}.$$

Combining $|\tilde{f}|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_E$ and the inequalities above, it implies that

$$\|\tilde{f} - p_{f,i}\|_{L^\infty(\Omega_{m_*,i})} \le C_1(\beta, d) C_E K^{-\beta}, \tag{4.21}$$

$$\|\tilde{f} - p_{f,i}\|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(\beta, d) C_E. \tag{4.22}$$

Step 2. Define a partition of unity. We construct a partition of unity in order to combine the local estimates. Let $K \in \mathbb{N}$. For any $0 \le i \le K$, we define $h_i : \mathbb{R} \to \mathbb{R}$ by

$$h_i(x) = h\left(4K\left(x - \frac{8i+3}{8K}\right)\right) \text{ where } h(x) = \begin{cases} 1, & |x| < 3/2, \\ 0, & |x| > 2, \\ 4 - 2|x|, & 3/2 \le |x| \le 2. \end{cases}$$

One can verify that $\{h_i\}_{i=1}^K$ is a partition of unity of $[0,1]$ and $h_i(x) = 1$ for any $x \in \left[\frac{i}{K}, \frac{3+4i}{4K}\right]$. Considering the multidimensional case, for any $x = [x_1, x_2, \cdots, x_d]^\top \in \mathbb{R}^d$ and any $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K\}^d$, let us define

$$h_i(x) := \prod_{j=1}^{d} h_{i_j}(x_j).$$

Then a partition of unity of $[0,1]^d$ is defined by $\{h_i : i \in \{0, 1, \cdots, K\}^d\}$. Moreover, $h_i(x) = 1$ for any $x \in \Omega_{m_*,i} = \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right]$ and $i = [i_1, i_2, \cdots, i_d]^\top \in \{0, 1, \cdots, K\}^d$. By the definition of $h_i(x)$ on $\Omega_{m_*,i}$ and equation 4.21, equation 4.22 , it yields that

$$\|h_i(\tilde{f} - p_{f,i})\|_{L^\infty(\Omega_{m_*,i})} \le \|\tilde{f} - p_{f,i}\|_{L^\infty(\Omega_{m_*,i})} \le C_1(\beta, d) C_E K^{-\beta},$$
$$\|h_i(\tilde{f} - p_{f,i})\|_{W^{1,\infty}(\Omega_{m_*,i})} \le \|\tilde{f} - p_{f,i}\|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(\beta, d) C_E.$$

Step 3. Get global estimates. To deduce the global estimates, we start with defining f_{K,m_*} over Ω_{m_*} by

$$\begin{aligned} f_{K,m_*} &:= \sum_{i \in \{0,1,\cdots,K\}^d} h_i p_{f,i} \\ &= \sum_{i \in \{0,1,\cdots,K\}^d} \sum_{\|\alpha\|_1 < \lfloor\beta\rfloor} h_i c_{f,i,\alpha} x^\alpha \\ &= \sum_{\|\alpha\|_1 < \lfloor\beta\rfloor} \sum_{i \in \{0,1,\cdots,K\}^d} h_i c_{f,i,\alpha} x^\alpha \\ &=: \sum_{\|\alpha\|_1 < \lfloor\beta\rfloor} g_{f,\alpha,m_*}(x) x^\alpha, \end{aligned}$$

where $g_{f,\alpha,m_*}(x)$ is a step function on Ω_{m_*} considering that $g_{f,\alpha,m_*}(x) \equiv c_{f,i,\alpha}$ for any $x \in \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right]$ and $i = [i_1, i_2, \cdots, i_d]^\top$. Then for any $x \in \Omega_{m_*}$, it holds that $|g_{f,\alpha,m_*}(x)| \le C_2(\beta, d)$. Furthermore, the following error bounds hold

$$\|f_{K,m_*} - f\|_{L^\infty(\Omega_{m_*})} \le \max_{i \in \{0,1,\cdots,K\}^d} \|h_i(\tilde{f} - p_{f,i})\|_{L^\infty(\Omega_{m_*,i})} \le C_1(\beta, d) C_E K^{-\beta},$$
$$\|f_{K,m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \le \max_{i \in \{0,1,\cdots,K\}^d} \|h_i(\tilde{f} - p_{f,i})\|_{W^{1,\infty}(\Omega_{m_*,i})} \le C_1(\beta, d) C_E.$$

We complete the proof. □

Lemma 4.38. *Given any $f \in W^{\beta,\infty}((0,1)^d)$ with $\|f\|_{W^{\beta,\infty}((0,1)^d)} \le 1$, for any $N, L \in \mathbb{N}$ and any $m \in \{1,2\}^d$, there exists a deep ReLU network ϕ_m with width $\mathcal{O}(\beta^2 d^{\beta-1} N \log N)$ and depth $\mathcal{O}(\beta^2 L \log L)$ such that*

$$\|\psi_m - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim (NL)^{-2\beta/d},$$

where we omit constants in β and d.

Proof of Lemma 4.38. The idea of proof is similar to those of [Hon and Yang, 2022, Theorem 3.1] and [Yang et al., 2023, Theorem 7]. For completeness, we provide a concrete proof in the following. Without loss of generality, we consider $m = m_* := [1,1,\cdots,1]^\top$. Given $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, by Lemma 4.37, we have

$$\|f_{K,m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \lesssim 1,$$
$$\|f_{K,m_*} - f\|_{L^\infty(\Omega_{m_*})} \lesssim K^{-\beta} \lesssim (NL)^{-2\beta/d},$$

where $f_{K,m_*} = \sum_{\|\alpha\|_1 < \lfloor\beta\rfloor} g_{f,\alpha,m_*}(x) x^\alpha$ for $x \in \prod_{j=1}^d \left[\frac{i_j}{K}, \frac{3+4i_j}{4K}\right]$ and $i = [i_1, i_2, \cdots, i_d]^\top \in \{0,1,\cdots,K-1\}^d$. The insight is to approximate f_{K,m_*} with deep ReLU networks. Let $\delta = 1/(4K) \le 1/(3K)$ in Lemma 3.81. Then by Lemma 3.81, there exists a deep ReLU network $\phi_1(x)$ with width $4N+5$ and depth $4L+4$ such that

$$\phi_1(x) = k, \quad x \in \left[\frac{k}{K}, \frac{k+1}{K} - \frac{1}{4K}\right], \quad k = 0,1,\cdots,K-1.$$

We further define

$$\phi_2(x) = \left[\frac{\phi_1(x_1)}{K}, \frac{\phi_1(x_2)}{K}, \cdots, \frac{\phi_1(x_d)}{K}\right]^\top.$$

For each $p = 0,1,\cdots,K^d - 1$, there exists a bijection

$$\eta(p) = [\eta_1, \eta_2, \cdots, \eta_d]^\top \in \{0,1,\cdots,K-1\}^d$$

satisfying $\sum_{j=1}^d \eta_j K^{j-1} = p$. We also define

$$\xi_{\alpha,p} := \frac{g_{f,\alpha,m_*}(\eta(p)/K) + C_2(\beta,d)}{2C_2(\beta,d)} \in [0,1].$$

Then, due to Lemma 3.82, there exists a deep ReLU network $\tilde{\phi}_\alpha$ with width $16\lceil\beta\rceil(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that $|\tilde{\phi}_\alpha(p) - \xi_{\alpha,p}| \le (NL)^{-2\lceil\beta\rceil}$ for $p = 0,1,\cdots,K^d-1$. Let us define

$$\phi_\alpha(x) := 2C_2(\beta,d)\tilde{\phi}_\alpha\left(\sum\nolimits_{j=1}^d \eta_j K^j\right) - C_2(\beta,d).$$

Then it is clear that

$$|\phi_\alpha(\eta(p)/N) - g_{f,\alpha,m_*}(\eta(p)/N)| = 2C_2(\beta,d)|\tilde{\phi}_\alpha(x) - \xi_{\alpha,p}| \le 2C_2(\beta,d)(NL)^{-2\lceil\beta\rceil}.$$

Since $\phi_\alpha \circ \phi_2(x) - g_{f,\alpha,m_*}(x)$ is a step function whose first order weak derivative is 0 in Ω_{m_*}, it further implies that

$$\begin{aligned}\|\phi_\alpha \circ \phi_2(x) - g_{f,\alpha,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})} &= \|\phi_\alpha \circ \phi_2(x) - g_{f,\alpha,m_*}(x)\|_{L^\infty(\Omega_{m_*})} \\ &\le 2C_2(\beta,d)(NL)^{-2\lceil\beta\rceil}.\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}\|\phi_\alpha \circ \phi_2(x)\|_{W^{1,\infty}(\Omega_{m_*})} &\le \|\phi_\alpha \circ \phi_2(x) - g_{f,\alpha,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})} + \|g_{f,\alpha,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})} \\ &\le 3C_2(\beta,d).\end{aligned}$$

According to Lemma 4.33, there exists a deep ReLU network $\phi_{3,\alpha}$ with width $9(N+1)+\lceil\beta\rceil-1$ and depth $14\lceil\beta\rceil^2 L$ such that $\|\phi_{3,\alpha}\|_{W^{1,\infty}((0,1)^d)} \le 18$ and that

$$\|\phi_{3,\alpha}(x) - x^\alpha\|_{W^{1,\infty}((0,1)^d)} \le 10\lceil\beta\rceil(N+1)^{-7\lceil\beta\rceil L}.$$

Let $C_3(\beta,d) := \max\{3C_2(\beta,d), 18\}$. Then it holds that

$$\max\{\|\phi_\alpha \circ \phi_2(x)\|_{W^{1,\infty}(\Omega_{m_*})}, \|\phi_{3,\alpha}(x)\|_{W^{1,\infty}(\Omega_{m_*})}\} \le C_3(\beta,d).$$

Due to Lemma 3.75, there exists a deep ReLU network ϕ_4 with width $15(N+1)$ and depth $4\lceil\beta\rceil(L+1)$ such that $\|\phi_4\|_{W^{1,\infty}((-C_3,C_3)^d)} \le 12C_3(\beta,d)^2$ and that

$$\|\phi_4(x,y) - xy\|_{W^{1,\infty}((-C_3,C_3)^d)} \le 6C_3(\beta,d)^2(N+1)^{-2\lceil\beta\rceil(L+1)}.$$

Then we are ready to construct the deep ReLU network ϕ_{m_*} to approximate f_{K,m_*} over Ω_{m_*} by

$$\psi_{m_*}(x) := \sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} \phi_4\Big(\phi_\alpha \circ \phi_2(x), \phi_{3,\alpha}(x)\Big).$$

We establish the approximation bounds of ψ_{m_*} with both the L^∞ norm and the $W^{1,\infty}$

norm in the following. The $W^{1,\infty}$ error can be decomposed as

$$
\begin{aligned}
&\|\psi_{m_*}(x) - f_{K,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})} \\
&= \Big\| \sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} \phi_4\Big(\phi_\alpha \circ \phi_2(x), \phi_{3,\alpha}(x)\Big) - f_{K,m_*}(x) \Big\|_{W^{1,\infty}(\Omega_{m_*})} \\
&\le \sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} \Big\|\phi_4\Big(\phi_\alpha \circ \phi_2(x), \phi_{3,\alpha}(x)\Big) - g_{f,\alpha,m_*}(x)x^\alpha \Big\|_{W^{1,\infty}(\Omega_{m_*})} \\
&\le \underbrace{\sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} \Big\|\phi_4\Big(\phi_\alpha \circ \phi_2(x), \phi_{3,\alpha}(x)\Big) - \phi_\alpha \circ \phi_2(x)\phi_{3,\alpha}(x) \Big\|_{W^{1,\infty}(\Omega_{m_*})}}_{:=\mathcal{E}_1} \\
&\quad + \underbrace{\sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} \Big\|\phi_\alpha \circ \phi_2(x)\phi_{3,\alpha}(x) - g_{f,\alpha,m_*}(x)\phi_{3,\alpha}(x) \Big\|_{W^{1,\infty}(\Omega_{m_*})}}_{:=\mathcal{E}_2} \\
&\quad + \underbrace{\sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} \Big\|g_{f,\alpha,m_*}(x)\phi_{3,\alpha}(x) - g_{f,\alpha,m_*}(x)x^\alpha \Big\|_{W^{1,\infty}(\Omega_{m_*})}}_{:=\mathcal{E}_3} \\
&\le \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3.
\end{aligned}
$$

We note that $\sum_{\|\alpha\|_1<\lfloor\beta\rfloor} 1 \le \beta d^{\beta-1}$. By Lemma 4.31, the term $\mathcal{E}_1$ can be bounded by

$$
\begin{aligned}
\mathcal{E}_1 \le &\sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} 2\sqrt{d} \max\Big\{\|\phi_4(x,y) - xy\|_{L^\infty((-C_3,C_3)^d)}, |\phi_4(x,y) - xy|_{W^{1,\infty}((-C_3,C_3)^d)} \times \\
&\max\{|\phi_\alpha \circ \phi_2(x)|_{W^{1,\infty}(\Omega_{m_*})}, |\phi_{3,\alpha}(x)|_{W^{1,\infty}(\Omega_{m_*})}\}\Big\} \\
\le &\sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} 2\sqrt{d} \max\Big\{\|\phi_4(x,y) - xy\|_{L^\infty((-C_3,C_3)^d)}, C_3(\beta,d)|\phi_4(x,y) - xy|_{W^{1,\infty}((-C_3,C_3)^d)}\Big\} \\
\le &12\beta d^\beta C_3(\beta,d)^2(C_3(\beta,d)+1)(N+1)^{-2\lceil\beta\rceil(L+1)}.
\end{aligned}
$$

By Lemma 4.32, the term $\mathcal{E}_2$ can be bounded by

$$
\begin{aligned}
\mathcal{E}_2 &\le \sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} 2\|\phi_\alpha \circ \phi_2(x) - g_{f,\alpha,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})}\|\phi_{3,\alpha}(x)\|_{W^{1,\infty}(\Omega_{m_*})} \\
&\le 72\beta d^{\beta-1} C_2(\beta,d)(NL)^{-2\lceil\beta\rceil}.
\end{aligned}
$$

Similarly, by Lemma 4.32, the term $\mathcal{E}_3$ can be bounded by

$$
\begin{aligned}
\mathcal{E}_3 &\le \sum\nolimits_{\|\alpha\|_1<\lfloor\beta\rfloor} 2\|\phi_{3,\alpha}(x) - x^\alpha\|_{W^{1,\infty}(\Omega_{m_*})}\|g_{f,\alpha,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})} \\
&\le 20\beta^2 d^{\beta-1} C_2(\beta,d)(N+1)^{-7\lceil\beta\rceil L}.
\end{aligned}
$$

Using that

$$
(N+1)^{-7\lceil\beta\rceil L} \le (N+1)^{-2\lceil\beta\rceil(L+1)} \le (NL)^{-2\lceil\beta\rceil},
$$

we derive the following error bound

$$\|\psi_{m_*}(x) - f_{K,m_*}(x)\|_{W^{1,\infty}(\Omega_{m_*})} \le \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 \le C_4(\beta, d)(NL)^{-2\lceil\beta\rceil},$$

where $C_4(\beta,d) := 12\beta d^{\beta} C_3(\beta,d)^2(C_3(\beta,d)+1)+72\beta d^{\beta-1}C_2(\beta,d)+20\beta^2 d^{\beta-1}C_2(\beta,d)$. By the triangle inequalities for $\|\cdot\|_{L^\infty(\Omega_{m_*})}$ and $\|\cdot\|_{W^{1,\infty}(\Omega_{m_*})}$, it is straightforward to show that

$$\|\psi_{m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \le \|\psi_{m_*} - f_{K,m_*}\|_{W^{1,\infty}(\Omega_{m_*})} + \|f_{K,m_*} - f\|_{W^{1,\infty}(\Omega_{m_*})} \lesssim 1,$$
$$\|\psi_{m_*} - f\|_{L^{\infty}(\Omega_{m_*})} \le \|\psi_{m_*} - f_{K,m_*}\|_{L^{\infty}(\Omega_{m_*})} + \|f_{K,m_*} - f\|_{L^{\infty}(\Omega_{m_*})} \lesssim (NL)^{-2\beta/d}.$$

In the end, we calculate the width and depth of the deep ReLU network implementing $\psi_{m_*} = \sum_{\|\alpha\|_1<\lfloor\beta\rfloor} \phi_4\Big(\phi_\alpha \circ \phi_2(x), \phi_{3,\alpha}(x)\Big)$. Recall that (1) ϕ_α has width $\mathcal{O}(N\log N)$ and depth $\mathcal{O}(L\log L)$; (2) ϕ_2 has width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$; (3) $\phi_{3,\alpha}$ has width $\mathcal{O}(N\vee\beta)$ and depth $\mathcal{O}(\beta^2 L)$; (4) ϕ_4 has with width $\mathcal{O}(N)$ and depth $\mathcal{O}(\beta L)$. Thus, the deep ReLU network of ψ_{m_*} is constructed with width $\mathcal{O}(\beta^2 d^{\beta-1} N\log N)$ and depth $\mathcal{O}(\beta^2 L\log L)$ for approximating f on the domain Ω_{m_*}. This completes the proof. □

Proof of Lemma 3.49. We proceed in a similar way as the proof of [Yang et al., 2023, Theorem 3]. By Lemma 3.54, there exists a sequence of ReLU networks $\{\phi_m\}_{m\in\{1,2\}^d}$ such that for any $m\in\{1,2\}^d$,

$$\|\phi_m - g_m\|_{W^{1,\infty}((0,1)^d)} \le 50 d^{5/2}(N+1)^{-4d\beta L}.$$

Each ϕ_m is implemented by a deep ReLU network with width $\mathcal{O}(dN)$ and depth $\mathcal{O}(d^2\beta L)$. By Lemma 4.38, there exists a collection of ReLU networks $\{\psi_m\}_{m\in\{1,2\}^d}$ such that for any $m\in\{1,2\}^d$,

$$\|\psi_m - f\|_{W^{1,\infty}(\Omega_m)} \lesssim 1, \quad \|\psi_m - f\|_{L^{\infty}(\Omega_m)} \lesssim (NL)^{-2\beta/d}$$

where we omit constants in d. Each ψ_m is implemented by a deep ReLU network with width $\mathcal{O}(\beta^2 d^{\beta-1} N\log N)$ and depth $\mathcal{O}(\beta^2 L\log L)$. Before proceeding, it is useful to estimate $\|\phi_m\|_{L^\infty(\Omega_m)}$, $\|\phi_m\|_{W^{1,\infty}(\Omega_m)}$, $\|\psi_m\|_{L^\infty(\Omega_m)}$, and $\|\psi_m\|_{W^{1,\infty}(\Omega_m)}$ as follows

$$\|\phi_m\|_{L^\infty(\Omega_m)} \le \|\phi_m\|_{L^\infty([0,1]^d)} \le \|g_m\|_{L^\infty([0,1]^d)} + \|\phi_m - g_m\|_{L^\infty([0,1]^d)} \le 1 + 50d^{5/2} \lesssim d^{5/2},$$

$$\begin{aligned}\|\phi_m\|_{W^{1,\infty}(\Omega_m)} &\le \|\phi_m\|_{W^{1,\infty}([0,1]^d)} \le \|g_m\|_{W^{1,\infty}([0,1]^d)} + \|\phi_m - g_m\|_{W^{1,\infty}([0,1]^d)} \\ &\le 4\lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor + 50 d^{5/2},\end{aligned}$$

$$\|\psi_m\|_{L^\infty(\Omega_m)} \le \|f\|_{L^\infty(\Omega_m)} + \|\psi_m - f\|_{L^\infty(\Omega_m)} \lesssim 1,$$

$$\|\psi_m\|_{W^{1,\infty}(\Omega_m)} \le \|f\|_{W^{1,\infty}([0,1]^d)} + \|\psi_m - f\|_{W^{1,\infty}([0,1]^d)} \lesssim 1.$$

Let $B_1 := \max_{m\in\{1,2\}^d}\{\|\phi_m\|_{L^\infty(\Omega_m)}, \|\psi_m\|_{L^\infty(\Omega_m)}\}$, then it yields that $B_1 \lesssim d^{5/2}$ by the estimates of $\|\phi_m\|_{L^\infty(\Omega_m)}$ and $\|\psi_m\|_{W^{1,\infty}(\Omega_m)}$. Let $B_2 := \max_{m\in\{1,2\}^d}\{\|\phi_m\|_{W^{1,\infty}(\Omega_m)}, \|\psi_m\|_{W^{1,\infty}(\Omega_m)}\}$. Similarly, it yields that $B_2 \lesssim (NL)^{2/d} + d^{5/2}$. By Lemma 3.75, for any $N, L \in \mathbb{N}$, there exists a deep ReLU network $\phi_{\times,B_1}$ with width $15(N+1)$ and depth $16\beta L$ such that $\|\phi_{\times,B_1}\|_{W^{1,\infty}((-B_1,B_1)^2)} \le 12B_1^2$ and

$$\|\phi_{\times,B_1}(x,y) - xy\|_{W^{1,\infty}((-B_1,B_1)^2)} \le 6B_1^2(N+1)^{-8\beta L}.$$

To obtain a global estimate on $[0,1]^d$, we combine the local estimate $\{\psi_m\}_{m\in\{1,2\}^d}$ and the approximate partition of unity $\{\phi_m\}_{m\in\{1,2\}^d}$. Let us construct the global approximation function ϕ by

$$\phi(x) := \sum_{m\in\{1,2\}^d} \phi_{\times,B_1}(\phi_m(x), \psi_m(x)). \tag{4.23}$$

Next, we bound the error of the global approximation estimate by

$$\begin{aligned}
\|f-\phi\|_{L^\infty([0,1]^d)} =& \|\sum\nolimits_{m\in\{1,2\}^d} g_m f - \phi\|_{L^\infty([0,1]^d)} \\
\le& \underbrace{\|\sum\nolimits_{m\in\{1,2\}^d}[g_m f - \phi_m\psi_m]\|_{L^\infty([0,1]^d)}}_{=:\mathcal{R}_1} \\
&+ \underbrace{\|\sum\nolimits_{m\in\{1,2\}^d}[\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{L^\infty([0,1]^d)}}_{=:\mathcal{R}_2}
\end{aligned}$$

and

$$\begin{aligned}
\|f-\phi\|_{W^{1,\infty}((0,1)^d)} =& \|\sum\nolimits_{m\in\{1,2\}^d} g_m f - \phi\|_{W^{1,\infty}((0,1)^d)} \\
\le& \underbrace{\|\sum\nolimits_{m\in\{1,2\}^d}[g_m f - \phi_m\psi_m]\|_{W^{1,\infty}((0,1)^d)}}_{=:\mathcal{R}_3} \\
&+ \underbrace{\|\sum\nolimits_{m\in\{1,2\}^d}[\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{W^{1,\infty}((0,1)^d)}}_{=:\mathcal{R}_4}.
\end{aligned}$$

It remains to bound $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$, and $\mathcal{R}_4$, respectively. For the term $\mathcal{R}_1$, it holds

$$
\begin{aligned}
\mathcal{R}_1 &\leq \sum_{m\in\{1,2\}^d} \|g_m f - \phi_m \psi_m\|_{L^\infty([0,1]^d)} \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{L^\infty([0,1]^d)} + \|\phi_m (f - \psi_m)\|_{L^\infty([0,1]^d)}\right] \\
&= \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{L^\infty([0,1]^d)} + \|\phi_m (f - \psi_m)\|_{L^\infty(\Omega_m)}\right] \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|g_m - \phi_m\|_{L^\infty([0,1]^d)} \|f\|_{L^\infty([0,1]^d)} + \|\phi_m\|_{L^\infty(\Omega_m)} \|f - \psi_m\|_{L^\infty(\Omega_m)}\right] \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|g_m - \phi_m\|_{W^{1,\infty}([0,1]^d)} \|f\|_{W^{1,\infty}([0,1]^d)} + \|\phi_m\|_{L^\infty(\Omega_m)} \|f - \psi_m\|_{L^\infty(\Omega_m)}\right] \\
&\leq 2^d [50 d^{5/2} (N+1)^{-4d\beta L} + (1 + 50 d^{5/2})(NL)^{-2\beta/d}] \\
&\lesssim (NL)^{-2\beta/d}
\end{aligned}
$$

where we use $(NL)^{2\beta/d} \leq (N+1)^{4d\beta L}$ to derive the last inequality and hide a prefactor in d. For the term $\mathcal{R}_3$, it holds

$$
\begin{aligned}
\mathcal{R}_3 &\leq \sum_{m\in\{1,2\}^d} \|g_m f - \phi_m \psi_m\|_{W^{1,\infty}((0,1)^d)} \\
&\leq \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{W^{1,\infty}((0,1)^d)} + \|\phi_m (f - \psi_m)\|_{W^{1,\infty}((0,1)^d)}\right] \\
&= \sum_{m\in\{1,2\}^d} \left[\|(g_m - \phi_m) f\|_{W^{1,\infty}((0,1)^d)} + \|\phi_m (f - \psi_m)\|_{W^{1,\infty}(\Omega_m)}\right] \\
&\leq \sum_{m\in\{1,2\}^d} \Big[\|g_m - \phi_m\|_{W^{1,\infty}((0,1)^d)} \|f\|_{W^{1,\infty}((0,1)^d)} \\
&\qquad + \|\phi_m\|_{W^{1,\infty}(\Omega_m)} \|f - \psi_m\|_{L^\infty(\Omega_m)} + \|\phi_m\|_{L^\infty(\Omega_m)} \|f - \psi_m\|_{W^{1,\infty}(\Omega_m)}\Big] \\
&\lesssim 2^d [50 d^{5/2} (N+1)^{-4d\beta L} + (4\lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor + 50 d^{5/2})(NL)^{-2\beta/d} + (1 + 50 d^{5/2})] \\
&\lesssim 1
\end{aligned}
$$

where we use $(NL)^{2\beta/d} \leq (N+1)^{4d\beta L}$ to derive the last inequality and hide a prefactor

in d. For the terms $\mathcal{R}_2$ and $\mathcal{R}_4$, it holds

$$
\begin{aligned}
\mathcal{R}_2 \le \mathcal{R}_4 &\le \sum_{m\in\{1,2\}^d} \|[\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{W^{1,\infty}((0,1)^d)} \\
&\le \sum_{m\in\{1,2\}^d} \|[\phi_m\psi_m - \phi_{\times,B_1}(\phi_m(x),\psi_m(x))]\|_{W^{1,\infty}(\Omega_m)} \\
&\le \sum_{m\in\{1,2\}^d} 2\sqrt{d}\max\Big\{\|\phi_{\times,B_1}(x,y)-xy\|_{L^\infty((-B_1,B_1)^2)}, \\
&\qquad\qquad |\phi_{\times,B_1}(x,y)-xy|_{W^{1,\infty}((-B_1,B_1)^2)}\cdot\max\{|\phi_m|_{W^{1,\infty}(\Omega_m)},|\psi_m|_{W^{1,\infty}(\Omega_m)}\}\Big\} \\
&\le \sum_{m\in\{1,2\}^d} 2\sqrt{d}\|\phi_{\times,B_1}(x,y)-xy\|_{W^{1,\infty}((-B_1,B_1)^2)}\cdot\max\{\|\phi_m\|_{W^{1,\infty}(\Omega_m)},\|\psi_m\|_{W^{1,\infty}(\Omega_m)}\} \\
&\le \sum_{m\in\{1,2\}^d} 12\sqrt{d}B_1^2(N+1)^{-8\beta L}B_2 \\
&\lesssim 2^d\sqrt{d}d^5(N+1)^{-8\beta L}((NL)^{2\beta/d}+d^{5/2}) \\
&\lesssim 2^d\sqrt{d}d^5(d^{5/2}(NL)^{2\beta/d})(N+1)^{-8\beta L} \\
&\lesssim (NL)^{2\beta/d}(N+1)^{-8\beta L} \\
&\lesssim (NL)^{-2\beta/d}
\end{aligned}
$$

where we use $(NL)^{2\beta/d} \le (N+1)^{4\beta L}$ in the last inequality and hide constants depending only on d. Combining the estimates of $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$, and $\mathcal{R}_4$, we have

$$\|f-\phi\|_{L^\infty([0,1]^d)} \le \mathcal{R}_1 + \mathcal{R}_2 \lesssim (NL)^{-2\beta/d}$$

and

$$\|f-\phi\|_{W^{1,\infty}([0,1]^d)} \le \mathcal{R}_3 + \mathcal{R}_4 \lesssim 1 + (NL)^{-2\beta/d} \lesssim 1.$$

It is easy to see

$$\|\phi\|_{W^{1,\infty}([0,1]^d)} \le \|f\|_{W^{1,\infty}([0,1]^d)} + \|f-\phi\|_{W^{1,\infty}([0,1]^d)} \lesssim 1.$$

Lastly, we need to calculate the complexity of the deep ReLU network. By the definition of ϕ in (4.23), we know that ϕ consists of $\mathcal{O}(2^d)$ parallel subnetworks listed as follows:

- $\phi_{\times,B_1}$ with width $\mathcal{O}(N)$ and depth $\mathcal{O}(\beta L)$;
- ϕ_m with width $\mathcal{O}(dN)$ and depth $\mathcal{O}(d^2\beta L)$;

- ψ_m with width $\mathcal{O}(\beta^2 d^{\beta-1} N \log N)$ and depth $\mathcal{O}(\beta^2 L \log L)$.

Hence, the deep ReLU network implementing the function ϕ has width $\mathcal{O}(2^d \beta^2 d^{\beta-1} N \log N)$ and depth $\mathcal{O}(d^2 \beta^2 L \log L)$. □

Proof of Corollary 3.50. The proof is completed by employing Lemma 3.49 on

$$\bar{f} := f / \|f\|_{W^{1,\infty}((0,1)^d)}.$$

□

Proof of Lemma 4.20

This lemma is yielded by applying Corollary 4.35 to the target function f^*.

Proof of Lemma 4.25

This lemma is also yielded by applying Corollary 4.35 to the target function M^*.

4.9.4 Proofs of stochastic error bounds

In this section, we present proofs of lemmas for bounding the stochastic errors.

Proof of Lemma 4.21

The proof is almost identical to that of Lemma 3.38.

Proof of Lemma 4.26

The proof follows from that of Lemma 3.38.

Chapter 5

Conclusions and Discussions

In this thesis, we have investigated the theoretical properties of both ODE-based and SDE-based generative models from the viewpoints of regularity, approximation, and convergence analyses.

Through a unified framework and rigorous analysis, we have established the well-posedness of the Gaussian interpolation flows, shedding light on their capabilities and limitations. We have examined the Lipschitz regularity of the corresponding flow maps for several rich classes of probability measures. When applied to generative modeling based on Gaussian denoising, we have shown that GIFs possess auto-encoding and cycle consistency properties at the population level. Additionally, we have established stability error bounds for the errors accumulated during the process of learning GIFs.

We have established non-asymptotic error bounds for the CNF distribution estimator trained via flow matching, using the Wasserstein-2 distance. Assuming that the target distribution belongs to several rich classes of probability distributions, we have established Lipschitz regularity properties of the velocity field for simulation-free CNFs defined with linear interpolation. To meet the regularity requirements of flow matching estimators, we have developed L^∞ approximation bounds of deep ReLU networks for Lipschitz functions, along with Lipschitz regularity control of the constructed deep ReLU networks. By integrating the regularity results, the deep approximation bounds, and perturbation analyses of ODE flows, we have shown that the convergence rate of the CNF distribution estimator is $\widetilde{\mathcal{O}}(n^{-1/(d+5)})$, up to a polylogarithmic prefactor of n. Our error analysis framework can be extended to study more general CNFs based on interpolation, beyond the CNFs constructed with linear interpolation.

We have proved that a pre-trained large diffusion model can gain a faster convergence rate from the Bayesian fine-tuning procedure when adapted to perform conditional generation tasks. This improvement in the convergence rate justifies that a pre-trained large diffusion model would perform better on a downstream conditional gen-

eration task than a standard conditional diffusion model, whenever an appropriate fine-tuning procedure is implemented.

However, the theoretical results established in this thesis do have a few limitations due to the assumptions and the techniques used. Several questions deserve further investigation. Firstly, it would be interesting to consider target distributions with general regularity properties and investigate the resulting high-order smoothness properties of the corresponding velocity fields. Accordingly, the well-posedness of the flow models deserves further exploration if we relax the assumptions on the target distribution. Secondly, the inevitability of the time singularity of the velocity field remains unclear and warrants further analysis, as we have not provided a lower bound on the Lipschitz constant in the time variable. This is a challenging problem that requires more effort and careful analysis. Thirdly, it would be interesting to derive general non-asymptotic error bounds and convergence rates for CNF distribution estimators under general smoothness conditions. For this purpose, we need to combine the general smoothness properties of velocity fields with the deep neural network approximation theory. Fourthly, it remains interesting to conduct rigorous analyses of one-step flow models, a nascent family of ODE-based generative models designed for fast generation and computational efficiency [Song et al., 2023a, Kim et al., 2023, Song and Dhariwal, 2023]. Lastly, the analysis of fine-tuning is based on the regularity assumptions of the denoising functions. It is worth considering a new framework for analyzing the fine-tuning in a more practical setting.

Bibliography

Robert A. Adams and John J.F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, second edition, 2003.

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

Michael S. Albergo, Nicholas M. Boffi, Michael Lindsey, and Eric Vanden-Eijnden. Multimarginal generative modeling with stochastic interpolants. *arXiv preprint arXiv:2310.03695*, 2023a.

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

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

Luigi Ambrosio and Gianluca Crippa. Continuity equations and ODE flows with non-smooth velocity. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 144(6):1191–1244, 2014.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: In metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Luigi Ambrosio, Sebastiano N. Golo, and Francesco S. Cassano. Classical flows of vector fields with exponential or sub-exponential summability. *Journal of Differential Equations*, 372:458–504, 2023.

Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining deep generative models via discriminator gradient flow. In *International Conference on Learning Representations*, 2021.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*, volume 9. Cambridge University Press, 1999.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Seminaire de probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Keith Ball, Franck Barthe, and Assaf Naor. Entropy jumps in the presence of a spectral gap. *Duke Mathematical Journal*, 119(1):41 – 63, 2003.

Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear VC dimension bounds for piecewise polynomial networks. In *Advances in Neural Information Processing Systems*, volume 11, 1998.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 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, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. 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.

Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural ODEs. In *Advances in Neural Information Processing Systems*, volume 34, pages 21325–21337. Curran Associates, Inc., 2021.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Sergey G. Bobkov and Michel Ledoux. From Brunn-Minkowski to Brascamp-Lieb and to logarithmic Sobolev inequalities. *Geometric and Functional Analysis*, 10(5):1028–1052, 2000.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Herm J. Brascamp and Elliott H. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4): 366–389, 1976.

Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*, chapter 4, pages 93–127. Springer New York, New York, NY, third edition, 2008.

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

Luis A. Caffarelli. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563, 2000.

Tony T. Cai and Yihong Wu. Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4):2217–2232, 2014.

Patrick Cattiaux and Arnaud Guillin. Semi log-concave Markov diffusions. In Catherine Donati-Martin, Antoine Lejay, and Alain Rouault, editors, *Séminaire de probabilités XLVI*, pages 231–292. Springer International Publishing, Cham, 2014.

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

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4735–4763. PMLR, 2023a.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative sdversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 03 2022.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *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.

Ricky T.Q. Chen and Yaron Lipman. Riemannian flow matching on general geometries. *arXiv preprint arXiv:2302.03660*, 2023.

Ricky T.Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *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 *Thirty-seventh Conference on Neural Information Processing Systems*, 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.

Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4462–4484. PMLR, 2023e.

Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023a.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *arXiv preprint arXiv:2310.17582*, 2023b.

Sinho Chewi and Aram-Alexandre Pooladian. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *arXiv preprint arXiv:2203.04954*, 2022.

Kai Lai Chung. *A course in probability theory*. Academic press, 2001.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-Gaussian distributions. In *The Twelfth International Conference on Learning Representations*, 2024.

Maria Colombo, Alessio Figalli, and Yash Jhaveri. Lipschitz changes of variables between perturbations of log-concave measures. *Annali della Scuola Normale Superiore di Pisa. Classe di scienze*, 17(4):1491–1519, 2017.

Dario Cordero-Erausquin. Transport inequalities for log-concave measures, quantitative forms, and applications. *Canadian Journal of Mathematics*, 69(3):481–501, 2017.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity. *arXiv preprint arXiv:2310.03575*, 2023.

Yin Dai, Yuan Gao, Jian Huang, Yuling Jiao, Lican Kang, and Jin Liu. Lipschitz transport maps via the Föllmer flow. *arXiv preprint arXiv:2309.03490*, 2023.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Ludwig Danzer, Branko Grünbaum, and Victor Klee. Helly's theorem and its relatives. In *Proceedings of Symposia in Pure Mathematics: Convexity*, volume VII, pages 101–180, Providence, RI, 1963. American Mathematical Society.

Ingrid Daubechies, Ronald DeVore, Simon Foucart, Boris Hanin, and Guergana Petrova. Nonlinear approximation and (deep) ReLU networks. *Constructive Approximation*, 55 (1):127–172, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc., 2021.

Pierre Del Moral and Sumeetpal S. Singh. Backward Itô–Ventzell and stochastic interpolation formulae. *Stochastic Processes and their Applications*, 154:197–250, 2022.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Ronald A. DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63:469–478, 1989.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

Todd Dupont and Ridgway Scott. Polynomial approximation of functions in Sobolev spaces. *Mathematics of Computation*, 34(150):441–463, 1980.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017.

Alex Dytso, Martina Cardone, and Ian Zieder. Meta derivative identity for the conditional expectation. *IEEE Transactions on Information Theory*, 69(7):4284–4302, 2023a.

Alex Dytso, H. Vincent Poor, and Shlomo Shamai Shitz. Conditional mean estimation in Gaussian noise: A meta derivative identity with applications. *IEEE Transactions on Information Theory*, 69(3):1883–1898, 2023b.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Ronen Eldan and James R. Lee. Regularization under diffusion and anticoncentration of the information content. *Duke Mathematical Journal*, 167(5):969–993, 2018.

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. *arXiv preprint arXiv:2403.03206*, 2024.

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, second edition, 2010.

Jianqing Fan and Yihong Gu. Factor augmented sparse throughput deep ReLU neural networks for high dimensional regression. *Journal of the American Statistical Association*, 119(548):2680–2694, 2024.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 6185–6215. PMLR, 2022.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Max Fathi, Dan Mikulincer, and Yair Shenfeld. Transportation onto log-Lipschitz perturbations. *arXiv preprint arXiv:2305.03786*, 2023.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ODE: The world of Jacobian and kinetic regularization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 3154–3164. PMLR, 2020.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv preprint arXiv:2311.11003*, 2023.

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

Yuan Gao, Jian Huang, Yuling Jiao, Jin Liu, Xiliang Lu, and Zhijian Yang. Deep generative learning via Euler particle transport. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145, pages 336–368. PMLR, 2022.

Yuan Gao, Jian Huang, , and Yuling Jiao. Gaussian interpolation flows. *Journal of Machine Learning Research*, 25(253):1–52, 2024a.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Will Grathwohl, Ricky T.Q. Chen, Jesse Bettencourt, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97 (4):1061–1083, 1975.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05): 803–859, 2020.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Ernst Hairer, Gerhard Wanner, and Syvert P. Nørsett. *Solving Ordinary Differential Equations I: Nonstiff Problems*, chapter I, pages 1–128. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

Philip Hartman. *Dependence on Initial Conditions and Parameters*, chapter V, pages 93–116. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002a.

Philip Hartman. *Existence*, chapter II, pages 8–23. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002b.

Charles P. Hatsell and Loren W. Nolte. Some geometric properties of the likelihood ratio (corresp.). *IEEE Transactions on Information Theory*, 17(5):616–618, 1971.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Sean Hon and Haizhao Yang. Simultaneous neural network approximation for smooth functions. *Neural Networks*, 154:152–164, 2022.

Ding Huang, Jian Huang, Ting Li, and Guohao Shen. Conditional stochastic interpolation for generative learning. *arXiv preprint arXiv:2312.05579*, 2023.

Ding Huang, Ting Li, and Jian Huang. Bayesian power steering: An effective approach for domain adaptation of diffusion models. In *Forty-first International Conference on Machine Learning*, 2024.

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.

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.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Rie Johnson and Tong Zhang. Composite functional gradient learning of generative adversarial models. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2371–2379. PMLR, 2018.

Rie Johnson and Tong Zhang. A framework of composite functional gradient methods for generative adversarial models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):17–32, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pages 26565–26577. Curran Associates, Inc., 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. *arXiv preprint arXiv:2310.02279*, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Bo'az Klartag. High-dimensional distributions with convexity properties. In *Proceedings of the Fifth European Congress of Mathematics*, pages 401–417, Amsterdam, 14 July–18 July 2010. European Mathematical Society Publishing House.

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.

Jeroen S.W. Lamb and John A.G. Roberts. Time-reversal symmetry in dynamical systems: A survey. *Physica D: Nonlinear Phenomena*, 112(1-2):1–39, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015.

Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *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 *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pages 946–985. PMLR, 2023.

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*, 2024.

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

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. *arXiv preprint arXiv:2209.14577*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 4104–4113. PMLR, 2019.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion ODEs by high order denoising score matching. In *Proceedings of the 39th International Conference on Machine Learning*, pages 14429–14460. PMLR, 2022a.

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 *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787. Curran Associates, Inc., 2022b.

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.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6672–6681. PMLR, 2020.

Pierre Marion. Generalization bounds for neural ordinary differential equations and deep residual networks. *arXiv preprint arXiv:2305.06648*, 2023.

Pierre Marion, Yu-Han Wu, Michael E Sander, and Gérard Biau. Implicit regularization of deep residual networks towards neural ODEs. *arXiv preprint arXiv:2309.01213*, 2023.

Youssef Marzouk, Zhi Ren, Sven Wang, and Jakob Zech. Distribution learning via neural differential equations: A nonparametric statistical perspective. *arXiv preprint arXiv:2309.01043*, 2023.

Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

Dan Mikulincer and Yair Shenfeld. On the Lipschitz properties of transportation along heat flows. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2020-2022*, pages 269–290. Springer, 2023.

Dan Mikulincer and Yair Shenfeld. The Brownian transport map. *Probability Theory and Related Fields*, 190(1):379–444, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

Youssef Mroueh and Truyen Nguyen. On the convergence of gradient descent in GANs: MMD GAN as a gradient flow. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1720–1728. PMLR, 2021.

Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2976–2985. PMLR, 2019.

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.

Joe Neeman. Lipschitz changes of variables via heat flow. *arXiv preprint arXiv:2201.03403*, 2022.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 25858–25889. PMLR, 2023.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 26517–26582. PMLR, 2023.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. OT-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232, 2021.

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with ReLU networks: Estimators and minimax rates. *The Journal of Machine Learning Research*, 23(1):11251–11292, 2022.

Daniel P. Palomar and Sergio Verdú. Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, 52(1):141–154, 2005.

Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 2023.

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

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T.Q. Chen. Multisample flow matching: straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28100–28127. PMLR, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1530–1538. PMLR, 2015.

Herbert E. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, volume I, page 157–163, Berkeley and Los Angeles, 1956. University of California Press.

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.

Domenec Ruiz-Balet and Enrique Zuazua. Neural ODE control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

Ruslan Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2:361–385, 2015.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45 – 114, 2014.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the DDPM objective. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Neta Shaul, Ricky T.Q. Chen, Maximilian Nickel, Matthew Le, and Yaron Lipman. On kinetic optimal probability paths for generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 30883–30907. PMLR, 2023.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, Joel L Horowitz, and Jian Huang. Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*, 2021.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, Joel L. Horowitz, and Jian Huang. Estimation of non-crossing quantile regression process with deep requ neural networks. *arXiv preprint arXiv:2207.10442. Accepted by JMLR.*, 2022a.

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

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022b.

Jonathan W. Siegel. Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces. *Journal of Machine Learning Research*, 24(357):1–52, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265. PMLR, 2015.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11895–11907. Curran Associates, Inc., 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pages 12438–12448. Curran Associates, Inc., 2020.

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*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 23–29 Jul 2023a.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3D molecule generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

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

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2023.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. In *Advances in Neural Information Processing Systems*, volume 34, pages 3609–3621. Curran Associates, Inc., 2021.

Esteban G. Tabak and Cristina V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Cédric Villani. *Displacement interpolation.*, pages 113–162. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Andre Wibisono and Varun Jog. Convexity of mutual information along the heat flow. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1615–1619. IEEE, 2018a.

Andre Wibisono and Varun Jog. Convexity of mutual information along the Ornstein-Uhlenbeck flow. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pages 55–59. IEEE, 2018b.

Andre Wibisono and Kaylee Yingxi Yang. Convergence in KL divergence of the inexact Langevin algorithm with application to score-based generative models. *arXiv preprint arXiv:2211.01512*, 2022.

Andre Wibisono, Varun Jog, and Po-Ling Loh. Information and estimation in Fokker-Planck channels. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2673–2677. IEEE, 2017.

Yihong Wu and Sergio Verdú. Functional properties of minimum mean-square error and mutual information. *IEEE Transactions on Information Theory*, 58(3):1289–1301, 2011.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *International Conference on Machine Learning*, pages 53291–53327, 2024.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Invertible normalizing flow neural networks by JKO scheme. *arXiv preprint arXiv:2212.14424*, 2022.

Liu Yang and George E. Karniadakis. Potential flow generator with L_2 optimal transport regularity for generative models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):528–538, 2020.

Yahong Yang, Haizhao Yang, and Yang Xiang. Nearly optimal VC-dimension and pseudo-dimension bounds for deep neural network derivatives. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on Learning Theory*, pages 639–649. PMLR, 2018.

Linfeng Zhang, Weinan E, and Lei Wang. Monge-Ampère flow for generative modeling. *arXiv preprint arXiv:1809.10188*, 2018.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Improved techniques for maximum likelihood estimation for diffusion ODEs. *arXiv preprint arXiv:2305.03935*, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.