NRFlow: Towards Noise-Robust Generative Modeling via High-Order Flow Matching

Bo Chen[1] **Chengyue Gong**[2] **Xiaoyu Li**[3] **Yingyu Liang**[4,†] **Zhizhou Sha**[2] **Zhenmei Shi**[4] **Zhao Song**[5,*]
Mingda Wan[1] **Xugang Ye**[6]

[1]Middle Tennessee State University, [2]University of Texas at Austin, [3]University of New South Wales
[4]University of Wisconsin-Madison, [5]University of California, Berkeley, [6]TikTok
[*] magic.linuxkde@gmail.com, [†] yliang@cs.wisc.edu,yingyul@hku.hk

Abstract

Flow-based generative models have shown promise in various machine learning applications, but they often face challenges in handling noise and ensuring robustness in trajectory estimation. In this work, we propose NRFlow, a novel extension to flow-based generative modeling that incorporates second-order dynamics through acceleration fields. We develop a comprehensive theoretical framework to analyze the regularization effects of high-order terms and derive noise robustness guarantees. Our method leverages a two-part loss function to simultaneously train first-order velocity fields and high-order acceleration fields, enhancing both smoothness and stability in learned transport trajectories. These results highlight the potential of high-order flow matching for robust generative modeling in complex and noisy environments.

1 INTRODUCTION

Flow-based generative modeling [Lipman et al., 2023, Liu et al., 2023, Albergo and Vanden-Eijnden, 2023, Bose et al., 2024, Esser et al., 2024] has recently gained substantial traction in machine learning due to its capacity to learn expressive, invertible transformations that map simple source distributions to more complex target distributions. In particular, flow matching techniques [Lipman et al., 2023, Liu et al., 2023] have shown promising results in bridging the gap between traditional normalizing flows and score-based diffusion models. These methods typically construct a continuous time trajectory or "flow" that transports samples from a prior distribution, usually Gaussian distribution, to an unknown data distribution. By matching a parameterized velocity field to the ground-truth time derivatives along a path connecting these distributions, flow matching has demonstrated impressive empirical performance, as well as favorable theoretical properties.

Despite these advances, existing flow-based frameworks remain susceptible to perturbations such as noise contamination in the data or instability in the learned transport path [Wang et al., 2024a, Hu et al., 2024b]. This vulnerability arises because standard (i.e., first-order) methods predominantly focus on velocity alignment, thereby neglecting higher-order dynamics and their influence on smoothness and robustness. Several works in diffusion-based modeling [Chen, 2023, Hang and Gu, 2024, Lin et al., 2024] have suggested that carefully accounting for noise and incorporating additional constraints can lead to more stable solutions. However, a principled and comprehensive approach to integrating higher-order information within flow matching has yet to be fully explored.

In this paper, we propose NRFlow, a novel extension to the traditional flow-based generative framework that leverages acceleration fields in addition to velocity fields. Our approach is motivated by the observation that second-order information can be interpreted as a form of regularization, acting to enforce higher-order smoothing constraints on the learned trajectories. Concretely, we show—both formally and informally—that these second-order terms can mitigate noisy or imperfect training data by providing stronger regularity conditions, which in turn bolster model robustness. The core idea is straightforward but powerful: we decompose the learning objective into two parts, one for velocity matching and one for acceleration matching, and jointly train these terms to ensure smooth, stable flows.

Our main theoretical contributions center on establishing a rigorous noise-robustness guarantee that quantifies how noise in the observed data propagates through the learned second-order flow. Specifically, we derive a two-part loss function whose first-order component learns a velocity field approximating $\dot{x}_t$, while the second-order component targets $\ddot{x}_t$. We then prove that if these losses remain small, the Sobolev H^2-norm of the estimation error is bounded, effectively demonstrating the regularization effect of the

second-order term. Further, a discrete Gronwall-type analysis reveals that noise in the training distribution propagates sublinearly over time, thereby yielding improved stability compared with purely first-order flow matching.

In addition to the theoretical framework, we also propose a second-order inference algorithm that modifies the classic flow integration step by adding an acceleration update. Experimental results on a Gaussian mixture dataset highlight the effectiveness of our approach.

The summary of our contributions to the theoretical understanding of these architectures and their boundaries, showed as follows:

- We develop a unified second-order flow formulation offering a broad perspective on incorporating higher-order dynamics into generative models.
- We establish rigorous guarantees showing that NRFlow exhibits improved immunity to data noise, grounded in both an informal explanation of its regularization properties and a formal theorem bounding noise propagation.

2 RELATED WORK

Flow Matching. Flow Matching (FM) [Lipman et al., 2024] has recently gained prominence in generative modeling, particularly within the framework of Continuous Normalizing Flows (CNFs). FM offers a simulation-free approach to training CNFs by regressing vector fields along fixed conditional probability paths, thereby enhancing scalability and performance in generative tasks Lipman et al. [2023]. Building upon this foundation, Tong et al. [2024] developed Conditional Flow Matching (CFM), a family of simulation-free training objectives for CNFs. CFM facilitates conditional generative modeling and accelerates both training and inference processes. An exciting development in this area is the introduction of Rectified Flow, which refines flow-based methods by incorporating corrective adjustments to the learned vector fields, enabling more robust convergence and improved stability in generative modeling tasks. Rectified Flow not only enhances training efficiency but also synergizes effectively with other flow-matching methods, further extending the utility of FM in diverse applications. A notable variant within CFM, Optimal Transport Conditional Flow Matching (OT-CFM), approximates dynamic optimal transport in a simulation-free manner, leading to more efficient and stable training. Recent advancements in flow matching for generative modeling have introduced several innovative approaches. Haviv et al. [2024] proposed Wasserstein Flow Matching, extending traditional flow matching to families of distributions, enhancing its applicability in fields like computer graphics and genomics. Cao et al. [2025a] incorporates special relativity constraints in flow matching. Moreover, numerous recent works [Xu et al., 2022, Dax et al., 2023, Pooladian et al., 2023, Wang et al., 2024d,b,c, Chen and Lipman, 2024, Klein et al., 2024, Bansal et al., 2025] have significantly inspired and influenced our work.

Diffusion Models. Generative Models have long been a central topic in the field of Deep Learning [Kingma and Welling, 2014, Goodfellow et al., 2020, Liu et al., 2022, Corvi et al., 2023]. Empowered by the recent advances in Vision Transformers [Dosovitskiy et al., 2021, Zhang et al., 2021, Peebles and Xie, 2023, Bao et al., 2023], diffusion models have gained unprecedented success in generative modeling, producing high fidelity visual contents and has applications in a wide range of real-world scenarios, such as image generation [Song et al., 2021, Rombach et al., 2022, Cao et al., 2025c], video generation [Yang et al., 2024, Cao et al., 2025b, Guo et al., 2025a,b], text editing [Kawar et al., 2023, Garibi et al., 2024, Guo et al., 2025c], e-commerce [Wang et al., 2023, Zhao et al., 2024, Liu et al., 2024]. These approaches typically involve a forward process that systematically adds noise to an initial clean image and a corresponding reverse process that learns to remove noise step by step, thereby recovering the underlying data distribution in a probabilistic manner. Early works [Song and Ermon, 2019, Song et al., 2021, Dockhorn et al., 2022] established the theoretical foundations of this denoising strategy, introducing score-matching and continuous-time diffusion frameworks that significantly improved sample quality and diversity. Subsequent research has focused on more efficient training and sampling procedures [Lu et al., 2022, Wu et al., 2023, Shen et al., 2025a,b], aiming to reduce computational overhead and converge faster without sacrificing image fidelity. Other lines of work leverage latent spaces to learn compressed representations, thereby streamlining both training and inference [Rombach et al., 2022, Hu et al., 2024a]. This latent learning approach integrates naturally with modern neural architectures and can be extended to various modalities beyond images, showcasing the versatility of diffusion processes in modeling complex data distributions. In parallel, recent researchers have also explored multi-scale noise scheduling and adaptive step-size strategies to enhance convergence stability and maintain high-resolution detail in generated content in Lovelace et al. [2024], Feng et al. [2024], Rout et al. [2024], Jiang et al. [2025], Luo et al. [2024]. On the other hand, Rout et al. [2023], Hu et al. [2024a], Wen et al. [2024], Hu et al. [2025d] explores the Diffusion Models theoretically, pointing out future directions of Diffusion Models.

Large Language Models. Neural networks built upon the Transformer architecture [Vaswani et al., 2017] have swiftly risen to dominate modern machine learning approaches in natural language processing. Extensive Transformer models, trained on wide-ranging and voluminous datasets while encompassing billions of parameters, are often termed large language models (LLM) or foundation models [Bommasani et al., 2021]. Representative instances include BERT [Devlin

et al., 2019], PaLM [Chowdhery et al., 2023], Llama [Touvron et al., 2023], ChatGPT [OpenAI, 2024], GPT4 [OpenAI, 2023], among others. These LLMs have showcased striking general intelligence abilities [Bubeck et al., 2023] in various downstream tasks. Numerous adaptation methods have been developed to tailor LLMs for specific applications, such as adapters [Hu et al., 2022, Zhang et al., 2024, Gao et al., 2023a, Shi et al., 2023, Hu et al., 2025b, Cao and Song, 2025], calibration schemes [Zhao et al., 2021, Zhou et al., 2023], multitask fine-tuning [Gao et al., 2021, Von Oswald et al., 2023, Xu et al., 2024], prompt optimization [Gao et al., 2021, Lester et al., 2021, Hu et al., 2025c], scratchpad approaches [Nye et al., 2021], instruction tuning [Li and Liang, 2021, Chung et al., 2024, Mishra et al., 2022], symbol tuning [Wei et al., 2023], black-box tuning [Sun et al., 2022], in-context learning [Wei et al., 2022, Wu et al., 2025b,a] and reinforcement learning from human feedback (RLHF) [Ouyang et al., 2022]. Additional lines of research endeavor to boost model efficiency without sacrificing performance across diverse domains, for example, in Liang et al. [2025a], Li et al. [2024a], Chen et al. [2025d, 2024b], Ke et al. [2024]. An emerging focus in LLMs is the inherent theoretical limitation of these models, including the infeasibility of efficient computation under sufficiently large model weight magnitudes [Alman and Song, 2023, 2024a,b, 2025a,b], circuit complexity [Li et al., 2024b, 2025a, Chen et al., 2024a, Li et al., 2025b], the infeasibility to learn some Boolean functions under gradient descent [Chen et al., 2025c, Hu et al., 2025e, Kim and Suzuki, 2025], universal approximation Kratsios et al. [2022], Chen et al. [2025e], Liu et al. [2025], Hu et al. [2025a], and in-context learning [Wu et al., 2025b,a, Chen et al., 2025a].

Second Order Method. More recently, second-order methods have been applied to neural network optimization and used to solve a lot of problems. Martens et al. [2010] introduced Hessian-free optimization, using conjugate gradients to approximately solve the Newton update. Vinyals and Povey [2012] used a Krylov subspace descent method to directly approximate the Newton update. For natural gradient methods, which perform the steepest descent in the space of network outputs rather than parameters, Amari [1998] showed a connection to second-order optimization via the Fisher information matrix. Grosse and Martens [2016] later extended K-FAC to convolutional neural networks. Ba et al. [2022] combined natural gradient with trust region methods to further improve stability and performance. Despite these advances, second-order neural network optimization remains an active area of research, such as Song [2019], Deng et al. [2022], Song et al. [2023], Gao et al. [2025a,b, 2023c], Bian et al. [2023], Deng et al. [2023], Gao et al. [2023b], Shrivastava et al. [2023], Qin et al. [2023], Chen et al. [2025b,a], Liang et al. [2024], Chen et al. [2024a], Liang et al. [2025b,c], Ke et al. [2025]. Open problems include improving the scalability of Hessian approximations, handling non-convex optimization landscapes, and automating hyper-parameter selection.

Roadmap. In Section 3, we introduce essential computational techniques and key definitions of flow matching and our NRFlow. In Section 4, we provide a detailed regularity analysis and compute the upper bound of the excess risk. We also provide an inequality that quantifies the growth of estimation error under bounded noise. In Section 5, We design some preliminary experiments to demonstrate the validity of our theory and the results. We conclude in Section 6.

3 PRELIMINARY

In this section, we provide the foundational concepts, notations, and assumptions required for the subsequent theoretical developments. In Section 3.1, we begin by listing the key notations employed throughout this work. In this Section 3.2, we state the principal assumptions under which our analysis is conducted. Next, we introduce the flow-matching framework, along with its second-order extension, and highlight several important definitions in Section 3.3. Finally, in Section 3.4, we provide our second-order algorithms.

3.1 NOTATIONS

We use $\Pr[]$ to denote the probability. We use $\mathbb{E}[]$ to denote the expectation. We use $\mathrm{Var}[]$ to denote the variance. We use $\|x\|_p$ to denote the ℓ_p norm of a vector $x \in \mathbb{R}^n$, i.e. $\|x\|_1 := \sum_{i=1}^n |x_i|$, $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$, and $\|x\|_\infty := \max_{i\in[n]} |x_i|$. For variables a, b, We write $a \lesssim b$ to indicate that a is bounded above by b up to a multiplicative constant independent of the main parameters. We write $a \gtrsim b$ to indicate that a is bounded below by b up to a multiplicative constant independent of the main parameters. We denote $\dot{x}^{(k)}$ as the k-th order derivative field of x. We use $\|\cdot\|_{H^2(\Omega)}$ to denote the Sobolev norm in $W^{2,2}(\Omega)$, corresponding to $k = 2$ and $p = 2$.

3.2 ASSUMPTIONS

We now outline the principal assumptions that underlie our theoretical analysis. These assumptions concern smoothness, Lipschitz continuity, bounded noise, function-class complexity, and time discretization. First, we show the assumption of smoothness and boundness.

Assumption 3.1 (Smoothness and boundedness)**.** *We assume the true trajectory x_t^{true} and its first and second derivatives are sufficiently smooth and bounded. Specifically, $x_t^{\mathrm{true}} \in H^2(\Omega)$, and there exist constants $M_1, M_2 > 0$ such that*

$$\|\dot{x}_t^{\mathrm{true}}\|_{H^2(\Omega)} \le M_1, \quad \|\ddot{x}_t^{\mathrm{true}}\|_{H^2(\Omega)} \le M_2.$$

Assumption 3.2 (Lipschitz continuity)**.** *The learned fields $u_{1,\theta_1}(x,t)$ and $u_{2,\theta_2}(v,x,t)$ are L-Lipschitz continuous in*

spatial and temporal arguments. Formally, there exists $L > 0$ such that for all $x, y \in \mathbb{R}^d$ and $t, s \in [0, 1]$:

$$\|u_{1,\theta_1}(x,t) - u_{1,\theta_1}(y,t)\|_2 \le L\|x-y\|_2,$$
$$\|u_{2,\theta_2}(v,x,t) - u_{2,\theta_2}(v,y,t)\|_2 \le L\|x-y\|_2,$$

and

$$\|u_{2,\theta_2}(v,x,t) - u_{2,\theta_2}(v,y,t)\|_2 \le L\|x-y\|_2,$$
$$\|u_{2,\theta_2}(v,x,t) - u_{2,\theta_2}(v,x,s)\|_2 \le L\|t-s\|_2,$$

Similar conditions hold for time differences.

Assumption 3.3 (Bounded noise magnitude)**.** *There exists $\delta > 0$ such that $\|\eta_i\|_2 \le \delta$ or $\mathbb{E}[\|\eta_i\|_2^2] \le \delta^2$. This ensures that the noise does not grow without bounds.*

Assumption 3.4 (Rademacher complexity or VC dimension)**.** *There exist function classes $\mathcal{F}_1, \mathcal{F}_2$ such that*

$$u_{1,\theta_1}(\cdot,\cdot) \in \mathcal{F}_1, \quad u_{2,\theta_2}(\cdot,\cdot) \in \mathcal{F}_2.$$

The complexity of each class is measured by $\mathcal{C}(\mathcal{F}_1)$ and $\mathcal{C}(\mathcal{F}_2)$.

Assumption 3.5 (Bounded loss)**.** *There exists some constant $Q > 0$ such that for all $\theta \in \Theta$ and all $x \in \mathcal{X}$, the per-sample loss $l_\theta(x)$ satisfies $|l_\theta(x)| \le Q$.*

Assumption 3.6 (Time discretization)**.** *For the inference (deployment) stage, let $\Delta t = 1/L$ be the uniform step size, and define discrete times $t_l = l\Delta t$ for $l = 0, 1, \cdots, L$. The numerical scheme for forward integration is*

$$x_{l+1} = x_l + \Delta t \cdot u_{1,\theta_1}(x_l, t_l) + \frac{(\Delta t)^2}{2} u_{2,\theta_2}(u_{1,\theta_1}(x_l,t_l), x_l, t_l).$$

Remark 3.7. *In this paper, our derivation accounts for higher-order residual terms through discrete Gronwall-type analyses (see Lemma 4.6). Despite the discretization assumption potentially omitting explicit higher-order terms, we bound such cumulative effects over time by considering Lipschitz continuity and noise assumptions. This ensures that any leftover remainder does not cause unbounded error growth.*

3.3 FLOW MATCHING AND RECTIFIED FLOW

Next, we describe the general framework of flow matching and its second-order rectification. These concepts form the basis for our proposed method, as they integrate first and second-order information for trajectory estimation.

Definition 3.8 (Easy error)**.** *Let*

$$c_1(t) := \dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}, \quad c_2(t) := \ddot{x}_t^{\text{est}} - \ddot{x}_t^{\text{true}}.$$

Fact 3.9. *Let a field x_t be defined as*

$$x_t = \alpha_t x_0 + \beta_t x_1,$$

where α_t and β_t are functions of t, and x_0, x_1 are constants. Then, the first-order gradient $\dot{x}_t$ and the second-order gradient $\ddot{x}_t$ can be manually calculated as

$$\dot{x}_t = \dot{\alpha}_t x_0 + \dot{\beta}_t x_1 \text{ and } \ddot{x}_t = \ddot{\alpha}_t x_0 + \ddot{\beta}_t x_1.$$

Definition 3.10 (A variant of flow matching in Lipman et al. [2023])**.** *Given two distributions μ_0 and π_0 on $\mathbb{R}^d$, flow matching aims to learn a time-dependent velocity field*

$$v_\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$$

such that for any trajectory x_t transporting $x_0 \sim \mu_0$ to $x_1 \sim \pi_0$, we have

$$\dot{x}_t \sim v_\theta(x_t, t).$$

Remark 3.11. *In practice, one often samples (x_0, x_1) from (μ_0, π_0) and parameterizes x_t (e.g. via interpolation) at intermediate times to build a training objective that matches the velocity field to the true time derivative $\dot{x}_t$.*

Definition 3.12 (Second-order flow matching and loss)**.** *We additionally learn an acceleration field*

$$u_{2,\theta_2}(v,x,t), \quad \text{where} \quad v = u_{1,\theta_1}(x,t),$$

to approximate $\ddot{x}_t$. Hence, the two-part (second-order) loss is:

$$L_{2\text{nd}}(\theta_1,\theta_2) = \underbrace{\mathbb{E}[\|\dot{x}_t^{\text{true}} - u_{1,\theta_1}(x_t,t)\|_2^2]}_{L_{2,1,\theta_1}} + \underbrace{\mathbb{E}[\|\ddot{x}_t^{\text{true}} - u_{2,\theta_2}(u_{1,\theta_1}(x_t,t),x_t,t)\|_2^2]}_{L_{2,2,\theta_2,\theta_1}}.$$

Here, $\dot{x}_t^{\text{true}}$ and $\ddot{x}_t^{\text{true}}$ are observed (or numerically approximated) true velocity and acceleration, while u_{1,θ_1} and u_{2,θ_2} are the networks to be trained.

Definition 3.13 (Trajectory and time parameterization)**.** *Consider a continuous trajectory $\{x_t\}_{t\in[0,1]} \in \mathbb{R}^d$ connecting an initial distribution μ_0 to a target data distribution π_0. We assume $x_0 \sim \mu_0$ and $x_1 \sim \pi_0$.*

Definition 3.14 (First and second order flow)**.** *A first-order rectified flow is characterized by a velocity field $u_{1,\theta_1}(x,t)$ approximating $\dot{x}_t$. A second-order rectified flow further involves an acceleration field $u_{2,\theta_2}(v,x,t)$, where $v = u_{1,\theta_1}(x,t)$ approximates $\dot{x}_t$, and u_{2,θ_2} approximates $\ddot{x}_t$.*

Definition 3.15 (Sobolev space)**.** *For a domain $\Omega \subset \mathbb{R}^d$, the Sobolev space $H^2(\Omega)$ is defined as*

$$H^2(\Omega) = \{f \in L^2(\Omega) : D^\alpha f \in L^2(\Omega), \forall |\alpha| \le 2\}.$$

We assume the trajectory $x_t(\omega)$ or its corresponding fields lie in such spaces, ensuring sufficient smoothness.

Definition 3.16 (Noisy data and noise proportion)**.** *Let the training dataset $X = \{x_i\}_{i=1}^N$ be drawn from π_0 but corrupted by noise. We denote the noise proportion as $\epsilon = N_{\text{noisy}}/N$. A noisy sample can be modeled as*

$$x_i^{\text{noisy}} = x_i^{\text{clean}} + \eta_i,$$

where η_i satisfies certain boundedness conditions.

Definition 3.17 (Error)**.** *As we assume in Assumption 3.6, we define the error in H^2-norm.*

$$e_k := \|x_k^{\text{est}} - x_k^{\text{true}}\|_{H^2(\Omega)}$$

Definition 3.18 (Second-order loss function)**.** *The loss function for the second-order method contains two parts. We define the first part which is trying to using $\dot{x}_t$ in Fact 3.9, x_t and t to learn function $u_{1,t}$, thus the loss is*

$$L_{2,1,\theta_1} := \|\dot{x}_t - u_{1,\theta_1}(x_t, t)\|_2^2$$

Next, we define the second part, which is trying to use $\ddot{x}_t, u_{1,\theta_1}(x_t, t)$, x_t and t to learn u_{2,θ_2} function, thus the loss is

$$L_{2,2,\theta_2,\theta_1} := \|\ddot{x}_t - u_{2,\theta_2}(u_{1,\theta_1}(x_t, t), x_t, t)\|_2^2$$

Overall, the total loss is

$$L_{2,\theta} := L_{2,1,\theta_1} + L_{2,2,\theta_2,\theta_1}$$

Definition 3.19 (Empirical loss)**.** *We define the empirical second-order loss as $\widetilde{L}_{2,\theta} = \frac{1}{N}\sum_{i=1}^N l_\theta(x_i)$.*

Definition 3.20 (Population loss)**.** *We define the population second-order loss as $L_{2,\theta} = \mathbb{E}[l_\theta(X)]$*

3.4 PROPOSED METHOD

In this section, we now summarize the second-order algorithms that arise from the definitions above. Due to the space limitation, we delay the original first-order algorithm and our new third algorithms in the appendix.

Algorithm 1 Our new second-order training process

procedure 2NDORDERFORWARD()
  for each iteration **do**
    Random sample x_0 and time t, with target x_1
    $x_t \leftarrow \alpha_t \cdot x_0 + \sqrt{1-\alpha_t^2} \cdot x_1$
    Compute gradient with respect to $L_{2,\theta}$ ▷ see Def. 3.18
  end for
  return u_1, u_2 ▷ Two network functions
end procedure

Algorithm 2 Our new second-order inference algorithm

procedure 2NDORDERINFERENCE(u_1, u_2)
  $x_0 \sim \mathcal{N}(0, 1)$
  Initial $x \leftarrow x_0$
  for t from 0 to 1 with step $\Delta t = 0.01$ **do**
    $x \leftarrow x + \Delta t \cdot u_1(x, t) + \frac{(\Delta t)^2}{2} \cdot u_2(u_1(x, t), x, t)$
  end for
  return x
end procedure

Remark 3.21. *Since our approach could use a separate neural network for each higher-order term, the overall complexity scales by a constant factor corresponding to the number of higher-order terms. Thus, the time complexity of our method exactly matches the complexity of previous first-order flow-matching methods, and this constant factor will not substantially increase computational overhead beyond the original first-order algorithm.*

Remark 3.22. *Our model parameterizes all orders of time derivatives, rather than only parameterizing higher-order derivatives (e.g., acceleration) and using numerical integration to compute lower-order derivatives (e.g., velocity). This technical choice ensures the numerical stability of our model and avoids introducing extra numerical errors during integration. In computationally resource-scarce settings where fewer model parameters are needed, our model can also flexibly consider parameterizing only higher-order time derivatives while deriving lower-order derivatives through numerical integration.*

4 OUR RESULT

In Section 4.1, we first present a classical elliptic regularity lemma. In Section 4.2, we then show how controlling the second-order loss ensures bounded estimation error in a stronger Sobolev norm, thereby revealing a key regularization effect. In Section 4.3, we derive an excess risk bound, demonstrating that our method generalizes well under finite-sample conditions. In Section 4.4, we further analyze a discrete propagation inequality under noise. In Section 4.5, we combine these insights in our main theorem, proving that the learned trajectory remains robust against noise and sampling limitations.

4.1 ELLIPTIC REGULARITY

In this section, we introduce the first result, which is a classical result that characterizes the relationship between different Sobolev norms for sufficiently smooth functions.

Lemma 4.1 (Elliptic regularity in Evans [2010])**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a sufficiently smooth boundary. Suppose $h : \Omega \to \mathbb{R}$ belongs to $L^2(\Omega)$, has weak*

derivatives up to second order in $L^2(\Omega)$ and satisfies appropriate boundary conditions. Then, there exists a constant $C_{\text{reg}} > 0$, depending only on Ω and the boundary conditions, such that

$$\|h\|_{H^2(\Omega)} \leq C_{\text{reg}}(\|\nabla h\|_{L^2(\Omega)} + \|h\|_{L^2(\Omega)}).$$

The above result is fundamental in establishing norm equivalences in Sobolev spaces, which we will use to analyze the regularity of error terms in subsequent lemmas.

4.2 REGULARIZATION EFFECT

In this section, we now connect the second-order loss function with the Sobolev norm of the estimation error.

Lemma 4.2 (Regulazation effect)**.** *Let $\{(x_0, x_1, t)\}$ denote the sampling start point, endpoint, and time in the training set, and suppose the true trajectory $\ddot{x}_t \in H^2(\Omega)$, we consider*

$$L_{2,2,\theta_1,\theta_2} = \mathbb{E}[\|\ddot{x}_t^{\text{true}} - u_{2,\theta_2}(u_{1,\theta_1}(x_t^{\text{true}}, t), x_t^{\text{true}}, t)\|_2^2],$$

The second-order loss with respect to the true second derivative. Particularly, there exist $C_{\text{reg}} \in \mathbb{R}$ when $L_{2,2,\theta_1,\theta_2}$ is sufficiently small such that

$$\begin{aligned}&\|\dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}\|_{H^2(\Omega)}\\ &\leq C_{\text{reg}}(L_{2,2,\theta_1,\theta_2}^{1/2} + \|\dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}\|_{L^2(\Omega)}).\end{aligned}$$

Proof. First we let $h(\cdot) = \dot{x}_t^{\text{est}}(\cdot) - \dot{x}_t^{\text{true}}(\cdot)$, the problem depends on both t and x, we could it by $h(t, x)$. For clarity, we simply write $h(\cdot)$ and regard it as a function on Ω. Generally, one assumes $\dot{x}_t^{\text{est}}, \dot{x}_t^{\text{true}} \in H^2(\Omega)$ so that $h \in H^2(\Omega)$.

Applying Lemma 4.1 to $h(\cdot) = \dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}$, we have

$$\begin{aligned}&\|\dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}\|_{H^2(\Omega)}\\ &\leq C_{\text{reg}}(\|\nabla h\|_{L^2(\Omega)} + \|h\|_{L^2(\Omega)}).\end{aligned} \tag{1}$$

By the Definition of the loss function, a small $L_{2,2,\theta_1,\theta_2}$ implies

$$\|\dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}\|_{L^2(\Omega)} \lesssim L_{2,2,\theta_1,\theta_2}^{1/2} \tag{2}$$

Combining Eq.(1) and (2), we have

$$\begin{aligned}&\|\dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}\|_{H^2(\Omega)}\\ &\leq C_{\text{reg}}(L_{2,2,\theta_1,\theta_2}^{1/2} + \|\dot{x}_t^{\text{est}} - \dot{x}_t^{\text{true}}\|_{L^2(\Omega)}).\end{aligned}$$

Thus, we complete the proof. □

The above lemma highlights the importance of small second-order loss: it guarantees that the estimation error in the stronger Sobolev norm $H^2(\Omega)$ remains controlled.

4.3 EXCESS RISK

In this section, we introduce the following result which bounds the difference between the empirical and population loss, demonstrating that our method generalizes well under finite-sample conditions.

Lemma 4.3 (Symmetrization bound)**.** *Let $\{x_i\}_{i=1}^N$ and $\{x_i'\}_{i=1}^N$ be i.i.d. samples. For $\mathcal{G} = \{\ell_\theta : \theta \in \Theta\}$, we have:*

$$\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^N (g(x_i) - g(x_i'))| \leq \frac{2}{N}\mathop{\mathbb{E}}_{\sigma}[\sup_{g\in\mathcal{G}} \sum_{i=1}^N \sigma_i g(x_i)],$$

where $\{\sigma_i\}_{i=1}^N$ are Rademacher random variables, $\sigma_i \in \{+1, -1\}$ with equal probability.

Proof. For each σ_i has a symmetric distribution, we have:

$$|\sum_{i=1}^N (g(x_i) - g(x_i'))| \leq \mathop{\mathbb{E}}_{\sigma}[|\sum_{i=1}^N \sigma_i(g(x_i) - g(x_i'))|$$

Taking the supremum over $g \in \mathcal{G}$ and noting that $\{x_i\}$ and $\{x_i'\}$ have the same distribution, we can split the expression inside the absolute value:

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\sum_{i=1}^N (g(x_i) - g(x_i'))|\\ &\leq \mathbb{E}_\sigma[\sup_{g\in\mathcal{G}} |\sum_{i=1}^N \sigma_i(g(x_i) - g(x_i'))|].\end{aligned}$$

By the triangle inequality, we get:

$$|\sum_{i=1}^N \sigma_i(g(x_i) - g(x_i'))| \leq |\sum_{i=1}^N \sigma_i g(x_i)| + |\sum_{i=1}^N \sigma_i g(x_i')|.$$

Hence,

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\sum_{i=1}^N (g(x_i) - g(x_i'))|\\ &\leq \mathbb{E}_\sigma[\sup_{g\in\mathcal{G}} |\sum_{i=1}^N \sigma_i\, g(x_i)| + \sup_{g\in\mathcal{G}} |\sum_{i=1}^N \sigma_i g(x_i')|].\end{aligned}$$

Because $\{x_i'\}$ is drawn from the same distribution as $\{x_i\}$, the two supremum terms have the same expected value. Therefore, we can combine them as follows:

$$\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^N (g(x_i) - g(x_i'))| \leq \frac{2}{N}\mathop{\mathbb{E}}_{\sigma}[\sup_{g\in\mathcal{G}} \sum_{i=1}^N \sigma_i g(x_i)],$$

Thu,s we complete the proof. □

Lemma 4.4 (Theorem 6.11 in Shalev-Shwartz and Ben–David [2014])**.** *As we defined in Definition 3.15, 3.20 and*

3.19, if Assumption 3.4 holds, for $g \in \mathcal{G}$ where $\mathcal{G} = \{\ell_\theta : \theta \in \Theta\}$, we have

$$\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N} g(x_i') - \mathbb{E}[g(x)]| \leq O(\sqrt{\ln(1/\beta)/N})$$

Lemma 4.5 (Excess risk)**.** *As we defined in Definition 3.20, we have*

$$\begin{aligned}&\widetilde{L}_{2,\theta}\\ &= \frac{1}{N}\sum_{i=1}^{N}(\|\dot{x}_t^{\mathrm{true},i} - u_{1,\theta_1}(\cdot)\|_2^2 + \|\ddot{x}_t^{\mathrm{true},i} - u_{2,\theta_2}(\cdot)\|_2^2)\end{aligned} \tag{3}$$

and

$$L_{2,\theta} = \mathbb{E}[\|\dot{x}_t^{\mathrm{true}} - u_{1,\theta_1}(\cdot)\|_2^2 + \|\ddot{x}_t^{\mathrm{true}} - u_{2,\theta_2}(\cdot)\|_2^2] \tag{4}$$

Suppose $\mathcal{F}_1$ nad $\mathcal{F}_2$ have finite or at most polynomially growing complexities $\mathcal{C}(\mathcal{F}_1), \mathcal{C}(\mathcal{F}_2)$. Then for $\beta \in (0,1)$, with probability at least $1-\beta$, we have

$$|\widetilde{L}_{2,\theta} - L_{2,\theta}| \leq O((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \ln(1/\beta))/N)^{1/2}.$$

Proof. Let $\mathcal{G} = \{\ell_\theta : \theta \in \Theta\}$ represent the complexity of $\mathcal{G}$ the Rademacher/VC dimension, As we defined in Definition 3.12, 3.19 and 3.20, we calculate the empirical loss and population loss,

$$\begin{aligned}\widetilde{L}_{2,\theta} &= \frac{1}{N}\sum_{i=1}^{N} l_\theta(x_i)\\ &= \frac{1}{N}\sum_{i=1}^{N}(\|\dot{x}_t^{\mathrm{true},i} - u_{1,\theta_1}(\cdot)\|_2^2\\ &\quad + \|\ddot{x}_t^{\mathrm{true},i} - u_{2,\theta_2}(\cdot)\|_2^2)\end{aligned}$$

and

$$\begin{aligned}L_{2,\theta} &= \mathbb{E}[l_\theta(X)]\\ &= \mathbb{E}[\|\dot{x}_t^{\mathrm{true}} - u_{1,\theta_1}(\cdot)\|_2^2 + \|\ddot{x}_t^{\mathrm{true}} - u_{2,\theta_2}(\cdot)\|_2^2]\end{aligned}$$

let $\{x_i'\}_{i=1}^N$ be an i.i.d. sample from the same distribution as $\{x_i\}_{i=1}^N$, and let $\{\sigma_i\}_{i=1}^N$ be i.i.d. Rademacher random variables ($\sigma_i \in \{+1,-1\}$ with probability $1/2$ each). Then, for any $g \in \mathcal{G}$, we have

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N} g(x_i) - \mathbb{E}[g(x)]|\\ \leq &\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N}(g(x_i)) - g(x_i')|\\ &+ \sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N} g(x_i') - \mathbb{E}[g(x)]|\end{aligned} \tag{5}$$

We can upper bound the first term in Eq. (5),

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N}(g(x_i)) - g(x_i')|\\ &\leq \frac{2}{N}\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}\sum_{i=1}^{N}\sigma_i g(x_i)]\\ &= 2\widetilde{\mathcal{R}}_N(\mathcal{G})\\ &\leq 2\cdot O(\sqrt{C(\mathcal{G})/N})\\ &\leq O(\sqrt{(C(\mathcal{F}_1)+C(\mathcal{F}_2))/N})\end{aligned} \tag{6}$$

where the first step follows from Lemma 4.3, the second step comes from we define $\widetilde{\mathcal{R}}_N(G) := \mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i g(x_i)]$, the third step follows from Assumption 3.4, the forth step follows from the definition of $\mathcal{G}$.

We can upper bound the second term in Eq. (5) by using Lemma 4.4,

$$\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N} g(x_i') - \mathbb{E}[g(x)]| \leq O(\sqrt{\ln(1/\beta)/N}) \tag{7}$$

Loading Eq. (6) and Eq. (7), we can obtain

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{i=1}^{N} g(x_i) - \mathbb{E}[g(x)]|\\ &\leq O(\sqrt{(C(\mathcal{F}_1)+C(\mathcal{F}_2)+\ln(1/\beta))/N})\end{aligned}$$

Thus, we complete the proof. □

4.4 DISCRETE PROPAGATION UNDER NOISE

In this section, we show the lemma about discrete propagation under noise, which quantifies how noise in the trajectory affects the error propagation in a discrete setting.

Lemma 4.6 (Discrete propagation under noise)**.** *Suppose η_i satisfies $\|\eta_i\| \leq \delta$, there exist $C_{\mathrm{prop}} \in \mathbb{R}$ such that*

$$e_{l+1} \leq (1 + \Delta t \cdot C_{\mathrm{prop}})e_l + C_{\mathrm{prop}} \cdot \Delta t \cdot \delta \cdot \epsilon$$

unrolling for $l = 0, \cdots, L-1$, we have

$$e_L \leq e_0 \exp(C_{\mathrm{prop}}) + \frac{\delta\cdot\epsilon}{C_{\mathrm{prop}}}(\exp(C_{\mathrm{prop}}) - 1)$$

Proof. By Assumptions 3.2 and 3.6, the discrete updates for both the estimated and true systems can be written as:

$$\begin{aligned}x_{l+1} &= x_l + \Delta t \cdot u_{1,\theta_1}(x_l, t_l)\\ &\quad + \frac{(\Delta t)^2}{2}u_{2,\theta_2}(u_{1,\theta_1}(x_l,t_l), x_l, t_l).\end{aligned}$$

Subtracting the true system update from the estimated one gives:

$$\begin{aligned} & x_{l+1}^{\text{est}} - x_{l+1}^{\text{true}} \\ &= (x_l^{\text{est}} - x_l^{\text{true}}) + \Delta t \cdot (u_{1,\theta_1}(x_l^{\text{est}}, t_l) - \dot{x}_l^{\text{true}}) \\ &\quad + \frac{(\Delta t)^2}{2} \cdot (u_{2,\theta_2}(\cdot) - \ddot{x}^{\text{true}}). \end{aligned}$$

Taking the H^2-norm, we have:

$$\begin{aligned} & \|x_{l+1}^{\text{est}} - x_{l+1}^{\text{true}}\|_{H^2(\Omega)} \\ &\leq \|x_l^{\text{est}} - x_l^{\text{true}}\|_{H^2(\Omega)} \\ &\quad + \Delta t \cdot \|u_{1,\theta_1}(x_l^{\text{est}}, t_l) - \dot{x}_l^{\text{true}}\|_{H^2(\Omega)} \\ &\quad + O((\Delta t)^2). \end{aligned}$$

Since $\dot{x}_l^{\text{true}} \sim u_{1,\theta_1}(x_l^{\text{true}}, t_l)$ and the deviation is controlled by $\|x_l^{\text{est}} - x_l^{\text{true}}\|$ and noise $\delta\epsilon$, we can write:

$$\begin{aligned} e_{l+1} &= \|x_{l+1}^{\text{est}} - x_{l+1}^{\text{true}}\|_{H^2(\Omega)} \\ &\leq (1 + \Delta t \cdot C_{\text{prop}}) e_l + C_{\text{prop}} \Delta t \cdot \delta \cdot \epsilon, \end{aligned}$$

where C_{prop} depends on the Lipschitz constants of u_{1,θ_1} and u_{2,θ_2}. Repeatedly applying this inequality from $l = 0$ to $l = L - 1$, we have:

$$\begin{aligned} e_L = e_0 & \prod_{j=0}^{L-1} (1 + \Delta t \cdot C_{\text{prop}}) \\ & + \sum_{l=0}^{L-1} (C_{\text{prop}} \Delta t \cdot \delta \cdot \epsilon \prod_{j=l+1}^{L-1} (1 + \Delta t \cdot C_{\text{prop}})). \end{aligned}$$

Recognizing that: $\prod_{j=0}^{L-1}(1 + \Delta t \cdot C_{\text{prop}}) = \exp(C_{\text{prop}})$, we simplify the summation term using the geometric series formula:

$$\sum_{l=0}^{L-1} \prod_{j=l+1}^{L-1} (1 + \Delta t \cdot C_{\text{prop}}) = \frac{\exp(C_{\text{prop}}) - 1}{C_{\text{prop}}}.$$

Thus, we obtain:

$$e_L \leq e_0 \exp(C_{\text{prop}}) + \frac{\delta \cdot \epsilon}{C_{\text{prop}}} (\exp(C_{\text{prop}}) - 1).$$

This completes the proof. □

This result provides a discrete Gronwall-type inequality, quantifying the growth of error under bounded noise.

4.5 MAIN RESULT

In this section, we now state and prove our main result with the auxiliary lemmas in place, which establishes the robustness of the learned trajectory against noise and finite-sample effects.

Theorem 4.7 (Noise robustness)**.** [1] *Suppose all Assumption 3.1, 3.2, 3.3 and 3.6 holds, Let $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2)$ is the approximately optimal solution, then $\beta \in (0, 0.1)$, the final-time($t = 1$) trajectory estimate satisfies*

$$\begin{aligned} & \|x_{t=1}^{\text{est}} - x_{t=1}^{\text{true}}\|_{H^2(\Omega)} \\ &\leq C_1 \exp(C_2) \cdot (e_0 + \delta \cdot \epsilon) \\ &\quad + C_3 \cdot ((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \ln(1/\beta))/N)^{1/2}, \end{aligned}$$

where $e_0 = \|x_0^{\text{est}} - x_0^{\text{true}}\|$ is the initial error. C_1, C_2, C_3 depends on Lipschitz constant L, dimension d, sobolev embedding constant, Δt, $\exp(C_2)$ represents the discrete gronwall factor for the time interval $[0, 1]$.

Table 1: **Euclidean distance loss across three complex distribution datasets under the new trajectory setting.** Lower values indicate higher accuracy in distribution transfer. The optimal values are highlighted in **Bold**, and the second-best loss values (second-lowest) are represented by Underlined numbers for each dataset (row).

Loss terms	Five mode	3-round Spiral	Dot Hyperbola
O1 Liu et al. [2023]	1.755	17.338	18.096
O1 + O2 (Ours)	<u>0.956</u>	<u>15.514</u>	<u>3.823</u>
O1 + O2 + O3 (Ours)	**0.778**	**11.866**	**2.959**

5 EXPERIMENTS

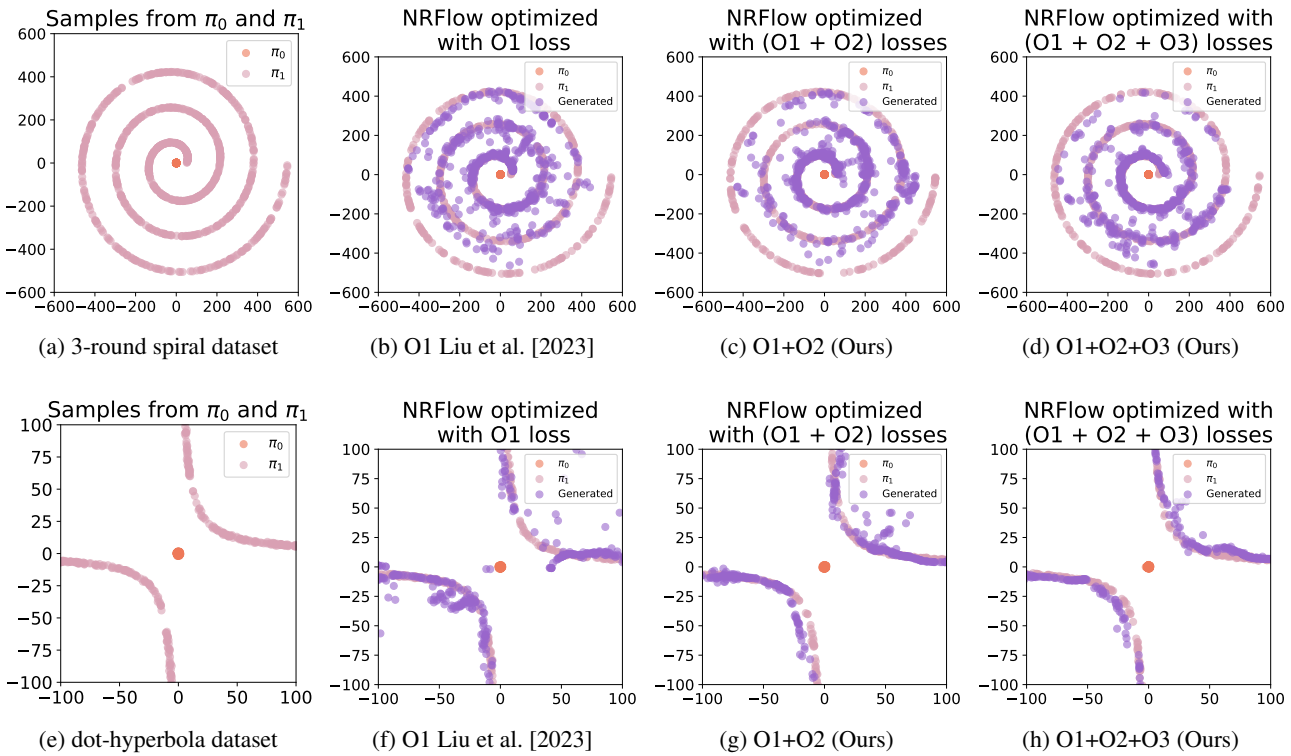
This section presents a series of experiments to evaluate the effectiveness of our NRFlow. Our results demonstrate that NRFlow significantly improves distribution generation, with the high-order loss playing a key role in enhancing model performance. In Section 5.1, we provide a detailed explanation for the setup of our experiments. In Section 5.2, we provide a comprehensive analysis of the effectiveness of our high-order supervision in our NRFlow model.

5.1 EXPERIMENT SETUP

We conduct comprehensive evaluations of NRFlow across diverse data distributions and multiple loss function combinations. Notably, the NRFlow implementation using only first-order loss (denoted as O1) corresponds exactly to the baseline Rectified Flow framework Liu et al. [2023]. Our proposed extensions consist of two configurations: second-order enhanced (O1+O2) and third-order augmented (O1+O2+O3) variants.

The evaluation employs two challenging synthetic datasets: a three-round spiral distribution and a dot-hyperbola distribution. Each dataset contains 100 sample points drawn

[1] We state the proof of Theorem 4.7 in Section B in our Appendix.



(a) 3-round spiral dataset (b) O1 Liu et al. [2023] (c) O1+O2 (Ours) (d) O1+O2+O3 (Ours)

(e) dot-hyperbola dataset (f) O1 Liu et al. [2023] (g) O1+O2 (Ours) (h) O1+O2+O3 (Ours)

Figure 1: **NRFlow on 3-round spiral dataset and dot-hyperbola dataset.** From left to right: the first column shows the 3-round spiral dataset and the dot-hyperbola dataset; the second column shows the results of NRFlow optimized with first-order loss (O1) Liu et al. [2023]; the third column shows the results of NRFlow optimized by first-order and second-order loss (O1+O2) (Ours); the fourth column are the results of NRFlow optimized by first-order, second-order and third-order loss (O1+O2+O3) (Ours). Our high-order NRFlows (third column and fourth column) show great capability in modeling complex distribution. Quantitative results are shown in Table 1.

from both source and target distributions. Our implementation utilizes a 2-layer fully connected network with 100 hidden units, trained using the Adam optimizer with a learning rate of 0.005. For the three-round spiral dataset, we employ full-batch training (batch size 1000) over 1000 optimization steps. The dot-hyperbola configuration uses an increased batch size of 1600 to account for its greater geometric complexity. Numerical integration is performed using an adaptive ODE solver throughout all experiments.

5.2 RESULTS ANALYSIS

Our primary objective involves learning optimal transport trajectories between source distributions (depicted in orange in Figure 1) and target distributions (shown in pick). Empirical results demonstrate that the baseline Rectified Flow model (O1) exhibits significant limitations in target distribution modeling. As visualized in Figure 1 (second column), the first-order model generates substantial out-of-distribution artifacts for both synthetic datasets. This observation is quantitatively confirmed in Table 1, where O1 has the highest Euclidean distance metrics among all settings.

The introduction of second-order regularization (O1+O2) yields marked improvements. Final optimization with third-order constraints (O1+O2+O3) produces the most accurate distribution alignment, achieving near-perfect coverage of target domains. These results conclusively demonstrate that high-order supervision progressively enhances the model's ability to capture complex distributional geometries, with each additional regularization term contributing to statistically significant performance gains.

6 CONCLUSION

In summary, we introduced NRFlow, which augments traditional flow-based generative models by the second-order term. Our theoretical results demonstrate that these higher-order terms act as an effective regularizer, providing improved noise robustness and smoother trajectories under bounded perturbations. A discrete Gronwall analysis further shows that error propagation remains controlled, reinforcing the framework's stability. These findings highlight the promise of second-order methods for robust generative modeling.

Acknowledgements

The authors would like to thank the anonymous reviewer of UAI 2025 for their highly insightful suggestions.

References

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

Josh Alman and Zhao Song. Fast attention requires bounded entries. In *NeurIPS*, 2023.

Josh Alman and Zhao Song. The fine-grained complexity of gradient computation for training large language models. In *NeurIPS*, 2024a.

Josh Alman and Zhao Song. How to capture higher-order correlations? generalizing matrix softmax attention to kronecker computation. In *ICLR*, 2024b.

Josh Alman and Zhao Song. Fast rope attention: Combining the polynomial method and fast fourier transform. *arXiv preprint arXiv:2505.11892*, 2025a.

Josh Alman and Zhao Song. Only large weights (and not skip connections) can prevent the perils of rank collapse. *arXiv preprint arXiv:2505.16284*, 2025b.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 1998.

Jimmy Ba, Roger Grosse, and James Martens. Distributed second-order optimization using Kronecker-factored approximations. In *ICLR*, 2022.

Vansh Bansal, Saptarshi Roy, Purnamrita Sarkar, and Alessandro Rinaldo. On the Wasserstein convergence and straightness of rectified flow. *arXiv preprint arXiv:2410.14949*, 2025.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *CVPR*, 2023.

Song Bian, Zhao Song, and Junze Yin. Federated empirical risk minimization via second-order method. *arXiv preprint arXiv:2305.17482*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *ICLR*, 2024.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Yang Cao and Zhao Song. Sorsa: Singular values and orthonormal regularized singular vectors adaptation of large language models. *arXiv preprint arXiv:2409.00055*, 2025.

Yang Cao, Bo Chen, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Mingda Wan. Force matching with relativistic constraints: A physics-inspired approach to stable and efficient generative modeling. *arXiv preprint arXiv:2502.08150*, 2025a.

Yang Cao, Zhao Song, and Chiwun Yang. Video latent flow matching: Optimal polynomial projections for video interpolation and extrapolation. *arXiv preprint arXiv:2502.00500*, 2025b.

Yuefan Cao, Xuyang Guo, Jiayan Huo, Yingyu Liang, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Zhen Zhuang. Text-to-image diffusion models cannot count, and prompt refinement cannot help. *arXiv preprint arXiv:2503.06884*, 2025c.

Bo Chen, Xiaoyu Li, Yingyu Liang, Jiangxuan Long, Zhenmei Shi, and Zhao Song. Circuit complexity bounds for RoPE-based transformer architecture. *arXiv preprint arXiv:2411.07602*, 2024a.

Bo Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Bypassing the exponential dependency: Looped transformers efficiently learn in-context by multi-step gradient descent. In *AISTATS*, 2025a.

Bo Chen, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. HSR-enhanced sparse attention acceleration. In *CPAL*, 2025b.

Bo Chen, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Provable failure of language models in learning majority boolean logic via gradient descent. *arXiv preprint arXiv:2504.04702*, 2025c.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *ICLR*, 2024.

Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

Yifang Chen, Jiayan Huo, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Fast gradient computation for RoPE attention in almost linear time. *arXiv preprint arXiv:2412.17316*, 2024b.

Yifang Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. The computational limits of state-space models and mamba via the lens of circuit complexity. In *CPAL*, 2025d.

Yifang Chen, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Fundamental limits of visual autoregressive transformers: Universal approximation abilities. In *ICML*, 2025e.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *CVPR*, 2023.

Maximilian Dax, Jonas Wildberger, Simon Buchholz, Stephen R Green, Jakob H Macke, and Bernhard Scholkopf. Flow matching for scalable simulation-based inference. In *NeurIPS*, 2023.

Yichuan Deng, Zhao Song, Yitan Wang, and Yuanyuan Yang. A nearly optimal size coreset algorithm with nearly linear time. *arXiv preprint arXiv:2210.08361*, 2022.

Yichuan Deng, Sridhar Mahadevan, and Zhao Song. Randomized and deterministic attention sparsification algorithms for over-parameterized feature dimension. *arXiv preprint arXiv:2304.04397*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. In *ICLR*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

Shibo Feng, Chunyan Miao, Zhong Zhang, and Peilin Zhao. Latent diffusion transformer for probabilistic time series forecasting. In *AAAI*, 2024.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023a.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL*, 2021.

Yeqi Gao, Sridhar Mahadevan, and Zhao Song. An over-parameterized exponential regression. *arXiv preprint arXiv:2303.16504*, 2023b.

Yeqi Gao, Zhao Song, and Junze Yin. Gradientcoin: A peer-to-peer decentralized large language models. *arXiv preprint arXiv:2308.10502*, 2023c.

Yeqi Gao, Zhao Song, Weixin Wang, and Junze Yin. A fast optimization view: Reformulating single layer attention in llm based on tensor and svm trick, and solving it in matrix multiplication time. In *UAI*, 2025a.

Yeqi Gao, Zhao Song, and Junze Yin. An iterative algorithm for rescaled hyperbolic functions regression. In *AISTATS*, 2025b.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *ECCV*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *ICML*, 2016.

Xuyang Guo, Zekai Huang, Jiayan Huo, Yingyu Liang, Zhenmei Shi, Zhao Song, and Jiahao Zhang. Can you count to nine? a human evaluation benchmark for counting limits in modern text-to-video models. *arXiv preprint arXiv:2504.04051*, 2025a.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025b.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vtextbench: A human evaluation benchmark for textual control in video generation models. *arXiv preprint arXiv:2505.04946*, 2025c.

Tiankai Hang and Shuyang Gu. Improved noise schedule for diffusion training. *arXiv preprint arXiv:2407.03297*, 2024.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. *arXiv preprint arXiv:2411.00698*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Jerry Yao-Chieh Hu, Weimin Wu, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (DiTs). In *NeurIPS*, 2024a.

Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. *arXiv preprint arXiv:2504.15956*, 2025a.

Jerry Yao-Chieh Hu, Maojiang Su, En-Jui Kuo, Zhao Song, and Han Liu. Computational limits of low-rank adaptation (lora) fine-tuning for transformer models. In *ICLR*, 2025b.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. In *ICLR*, 2025c.

Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. In *ICLR*, 2025d.

Jerry Yao-Chieh Hu, Xiwen Zhang, Maojiang Su, Zhao Song, and Han Liu. Minimalist softmax attention provably learns constrained boolean functions. *arXiv preprint arXiv:2505.19531*, 2025e.

Vincent Hu, Di Wu, Yuki Asano, Pascal Mettes, Basura Fernando, Björn Ommer, and Cees Snoek. Flow matching for conditional text generation in a few sampling steps. In *EACL*, 2024b.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *ECCV*, 2025.

Leonid Kantorovitch. On the translocation of masses. *Management science*, 1958.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

Yekun Ke, Xiaoyu Li, Yingyu Liang, Zhenmei Shi, and Zhao Song. Advancing the understanding of fixed point iterations in deep neural networks: A detailed analytical study. *arXiv preprint arXiv:2410.11279*, 2024.

Yekun Ke, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. On computational limits and provably efficient criteria of visual autoregressive models: A fine-grained complexity analysis. *arXiv preprint arXiv:2501.04377*, 2025.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *ICLR*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Leon Klein, Andreas Kramer, and Frank Noe. Equivariant flow matching. In *NeurIPS*, 2024.

Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *ICLR*, 2022.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

Chenyang Li, Yingyu Liang, Zhenmei Shi, Zhao Song, and Tianyi Zhou. Fourier circuits in neural networks and transformers: A case study of modular arithmetic with multiple inputs. In *AISTATS*, 2025a.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Xiaoyu Li, Jiangxuan Long, Zhao Song, and Tianyi Zhou. Fast second-order method for neural network under small treewidth setting. In *IEEE BigData*, 2024a.

Xiaoyu Li, Yingyu Liang, Zhenmei Shi, Zhao Song, Wei Wang, and Jiahao Zhang. On the computational capability of graph neural networks: A circuit complexity bound perspective. *arXiv preprint arXiv:2501.06444*, 2025b.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *ICLR*, 2024b.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Yufa Zhou. Multi-layer transformers gradient can be approximated in almost linear time. *arXiv preprint arXiv:2408.13233*, 2024.

Yingyu Liang, Jiangxuan Long, Zhenmei Shi, Zhao Song, and Yufa Zhou. Beyond linear approximations: A novel pruning approach for attention matrix. In *ICLR*, 2025a.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Differential privacy mechanisms in neural tangent kernel regression. In *WACV*, 2025b.

Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Yufa Zhou. Looped relu mlps may be all you need as practical programmable computers. In *AISTATS*, 2025c.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Chengyi Liu, Jiahao Zhang, Shijie Wang, Wenqi Fan, and Qing Li. Score-based generative diffusion models for social recommendations. *arXiv preprint arXiv:2412.15579*, 2024.

Feng Liu, Hanyang Wang, Jiahao Zhang, Ziwang Fu, Aimin Zhou, Jiayin Qi, and Zhibin Li. Evogan: An evolutionary computation assisted gan. *Neurocomputing*, 2022.

Hude Liu, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. Attention mechanism, max-affine partition, and universal approximation. *arXiv preprint arXiv:2504.19901*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. In *NeurIPS*, 2024.

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 *NeurIPS*, 2022.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2024.

James Martens et al. Deep learning via hessian-free optimization. In *ICML*, 2010.

Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 1997.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

G. Monge. *Memoire sur la théorie des déblais et des remblais*. Imprimerie royale, 1781.

Maxwell Nye, Anders Johan Andreassen, Gur AriGuy, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. Introducing ChatGPT, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *ICML*, 2023.

Lianke Qin, Zhao Song, and Baocheng Sun. Is solving graph neural tangent kernel equivalent to training graph neural network? *arXiv preprint arXiv:2309.07452*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Litu Rout, Advait Parulekar, Constantine Caramanis, and Sanjay Shakkottai. A theoretical justification for image inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2302.01217*, 2023.

Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *CVPR*, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Neta Shaul, Ricky TQ Chen, Maximilian Nickel, Matthew Le, and Yaron Lipman. On kinetic optimal probability paths for generative models. In *ICML*, 2023.

Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Yanyu Li, Yifan Gong, Kai Zhang, Hao Tan, Jason Kuen, Henghui Ding, et al. LazyDiT: Lazy learning for the acceleration of diffusion transformers. In *AAAI*, 2025a.

Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Jing Liu, Ruiyi Zhang, Ryan A Rossi, Hao Tan, Tong Yu, Xiang Chen, et al. Numerical pruning for efficient autoregressive models. In *AAAI*, 2025b.

Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *ICLR*, 2023.

Anshumali Shrivastava, Zhao Song, and Zhaozhuo Xu. A theoretical analysis of nearest neighbor search on approximate near neighbor graph. *arXiv preprint arXiv:2303.06210*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Zhao Song. *Matrix theory: optimization, concentration, and algorithms*. PhD thesis, The University of Texas at Austin, 2019.

Zhao Song, Weixin Wang, and Junze Yin. A unified scheme of resnet and softmax. *arXiv preprint arXiv:2309.13482*, 2023.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *ICML*, 2022.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

Cedric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Oriol Vinyals and Daniel Povey. Krylov subspace descent for deep learning. In *AISTATS*, 2012.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *ICML*, 2023.

Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *SIGIR*, 2023.

Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Yufei Xia, Jinzhu Li, Sheng Zhao, Jinyu Li, et al. An investigation of noise robustness for flow-matching-based zero-shot TTS. In *Interspeech*, 2024a.

Yilin Wang, Zeyuan Chen, Liangjun Zhong, Zheng Ding, Zhizhou Sha, and Zhuowen Tu. Dolfin: Diffusion layout transformers without autoencoder. In *ECCV*, 2024b.

Yilin Wang, Haiyang Xu, Xiang Zhang, Zeyuan Chen, Zhizhou Sha, Zirui Wang, and Zhuowen Tu. OmniControlNet: Dual-stage integration for conditional image generation. In *CVPR*, 2024c.

Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *CVPR*, 2024d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022.

Jerry Wei, Le Hou, Andrew Kyle Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. In *EMNLP*, 2023.

Yibo Wen, Chenwei Xu, Jerry Yao-Chieh Hu, and Han Liu. Alignab: Pareto-optimal energy alignment for designing nature-like antibodies. *arXiv preprint arXiv:2412.20984*, 2024.

Weimin Wu, Teng-Yun Hsiao, Jerry Yao-Chieh Hu, Wenxin Zhang, and Han Liu. In-context learning as conditioned associative memory retrieval. In *ICML*, 2025a.

Weimin Wu, Maojiang Su, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. In-context deep learning via transformer models. In *ICML*, 2025b.

Zike Wu, Pan Zhou, Kenji Kawaguchi, and Hanwang Zhang. Fast diffusion model. *arXiv preprint arXiv:2306.06991*, 2023.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, 2022.

Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, and Yingyu Liang. Towards few-shot adaptation of foundation models via multitask finetuning. In *ICLR*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jiahao Zhang, Feng Liu, and Aimin Zhou. Off-tanet: A lightweight neural micro-expression recognizer with optical flow features and integrated attention mechanism. In *PRICAI*, 2021.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024.

Jujia Zhao, Wenjie Wang, Yiyan Xu, Teng Sun, Fuli Feng, and Tat-Seng Chua. Denoising diffusion recommender model. In *SIGIR*, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *NeurIPS*, 2023.

NRFlow: Towards Noise-Robust Generative Modeling via High-Order Flow Matching (Supplementary Material)

Bo Chen[1] **Chengyue Gong**[2] **Xiaoyu Li**[3] **Yingyu Liang**[4,†] **Zhizhou Sha**[2] **Zhenmei Shi**[4] **Zhao Song**[5,*]
Mingda Wan[1] **Xugang Ye**[6]

[1]Middle Tennessee State University, [2]University of Texas at Austin, [3]University of New South Wales
[4]University of Wisconsin-Madison, [5]University of California, Berkeley, [6]TikTok
[*] magic.linuxkde@gmail.com, [†] yliang@cs.wisc.edu,yingyul@hku.hk

Roadmap. In Section A, we introduce some notations and basic concepts. We state the proof of Theorem 4.7 in Section B. In Section C, we extend our result to a third-order case. In Section D, we extend our result to k-th order. In Section E, we provide comprehensive experiments to evaluate our NRFlow under complex conditions.

A PRELIMINARY

In Section A.1, we introduce some notations we use in the appendix. In Section A.2, we introduce some basic concepts about flow matching. In Section A.3, we introduce the background of optimal transport.

A.1 NOTATIONS

We use $\Pr[]$ to denote the probability. We use $\mathbb{E}[]$ to denote the expectation. We use $\mathrm{Var}[]$ to denote the variance. We use $\|x\|_p$ to denote the ℓ_p norm of a vector $x \in \mathbb{R}^n$, i.e. $\|x\|_1 := \sum_{i=1}^n |x_i|$, $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$, and $\|x\|_\infty := \max_{i\in[n]} |x_i|$. For variables a, b, We write $a \lesssim b$ to indicate that a is bounded above by b up to a multiplicative constant independent of the main parameters. We write $a \gtrsim b$ to indicate that a is bounded below by b up to a multiplicative constant independent of the main parameters. We denote $\dot{x}^{(k)}$ as the k-th order derivative field of x. We use Dist as the function represents the probability distribution of a given random variable or random vector, mapping it to its corresponding measure on the probability space.

A.2 FLOW MATCHING

In this section, we restate and introduce some definitions of flow matching and the algorithm. We restate part of Definition 3.18 and introduce the loss function of flow matching.

Definition A.1 (Loss function)**.** *The loss function for the second order method contains two parts. We define the first part which is trying to using $\dot{x}_t$ in Fact 3.9, x_t and t to learn function $u_{1,t}$, thus the loss is*

$$L_{1\mathrm{st}} := \|\dot{x}_t - u_{1,\theta_1}(x_t, t)\|_2^2.$$

Here we restate Definition 3.10

Definition A.2 (A variant of flow matching in Lipman et al. [2023])**.** *Given two distributions μ_0 and π_0 on $\mathbb{R}^d$, flow matching aims to learn a time-dependent velocity field*

$$v_\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$$

such that for any trajectory x_t transporting $x_0 \sim \mu_0$ to $x_1 \sim \pi_0$, we have

$$\dot{x}_t \sim v_\theta(x_t, t).$$

We present the training algorithm and inference algorithm of flow matching.

Algorithm 3 Training algorithm of flow matching

procedure 1STORDERFORWARD()
  for each iteration **do**
    Random sample x_0 and time t, with target x_1
    $x_t \leftarrow \alpha_t \cdot x_0 + \sqrt{1-\alpha_t^2} \cdot x_1$
    Compute gradient with respect to $L_{1\text{st}}$ ▷ See Definition A.1
  end for
  return u_1 ▷ One network functions
end procedure

Algorithm 4 Inference algorithm of flow matching

procedure 1STORDERINFERENCE(u_1)
  $x_0 \sim \mathcal{N}(0, 1)$
  Initial $x \leftarrow x_0$
  for t from 0 to 1 with step $\Delta t = 0.01$ **do**
    $x \leftarrow x + \Delta t \cdot u_1(x, t)$
  end for
  return x
end procedure

A.3 OPTIMAL TRANSPORT

In this section, we introduce some background of optimal transport.

The optimal transport (OT) problem, as originally framed by Monge Monge [1781], seeks to minimize a cost functional:

$$\inf_{\mathcal{T}} \mathbb{E}[c(\mathcal{T}(x_0) - x_0)],$$
$$\text{s.t. } \text{Dist}(\mathcal{T}(x_0)) = \pi_0, \quad \text{Dist}(x_0) = \mu_0,$$

where the optimization is over deterministic mappings $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ that define a coupling (x_0, x_1) with $x_1 = \mathcal{T}(x_0)$, minimizing the cost c Villani [2009].

Kantorovich Kantorovitch [1958] extended Monge's problem by introducing the Monge-Kantorovich (MK) formulation, which allows for both deterministic and stochastic couplings (x_0, x_1) with marginal distributions μ_0 and π_0. Notably, when μ_0 is absolutely continuous with respect to the Lebesgue measure, the optimal coupling remains deterministic, reducing the problem to the set of mappings $\mathcal{T}$. This equivalence facilitates a dynamic interpretation, where the aim is to identify a continuous-time trajectory $\{x_t\}_{t\in[0,1]}$ from a collection of smooth interpolants $\mathcal{X}$, such that $x_0 \sim \mu_0$ and $x_1 \sim \pi_0$. For a convex cost function c, Jensen's inequality implies:

$$\mathbb{E}[c(x_1 - x_0)] \geq \inf_{\{x_t\}_{t\in[0,1]}\in\mathcal{X}} \mathbb{E}\left[\int_0^1 c(\dot{x}_t)\mathrm{d}t\right].$$

The infimum is achieved when x_t follows the displacement interpolant, $x_t = tx_1 + (1-t)x_0$, representing a geodesic in the Wasserstein space McCann [1997].

When the process is governed by ordinary differential equations (ODEs) of the form $\mathrm{d}x_t = v_t(x_t)\mathrm{d}t$, the evolution of the Lebesgue density ϵ_t of x_t satisfies the continuity equation:

$$\frac{\partial \epsilon_t}{\partial t} + \nabla \cdot (v_t \epsilon_t) = 0.$$

The Monge problem can then be reformulated dynamically as:

$$\inf_{\{v_t\}_{t\in[0,1]}, \{x_t\}_{t\in[0,1]}} \mathbb{E}\left[\int_0^1 c(v_t(x_t))\mathrm{d}t\right],$$
$$\text{s.t. } \frac{\partial \epsilon_t}{\partial t} + \nabla \cdot (v_t \epsilon_t) = 0,$$

$$\mu_0 = \frac{\mathrm{d}\mu_0}{\mathrm{d}\lambda}, \quad \pi_0 = \frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}.$$

Although this dynamic formulation provides deeper insights, solving it is computationally challenging. For cost functions like the ℓ_2 norm, this reduces to minimizing the kinetic energy of the flow, as shown by Shaul et al. [2023], where displacement interpolants are energy-optimal and correspond to straight-line flow paths.

B MISSING PROOF OF THEOREM 4.7

Here, we state the proof of Theorem 4.7.

Proof. Let $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2)$ denote the approximate optimal solution for the estimated loss in Eq. (3). By Lemma 4.5, we have

$$|\widetilde{L}_{2,\theta} - L_{2,\theta}| \leq O((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \ln(1/\beta))/N)^{1/2}.$$

Therefore, under the true distribution, $\dot{x}_t^{\mathrm{est}}$ and $\ddot{x}_t^{\mathrm{est}}$ approximate $\dot{x}_t^{\mathrm{true}}$ and $\ddot{x}_t^{\mathrm{true}}$ well in an L^2 sense.

As we defined $\ddot{x}_t^{\mathrm{true}} \in H^2(\Omega)$,we then apply Lemma 4.2, which leverages the Assumption 3.1. If $L_{2,2,\theta_1,\theta_2}$ is small, there exist C_{reg} such that

$$\begin{aligned} &\|\dot{x}_t^{\mathrm{est}} - \dot{x}_t^{\mathrm{true}}\|_{H^2(\Omega)} \\ &\leq C_{\mathrm{reg}}(L_{2,2,\theta_1,\theta_2}^{1/2} + \|\dot{x}_t^{\mathrm{est}} - \dot{x}_t^{\mathrm{true}}\|_{L^2(\Omega)}). \end{aligned}$$

Since Lemma 4.5 already guarantees that $\dot{x}_t^{\mathrm{est}}$ and $\ddot{x}_t^{\mathrm{est}}$ are close to the true $\dot{x}_t^{\mathrm{true}}$ and $\ddot{x}_t^{\mathrm{true}}$ in L^2, we conclude that the learned fields are also close in the stronger $H^2(\Omega)$ norm. As we assumed in Assumption 3.3 and 3.6, For or uniform time steps $\Delta t = 1/L$, the update for the estimate is

$$\begin{aligned} x_{l+1}^{\mathrm{est}} = x_l^{\mathrm{est}} &+ \Delta t \cdot u_{1,\widetilde{\theta}_1}(x_l^{\mathrm{est}}, t_l) \\ &+ \frac{(\Delta t)^2}{2} u_{2,\widetilde{\theta}_2}(u_{1,\widetilde{\theta}_1}(x_l^{\mathrm{est}}, t_l), x_l^{\mathrm{est}}, t_l). \end{aligned}$$

Similarly, the true trajectory x_{l+1}^{true} follows the same scheme but with the true velocity and acceleration (plus noise bounded by $\delta\epsilon$). Subtracting two updates and taking the H^2-norm, and invoking the Lipschitz condition in Assumption 3.2, yields Lemma 4.6, we have

$$\begin{aligned} e_{l+1} &= \|x_{l+1}^{\mathrm{est}} - x_{l+1}^{\mathrm{true}}\|_{H^2(\Omega)} \\ &\leq (1 + \Delta t \cdot C_{\mathrm{prop}}) e_l + C_{\mathrm{prop}} \Delta t \cdot \delta \cdot \epsilon. \end{aligned}$$

Here C_{prop} depends on Lipschitz constants and bounds on $\dot{x}$ and $\ddot{x}$. Iterating the above and a discrete Gronwell argument shows

$$\begin{aligned} e_L &= \|x_L^{\mathrm{est}} - x_L^{\mathrm{true}}\|_{H^2(\Omega)} \\ &\leq e_0 \exp(C_{\mathrm{prop}}) + \frac{\delta \cdot \epsilon}{C_{\mathrm{prop}}} (\exp(C_{\mathrm{prop}}) - 1). \end{aligned}$$

Since $\exp(C_{\mathrm{prop}})$ is just a constant factor, denote it by e^{C_2}. Combine all of these terms then yields the exact form of Theorem 4.7:

$$\begin{aligned} &\|x_{t=1}^{\mathrm{est}} - x_{t=1}^{\mathrm{true}}\|_{H^2(\Omega)} \\ &\leq C_1 \exp(C_2) \cdot (e_0 + \delta \cdot \epsilon) \\ &\quad + C_3 \cdot ((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \ln(1/\beta))/N)^{1/2}. \end{aligned}$$

Thus, we complete the proof. □

C EXTENSION ON THIRD-ORDER FLOW MATCHING

In this section, we extend the second-order flow-matching framework in Section 3.3 to incorporate third-order information. We first introduce additional assumptions in Section C.1 to ensure that the third derivative of the true trajectory is sufficiently smooth and bounded. In Section C.2, we introduce our third-order training algorithm and the inference algorithm. In Section C.3, we introduce the elliptic regularity for third-order cases. In Section C.4, we present the result of the regularization effect result for the third-order loss function. In Section C.5, we show the excess risk of third-order cases. In Section C.6, we show the lemma about discrete propagation under noise quantifies how noise in the trajectory affects the error propagation in a third-order discrete setting. In Section C.7, we prove our third-order main result.

C.1 PRELIMINARY

In this section, we introduce some additional definitions and assumptions specific to the third-order extension.

Assumption C.1 (Smoothness in higher Sobolev spaces)**.** *We assume $x_t^{\text{true}} \in H^3(\Omega)$, its derivatives up tothe third order lie in $L^2(\Omega)$ and satisfy suitable boundary conditions.*

$$\|\dot{x}_t^{\text{true}}\|_{H^3(\Omega)} \le M_1, \quad \|\ddot{x}_t^{\text{true}}\|_{H^3(\Omega)} \le M_2, \quad \|\dddot{x}_t^{\text{true}}\|_{H^3(\Omega)} \le M_3.$$

In addition, we assume the third derivative $\dddot{x}_t^{\text{true}}$ is continuous over $[0,1]$ and satisfies

$$\|\dddot{x}_t^{\text{true}}\|_\infty \le M_3.$$

The assumption above is critical to ensure the trajectory has sufficient regularity for third-order analysis.

Remark C.2. *Assumption C.1 extends Assumption 3.1 by requiring a bounded third derivative and ensuring the entire trajectory has appropriate regularity in Sobolev space $H^3(\Omega)$. This added smoothness is essential for deriving higher-order error bounds.*

We now define the discrete-time update rule for third-order systems.

Assumption C.3 (Time discretization for third-order update)**.** *Let $\Delta t = 1/L$ be the uniform step size, and define discrete times $t_l = l\Delta t$ for $l = 0, 1, \dots, L$. The third-order discrete update for the estimated system is:*

$$x_{l+1}^{\text{est}} = x_l^{\text{est}} + \Delta t u_{1,\theta_1}(x_l^{\text{est}}, t_l) + \frac{(\Delta t)^2}{2} u_{2,\theta_2}(u_{1,\theta_1}(x_l^{\text{est}}, t_l), x_l^{\text{est}}, t_l) + \frac{(\Delta t)^3}{6} u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_l^{\text{est}}, t_l).$$

The discrete update incorporates terms up to the third derivative, capturing the dynamics more accurately.

Assumption C.4. *The learned fields $u_{1,\theta_1}(x,t)$, $u_{2,\theta_2}(v,x,t)$ and $u_{3,\theta_3}(a,x,t)$ are L-Lipschitz continuous in spatial and temporal arguments. Formally, there exists $L > 0$ such that for all $x, y \in \mathbb{R}^d$ and $t, s \in [0,1]$:*

$$\begin{aligned}
\|u_{1,\theta_1}(x,t) - u_{1,\theta_1}(y,t)\|_2 &\le L\|x-y\|_2, \\
\|u_{2,\theta_2}(v,x,t) - u_{2,\theta_2}(v,y,t)\|_2 &\le L\|x-y\|_2, \\
\|u_{3,\theta_3}(a,x,t) - u_{3,\theta_3}(a,y,t)\|_2 &\le L\|x-y\|_2,
\end{aligned}$$

This is the natural extension of the second-order scheme in Assumption 3.6.

This assumption is necessary to control the propagation of errors through the system.

Definition C.5 (Third-order rectified flow)**.** *A* third-order rectified flow *is determined by three learned fields:*

$$\begin{aligned}
&u_{1,\theta_1}(x,t) \\
&u_{2,\theta_2}(v,x,t) \text{ where } \quad v = u_{1,\theta_1}(x,t), \\
&u_{3,\theta_3}(a,x,t) \text{ where } \quad a = u_{2,\theta_2}(v,x,t).
\end{aligned}$$

These fields aim to approximate $\dot{x}_t^{\text{true}}$, $\ddot{x}_t^{\text{true}}$, and $\dddot{x}_t^{\text{true}}$, respectively.

We now introduce the third-order analog of the velocity and acceleration fields. In addition to the velocity u_{1,θ_1} and acceleration u_{2,θ_2} fields, we define a field

$$u_{3,\theta_3}(a, x, t),$$

where $a = u_{2,\theta_2}(v, x, t)$ and $v = u_{1,\theta_1}(x, t)$. This field aims to approximate the third derivative $\dddot{x}_t^{\text{true}}$.

Here, we introduce the definition of the field of third-order flow.

Definition C.6 (Third-order flow field)**.** *A third-order rectified flow is characterized by a velocity field $u_{1,\theta_1}(x, t)$, an acceleration field $u_{2,\theta_2}(v, x, t)$, and a field*

$$u_{3,\theta_3}(a, x, t),$$

where

$$v = u_{1,\theta_1}(x, t), \quad a = u_{2,\theta_2}(u_{1,\theta_1}(x, t), x, t).$$

The function u_{3,θ_3} aims to approximate the $\dddot{x}_t^{\text{true}}$.

And we present the loss function of third-order flow as follows.

Definition C.7 (Third-order loss function)**.** *Let $\dot{x}_t^{\text{true}}$, $\ddot{x}_t^{\text{true}}$, and $\dddot{x}_t^{\text{true}}$ be the true velocity, acceleration, and of the trajectory x_t^{true}. We define the third-order loss as*

$$L_{3\text{rd}}(\theta_1, \theta_2, \theta_3) = \underbrace{\mathbb{E}[\|\dot{x}_t^{\text{true}} - u_{1,\theta_1}(x_t, t)\|_2^2]}_{L_{3,1,\theta_1}} + \underbrace{\mathbb{E}[\|\ddot{x}_t^{\text{true}} - u_{2,\theta_2}(u_{1,\theta_1}(x_t, t), x_t, t)\|_2^2]}_{L_{3,2,\theta_2,\theta_1}} + \underbrace{\mathbb{E}[\|\dddot{x}_t^{\text{true}} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_t, t)\|_2^2]}_{L_{3,3,\theta_3,\theta_2,\theta_1}},$$

where each expectation is taken over the possibly noisy samples of the continuous trajectory x_t^{true}.

Here's the empirical third-order loss.

Definition C.8 (Empirical third-order loss)**.** *Given a training dataset $\{(x_0^i, x_1^i)\}_{i=1}^N$ and time samples $\{t_i\}$, we define the empirical third-order loss:*

$$\widetilde{L}_{3\text{rd}} = \frac{1}{N}\sum_{i=1}^{N}[\|\dot{x}_t^{\text{true},i} - u_{1,\theta_1}(x_t^i, t_i)\|^2 + \|\ddot{x}_t^{\text{true},i} - u_{2,\theta_2}(u_{1,\theta_1}(x_t^i, t_i), x_t^i, t_i)\|^2 + \|\dddot{x}_t^{\text{true},i} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_t^i, t_i)\|^2].$$

C.2 PROPOSED THIRD-ORDER ALGORITHMS

We present the natural extension of the second-order methods in Section 3.4 to incorporate the jerk term. Here are our third-order training algorithm and inference algorithm.

Algorithm 5 Our third-order training algorithm

procedure 3RDORDERFORWARD()
  for each iteration **do**
    Random sample x_0 and time t, with target x_1
    $x_t \leftarrow \alpha_t \cdot x_0 + \sqrt{1 - \alpha_t^2} \cdot x_1$
    Compute gradient with respect to $L_{3\text{rd}}$ ▷ See Definition C.7
  end for
  return u_1, u_2, u_3 ▷ Three network functions
end procedure

C.3 ELLIPTIC REGULARITY

We now provide key lemmas and the main theorem establishing noise-robustness for third-order flow matching. In this section, we first introduce the elliptic regularity.

Lemma C.9 (Elliptic regularity in Evans [2010])**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain witha smooth boundary. Suppose a function $h : \Omega \to \mathbb{R}$ has weak derivatives up to order 3 in $L^2(\Omega)$ and satisfies relevant boundary conditions. Then there exists a constant $C_{\text{reg},3} > 0$ (depending on Ω) such that*

$$\|h\|_{H^3(\Omega)} \leq C_{\text{reg},3}(\|\nabla^2 h\|_{L^2(\Omega)} + \|\nabla h\|_{L^2(\Omega)} + \|h\|_{L^2(\Omega)}).$$

Algorithm 6 Our third-order inference algorithm

procedure 3RDORDERINFERENCE(u_1, u_2, u_3)
  $x_0 \sim \mathcal{N}(0, 1)$
  Initial $x \leftarrow x_0$
  for t from 0 to 1 with step $\Delta t = 0.01$ **do**
    $x \leftarrow x + \Delta t \cdot u_1(x, t) + \frac{(\Delta t)^2}{2} \cdot u_2(u_1(x, t), x, t) + \frac{(\Delta t)^3}{6} \cdot u_3(u_2(u_1(x, t), x, t), x_t, t)$
  end for
  return x
end procedure

C.4 REGULARIZATION EFFECT

In this section, we show the result of the regularization effect for the third-loss order.

Lemma C.10 (Regularization effect for third-order loss)**.** *As we defined in Definition C.8, then we define*

$$L_{3\mathrm{rd}} := \mathbb{E}[\| \dddot{x}_t^{\mathrm{true}} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_t^{\mathrm{true}}, t)\|^2].$$

If $\dddot{x}_t^{\mathrm{true}} \in H^3(\Omega)$ *and* $L_{3\mathrm{rd}}$ *is sufficiently small, then there exists a constant* $C_{\mathrm{reg},3}$ *such that*

$$\|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{H^3(\Omega)} \leq C_{\mathrm{reg},3}(L_{3\mathrm{rd}}^{1/2} + \|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{L^2(\Omega)}).$$

Proof. Applying Lemma C.9 to $h(\cdot) = \ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}$, we have

$$\begin{aligned} &\|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{H^3(\Omega)} \\ \leq\ & C_{\mathrm{reg},3}(\|\nabla^2 h\|_{L^2(\Omega)} + \|\nabla h\|_{L^2(\Omega)} + \|h\|_{L^2(\Omega)}). \end{aligned} \tag{8}$$

By the Definition of the loss function, a small $L_{2,2,\theta_1,\theta_2}$ implies

$$\|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{L^2(\Omega)} \lesssim L_{3\mathrm{rd}}^{1/2} \tag{9}$$

Combining Eq.(8) and (9), we have

$$\begin{aligned} &\|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{H^3(\Omega)} \\ \leq\ & C_{\mathrm{reg},3}(L_{3\mathrm{rd}}^{1/2} + \|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{L^2(\Omega)}). \end{aligned}$$

Thus, we complete the proof. □

C.5 EXCESS RISK

In this section, we first introduce some necessary tools that need to be used in Lemma C.10. Then, we show our result of excess risk for third-order flow. First, we restate the symmetrization bound again.

Lemma C.11 (Symmetrization bound, formal version of Lemma 4.3)**.** *Let* $\{x_i\}_{i=1}^N$ *and* $\{x_i'\}_{i=1}^N$ *be i.i.d. samples. For* $\mathcal{G} = \{\ell_\theta : \theta \in \Theta\}$*, we have:*

$$\sup_{g \in \mathcal{G}} \|\frac{1}{N}\sum_{i=1}^N (g(x_i) - g(x_i'))\| \leq \frac{2}{N}\mathop{\mathbb{E}}_{\sigma}[\sup_{g \in \mathcal{G}} \sum_{i=1}^N \sigma_i g(x_i)],$$

where $\{\sigma_i\}_{i=1}^N$ *are Rademacher random variables,* $\sigma_i \in \{+1, -1\}$ *with equal probability.*

Proof. For each σ_i has a symmetric distribution, we have:

$$\|\sum_{i=1}^N (g(x_i) - g(x_i'))\| \leq \mathop{\mathbb{E}}_{\sigma}[\|\sum_{i=1}^N \sigma_i (g(x_i) - g(x_i'))\|$$

Taking the supremum over $g \in \mathcal{G}$ and noting that $\{x_i\}$ and $\{x_i'\}$ have the same distribution, we can split the expression inside the absolute value:

$$\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} (g(x_i) - g(x_i')) \|$$
$$\leq \mathbb{E}_{\sigma}[\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sigma_i (g(x_i) - g(x_i')) \|].$$

By the triangle inequality, we get:

$$\| \sum_{i=1}^{N} \sigma_i (g(x_i) - g(x_i')) \| \leq \| \sum_{i=1}^{N} \sigma_i g(x_i) \| + \| \sum_{i=1}^{N} \sigma_i g(x_i') \|.$$

Hence,

$$\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} (g(x_i) - g(x_i')) \|$$
$$\leq \mathbb{E}_{\sigma}[\sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sigma_i \, g(x_i) \| + \sup_{g\in\mathcal{G}} \| \sum_{i=1}^{N} \sigma_i g(x_i')].$$

Because $\{x_i'\}$ is drawn from the same distribution as $\{x_i\}$, the two supremum terms have the same expected value. Therefore, we can combine them as follows:

$$\sup_{g\in\mathcal{G}} \| \frac{1}{N} \sum_{i=1}^{N} (g(x_i) - g(x_i')) \| \leq \frac{2}{N} \mathop{\mathbb{E}}_{\sigma}[\sup_{g\in\mathcal{G}} \sum_{i=1}^{N} \sigma_i g(x_i)],$$

Thus, we complete the proof. □

Here we restate Lemma 4.4.

Lemma C.12 (Formal version of Lemma 4.4)**.** *As we defined in Definition 3.15, C.7 and C.8, if Assumption 3.4 holds, for $g \in \mathcal{G}$ where $\mathcal{G} = \{\ell_\theta : \theta \in \Theta\}$, we have*

$$\sup_{g\in\mathcal{G}} \| \frac{1}{N} \sum_{i=1}^{N} g(x_i') - \mathbb{E}[g(x)] \| \leq O(\sqrt{\ln(1/\beta)/N})$$

We next present the result of excess risk for third-order flow.

Lemma C.13 (Excess risk)**.** *As we defined in Definition C.7 and C.8, we have*

$$\widetilde{L}_{3\mathrm{rd}} = \frac{1}{N} \sum_{i=1}^{N} [\| \dot{x}_t^{\mathrm{true},i} - u_{1,\theta_1}(x_t^i, t_i) \|^2 + \| \ddot{x}_t^{\mathrm{true},i} - u_{2,\theta_2}(u_{1,\theta_1}(x_t^i, t_i), x_t^i, t_i) \|^2 + \| \dddot{x}_t^{\mathrm{true},i} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_t^i, t_i) \|^2]$$

and

$$L_{3\mathrm{rd}}(\theta_1, \theta_2, \theta_3) = \underbrace{\mathbb{E}[\| \dot{x}_t^{\mathrm{true}} - u_{1,\theta_1}(x_t, t) \|^2]}_{L_{3,1,\theta_1}} + \underbrace{\mathbb{E}[\| \ddot{x}_t^{\mathrm{true}} - u_{2,\theta_2}(u_{1,\theta_1}(x_t, t), x_t, t) \|^2]}_{L_{3,2,\theta_2,\theta_1}} + \underbrace{\mathbb{E}[\| \dddot{x}_t^{\mathrm{true}} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_t, t) \|^2]}_{L_{3,3,\theta_3,\theta_2,\theta_1}},$$

Suppose $\mathcal{F}_1$ and $\mathcal{F}_2$ have finite or at most polynomially growing complexities $\mathcal{C}(\mathcal{F}_1), \mathcal{C}(\mathcal{F}_2)$. Then for $\beta \in (0, 1)$, with probability at least $1 - \beta$, we have

$$|\widetilde{L}_{3rd} - L_{3rd}| \leq O((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \ln(1/\beta))/N)^{1/2}.$$

Proof. Let $\mathcal{G} = \{\ell_\theta : \theta \in \Theta\}$ represent the complexity of $\mathcal{G}$ the Rademacher/VC dimension, As we defined in Definition C.5, C.7 and C.8 we calculate the empirical loss and population loss,

$$\widetilde{L}_{3\mathrm{rd}} = \frac{1}{N}\sum_{i=1}^{N}[\|\dot{x}_t^{\mathrm{true},i} - u_{1,\theta_1}(x_t^i, t_i)\|^2 + \|\ddot{x}_t^{\mathrm{true},i} - u_{2,\theta_2}(u_{1,\theta_1}(x_t^i, t_i), x_t^i, t_i)\|^2 + \|\dddot{x}_t^{\mathrm{true},i} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_t^i, t_i)\|^2]$$

and

$$L_{3\mathrm{rd}}(\theta_1,\theta_2,\theta_3) = \underbrace{\mathbb{E}[\|\dot{x}_t^{\mathrm{true}} - u_{1,\theta_1}(x_t,t)\|^2]}_{L_{3,1,\theta_1}} + \underbrace{\mathbb{E}[\|\ddot{x}_t^{\mathrm{true}} - u_{2,\theta_2}(u_{1,\theta_1}(x_t,t),x_t,t)\|^2]}_{L_{3,2,\theta_2,\theta_1}} + \underbrace{\mathbb{E}[\|\dddot{x}_t^{\mathrm{true}} - u_{3,\theta_3}(u_{2,\theta_2}(\cdot),x_t,t)\|^2]}_{L_{3,3,\theta_3,\theta_2,\theta_1}},$$

let $\{x_i'\}_{i=1}^N$ be an i.i.d. sample from the same distribution as $\{x_i\}_{i=1}^N$, and let $\{\sigma_i\}_{i=1}^N$ be i.i.d. Rademacher random variables ($\sigma_i \in \{+1,-1\}$ with probability $1/2$ each). Then, for any $g \in \mathcal{G}$, we have

$$\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N} g(x_i) - \mathbb{E}[g(x)]\| \le \sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}(g(x_i)) - g(x_i')\| + \sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N} g(x_i') - \mathbb{E}[g(x)]\| \tag{10}$$

We can upper bound the first term in Eq. (10),

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}(g(x_i)) - g(x_i')\| \\
\le\ & \frac{2}{N}\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}\sum_{i=1}^{N}\sigma_i g(x_i)] \\
=\ & 2\widetilde{\mathcal{R}}_N(\mathcal{G}) \\
\le\ & 2\cdot O(\sqrt{C(\mathcal{G})/N}) \\
\le\ & O(\sqrt{(C(\mathcal{F}_1)+C(\mathcal{F}_2))/N})
\end{aligned} \tag{11}$$

where the first step follows from Lemma C.11, the second step comes from we define $\widetilde{\mathcal{R}}_N(G) := \mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i g(x_i)]$, the third step follows from Assumption 3.4, the forth step follows from the definition of $\mathcal{G}$.

We can upper bound the second term in Eq. (10) by using Lemma C.12,

$$\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N} g(x_i') - \mathbb{E}[g(x)]\| \le O(\sqrt{\ln(1/\beta)/N}) \tag{12}$$

Loading Eq. (11) and Eq. (12), we can obtain

$$\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N} g(x_i) - \mathbb{E}[g(x)]\| \le O(\sqrt{(C(\mathcal{F}_1)+C(\mathcal{F}_2)+\ln(1/\beta))/N})$$

□

C.6 DISCRETE PROPAGATION

In this section, we show the lemma about discrete propagation under noise quantifies how noise in the trajectory affects the error propagation in a discrete setting for third-order flow.

Lemma C.14 (Discrete propagation with jerk)**.** *Under Assumptions C.1, C.3 and C.4 let*

$$e_l = \|x_l^{\mathrm{est}} - x_l^{\mathrm{true}}\|_{H^3(\Omega)}.$$

Then there is a constant $C_{\mathrm{prop},3} > 0$ such that

$$e_{l+1} \le (1 + \Delta t C_{\mathrm{prop},3}) e_l + C_{\mathrm{prop},3} \delta\epsilon \Delta t.$$

Unrolling from $l = 0$ to $l = L - 1$ with $\Delta t = 1/L$ yields

$$e_L \le e_0 \exp(C_{\mathrm{prop},3}) + \frac{\delta\epsilon}{C_{\mathrm{prop},3}} (\exp(C_{\mathrm{prop},3}) - 1).$$

Proof. By Assumptions C.3 and C.4, we have

$$x_{l+1}^{\mathrm{est}} = x_l^{\mathrm{est}} + \Delta t u_{1,\theta_1}(x_l^{\mathrm{est}}, t_l) + \frac{(\Delta t)^2}{2} u_{2,\theta_2}(u_{1,\theta_1}(\cdot), x_l^{\mathrm{est}}, t_l) + \frac{(\Delta t)^3}{6} u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_l^{\mathrm{est}}, t_l),$$
$$x_{l+1}^{\mathrm{true}} = x_l^{\mathrm{true}} + \Delta t \dot{x}_l^{\mathrm{true}} + \frac{(\Delta t)^2}{2} \ddot{x}_l^{\mathrm{true}} + \frac{(\Delta t)^3}{6} \dddot{x}_l^{\mathrm{true}}.$$

Subtracting the true update from the estimated one and taking the H^3-norm, the difference involves Lipschitz constants, the prior error e_l, and a noise term bounded by $\delta\epsilon$. One obtains

$$\|x_{l+1}^{\mathrm{est}} - x_{l+1}^{\mathrm{true}}\|_{H^3(\Omega)} \le (1 + \Delta t \cdot C_{\mathrm{prop},3}) \|x_l^{\mathrm{est}} - x_l^{\mathrm{true}}\|_{H^3(\Omega)} + C_{\mathrm{prop},3} \delta\epsilon \Delta t.$$

where $C_{\mathrm{prop},3}$ depends on Lipschitz constants of u_{1,θ_1}, u_{2,θ_2}, u_{3,θ_3}, and the boundedness of $\ddot{x}_t^{\mathrm{true}}$, $\dddot{x}_t^{\mathrm{true}}$.

Repeating this inequality from $l = 0$ to $l = L - 1$ and noting $\Delta t = 1/L$, by a discrete Gronwall argument we have

$$\begin{aligned} e_L &= \|x_L^{\mathrm{est}} - x_L^{\mathrm{true}}\|_{H^3(\Omega)} \\ &\le e_0 \exp(C_{\mathrm{prop},3}) + \sum_{l=0}^{L-1} (C_{\mathrm{prop},3} \delta\epsilon \Delta t \prod_{j=l+1}^{L-1} (1 + \Delta t C_{\mathrm{prop},3})), \\ &\le e_0 \exp(C_{\mathrm{prop},3}) + \frac{\delta\epsilon}{C_{\mathrm{prop},3}} (\exp(C_{\mathrm{prop},3}) - 1). \end{aligned}$$

This completes the proof. □

C.7 MAIN RESULT: THIRD-ORDER NOISE ROBUSTNESS

Combining the above, we obtain the final noise-robustness result for third-order flow matching in this section.

Theorem C.15 (Third-order noise robustness)**.** *Suppose Assumptions C.1, C.3, 3.2 and 3.3 hold. Let $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2, \widetilde{\theta}_3)$ be an approximately optimal solution minimizing the empirical loss $\widetilde{L}_{3\mathrm{rd},\theta_1,\theta_2,\theta_3}$. Then, with probability at least $1 - \beta$, for uniform time steps $t_l = l\Delta t$ with $\Delta t = 1/L$, we have*

$$\|x_L^{\mathrm{est}} - x_L^{\mathrm{true}}\|_{H^3(\Omega)} \le C_1' \exp(C_2')(e_0 + \delta\epsilon) + C_3'((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \mathcal{C}(\mathcal{F}_3) + \ln(1/\beta))/N)^{1/2},$$

where $e_0 = \|x_0^{\mathrm{est}} - x_0^{\mathrm{true}}\|_{H^3(\Omega)}$ denotes the initial error, and C_1', C_2', C_3' depend on Lipschitz constants, the dimension d, and Sobolev embedding constants in $H^3(\Omega)$.

Proof. Let $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2)$ denote the approximate optimal solution for the estimated loss, use Lemma C.13, we have

$$|\widetilde{L}_{3\mathrm{rd}} - L_{3\mathrm{rd}}| \le O((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \mathcal{C}(\mathcal{F}_3) \ln(1/\beta))/N)^{1/2}.$$

Therefore, under the true distribution, $\dot{x}_t^{\mathrm{est}}$ and $\ddot{x}_t^{\mathrm{est}}$ approximate $\dot{x}_t^{\mathrm{true}}$ and $\ddot{x}_t^{\mathrm{true}}$ well in an L^2 sense.

As we defined $\dddot{x}_t^{\mathrm{true}} \in H^3(\Omega)$,we then apply Lemma C.10, which leverages the Assumption C.1. If $L_{3\mathrm{rd}}$ is small, there exist $C_{\mathrm{reg},3}$ such that

$$\|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{H^3(\Omega)} \le C_{\mathrm{reg},3}(L_{3\mathrm{rd}}^{1/2} + \|\ddot{x}_t^{\mathrm{est}} - \ddot{x}_t^{\mathrm{true}}\|_{L^2(\Omega)}).$$

We conclude that the learned fields are also close to the stronger $H^2(\Omega)$ norm. As we assumed in Assumption 3.3 and C.3, For or uniform time steps $\Delta t = 1/L$, the update for the estimate is

$$x_{l+1}^{\text{est}} = x_l^{\text{est}} + \Delta t u_{1,\widetilde{\theta}_1}(x_l^{\text{est}}, t_l) + \frac{(\Delta t)^2}{2} u_{2,\widetilde{\theta}_2}(u_{1,\widetilde{\theta}_1}(x_l^{\text{est}}, t_l), x_l^{\text{est}}, t_l) + \frac{(\Delta t)^3}{6} u_{3,\widetilde{\theta}_3}(u_{2,\widetilde{\theta}_2}(\cdot), x_l^{\text{est}}, t_l),$$
$$x_{l+1}^{\text{true}} = x_l^{\text{true}} + \Delta t \dot{x}_l^{\text{true}} + \frac{(\Delta t)^2}{2} \ddot{x}_l^{\text{true}} + \frac{(\Delta t)^3}{6} \dddot{x}_l^{\text{true}}.$$

Subtracting two updates and taking the H^3-norm, and invoking the Lipschitz condition in Assumption C.4, yields Lemma C.14, we have

$$e_{l+1} = \|x_{l+1}^{\text{est}} - x_{l+1}^{\text{true}}\|_{H^3(\Omega)} \leq (1 + \Delta t \cdot C_{\text{prop},3}) e_l + C_{\text{prop},3} \Delta t \cdot \delta \cdot \epsilon.$$

Here $C_{\text{prop},3}$ depends on Lipschitz constants and bounds on $\dot{x}$ and $\ddot{x}$, $\dddot{x}$. Iterating the above and a discrete Gronwell argument shows

$$e_L = \|x_L^{\text{est}} - x_L^{\text{true}}\|_{H^3(\Omega)} \leq e_0 \exp(C_{\text{prop},3}) + \frac{\delta \cdot \epsilon}{C_{\text{prop},3}} (\exp(C_{\text{prop},3}) - 1).$$

Since $\exp(C_{\text{prop},3})$ is just a constant factor, denote it by e^{C_2}. Combine all of these terms then yields the exact form of Theorem C.15:

$$\|x_L^{\text{est}} - x_L^{\text{true}}\|_{H^3(\Omega)} \leq C_1' \exp(C_2')(e_0 + \delta\epsilon) + C_3'((\mathcal{C}(\mathcal{F}_1) + \mathcal{C}(\mathcal{F}_2) + \mathcal{C}(\mathcal{F}_3) + \ln(1/\beta))/N)^{1/2},$$

Thus, we complete the proof. □

D EXTENSION ON k-TH ORDER FLOW MATCHING

In this section, we extend the second-order flow-matching framework in Section 3.3 to incorporate third-order information. We first introduce additional assumptions in Section D.1 to ensure that the third derivative of the true trajectory is sufficiently smooth and bounded. In Section D.2, we introduce our third-order training algorithm and the inference algorithm. In Section D.3, we introduce the elliptic regularity for the third-order case. In Section D.4, we present the result of the regularization effect result for the third-order loss function. In Section D.5, we show the excess risk of the third-order case. In Section D.6, we show the lemma about discrete propagation under noise quantifies how noise in the trajectory affects the error propagation in a third-order discrete setting. In Section D.7, we prove our k-th order main result.

D.1 PRELIMINARY

In this section, we introduce some additional definitions and assumptions specific to the k-th order extension.

Assumption D.1 (Smoothness in higher Sobolev spaces)**.** *We assume the true trajectory $x_t^{\text{true}} \in H^k(\Omega)$ and that its derivatives up to the k-th order are sufficiently smooth and bounded. Formally, there exist constants $\{M_j\}_{j=1}^k > 0$ such that*

$$\|\dot{x}_t^{(j),\text{true}}\|_{H^k(\Omega)} \leq M_j, \quad \text{for} j = 1, \ldots, k,$$

where $\dot{x}_t^{(j),\text{true}}$ denotes the j-th order time derivative of x_t^{true}. We also require these derivatives to be continuous on $[0, 1]$ in the time variable.

Then, we introduce our assumption for time discretization under k-th order update.

Assumption D.2 (Time discretization for k-th order update)**.** *Let $\Delta t = 1/L$ be the uniform step size, and define discrete times $t_l = l\Delta t$ for $l = 0, 1, \ldots, L$. We consider the following k-th order discrete update for the estimated system:*

$$x_{l+1}^{\text{est}} = x_l^{\text{est}} + \sum_{j=1}^{k} \frac{(\Delta t)^j}{j!} u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\ldots u_{1,\theta_1}(x_l^{\text{est}}, t_l)\ldots), x_l^{\text{est}}, t_l), \tag{13}$$

where each u_{j,θ_j} is a learned field approximating the j-th order derivative $\dot{x}_t^{(j),\text{true}}$.

The k-th order Lipschitz continuity is also necessary, and we present it here.

Assumption D.3 (k-th order Lipschitz continuity)**.** *We assume the learned fields $(u_{j,\theta_j})_{j=1}^k$ are each L-Lipschitz continuous in their spatial and temporal arguments. Formally, there exists $L > 0$ such that for any $x, y \in \mathbb{R}^d$ and $t, s \in [0, 1]$*

$$\|u_{j,\theta_j}(\dots, x, t) - u_{j,\theta_j}(\dots, y, t)\|_2 \le L\|x - y\|_2$$
$$\|u_{j,\theta_j}(\dots, x, t) - u_{j,\theta_j}(\dots, x, s)\|_2 \le L\|t - s\|_2.$$

Next we introduce the definition of k-th order loss function.

Definition D.4 (k-th order flow)**.** *A k-order flow involves a sequence of learned fields u_{j,θ_j} for $j = 1, \dots, k$, each targeting the approximation of $(\dot{x}_t^{(j)})^{\text{True}}$.*

Here is the k-th order loss function.

Definition D.5 (k-th order loss function)**.** *The k-order loss function evaluates the accuracy of approximations for each derivative:*

$$L_{\text{k-order}}(\theta_1, \dots, \theta_k) = \sum_{j=1}^{k} \mathbb{E}[\|(\dot{x}_t^{(j)})^{\text{True}} - u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\dots), x_t, t)\|^2].$$

And we introduce empirical k-th order loss here.

Definition D.6 (Empirical k-th order loss)**.** *Given a training dataset $\{(x_0^i, x_1^i)\}_{i=1}^N$ with times $\{t_i\}$ and (approximate) ground-truth derivatives up to the k-th order, the empirical k-th order loss is*

$$\widetilde{L}_{\text{k-order}}(\theta_1, \dots, \theta_k) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} \|(\dot{x}_{t_i}^{(j)})^{\text{true},i} - u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\dots), x_{t_i}^i, t_i)\|^2.$$

D.2 PROPOSED k-TH ORDER ALGORITHMS

In this section, we show our k-th order training algorithm and inference algorithm. First, we show the k-th order training algorithm.

Algorithm 7 Our k-th order training algorithm

```
procedure K-THORDERFORWARD()
    for each iteration do
        Random sample x_0 and time t, with target x_1
        x_t ← α_t · x_0 + √(1 − α_t^2) · x_1
        Compute gradient with respect to L_{k-order}        ▷ See Definition D.5
    end for
    return u_1, u_2, ··· , u_k                              ▷ k network functions
end procedure
```

D.3 ELLIPTIC REGULARITY

In this section, we introduce the first result, which is a classical result that characterizes the relationship between different Sobolev norms for sufficiently smooth functions for k-th order flow.

Lemma D.7 (Elliptic regularity in Evans [2010])**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a sufficiently smooth boundary. Suppose $h : \Omega \to \mathbb{R}$ has weak derivatives up to order k in $L^2(\Omega)$. Then there is a constant $C_{\text{reg},k} > 0$ depending on Ω such that*

$$\|h\|_{H^k(\Omega)} \le C_{\text{reg},k}\Big(\sum_{m=0}^{k-1} \|\nabla^m h\|_{L^2(\Omega)}\Big).$$

Algorithm 8 Our k-th order inference algorithm

procedure K-THORDERINFERENCE($u_1, \ldots, u_k$)
  $x_0 \sim \mathcal{N}(0, 1)$
  Initialize $x \leftarrow x_0$
  for t from 0 to 1 with step $\Delta t = 0.01$ **do**
    $x \leftarrow x + \sum_{j=1}^{k} \frac{(\Delta t)^j}{j!} \cdot u_j(u_{j-1}(\ldots u_1(x, t)\ldots), x, t)$
  end for
  return x
end procedure

D.4 REGULARIZATION EFFECT

In this section, we now connect the second-order loss function with the Sobolev norm of the estimation error under the k-th order loss function.

Lemma D.8 (Regularization effect for k-th order loss)**.** *As we defined in Definition D.6, then we define*

$$L_{\mathrm{k-order}}(\theta_1, \ldots, \theta_k) = \sum_{j=1}^{k} \mathbb{E}[\|(\dot{x}_t^{(j)})^{\mathrm{True}} - u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\ldots), x_t, t)\|^2].$$

If $(\dot{x}_t^{\mathrm{true}})^{(k)} \in H^3(\Omega)$ *and* $L_{\mathrm{k-order}}$ *is sufficiently small, then there exists a constant* C_{regk} *such that*

$$\|(\dot{x}_t^{\mathrm{est}})^{(k-1)} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}\|_{H^k(\Omega)} \leq C_{\mathrm{regk}}(L_{\mathrm{k-order}}^{1/2} + \|(\dot{x}_t^{\mathrm{est}})^{(k-1)} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}\|_{L^2(\Omega)}).$$

Proof. Applying Lemma D.7 to $h(\cdot) = \ddot{x}_t^{\mathrm{est}} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}$, we have

$$\begin{aligned} &\|(\dot{x}_t^{\mathrm{est}})^{(k-1)} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}\|_{H^k(\Omega)} \\ \leq\ & C_{\mathrm{regk}}(\sum_{m=0}^{k-1} \|\nabla^m h\|_{L^2(\Omega)}). \end{aligned} \tag{14}$$

By the Definition of the loss function, a small $L_{2,2,\theta_1,\theta_2}$ implies

$$\|(\dot{x}_t^{\mathrm{est}})^{(k-1)} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}\|_{L^2(\Omega)} \lesssim L_{\mathrm{k-order}}^{1/2} \tag{15}$$

Combining Eq.(14) and (15), we have

$$\begin{aligned} &\|(\dot{x}_t^{\mathrm{est}})^{(k-1)} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}\|_{H^k(\Omega)} \\ \leq\ & C_{\mathrm{regk}}(L_{\mathrm{k-order}}^{1/2} + \|(\dot{x}_t^{\mathrm{est}})^{(k-1)} - (\dot{x}_t^{\mathrm{true}})^{(k-1)}\|_{L^2(\Omega)}). \end{aligned}$$

Thus, we complete the proof. □

D.5 EXCESS RISK

In this section, we present a result that quantifies the gap between the empirical and population loss, highlighting the strong generalization capabilities of our method even with a finite sample size.

Lemma D.9 (Excess risk for k-th order)**.** *As in Definition D.5 and D.6, let*

$$\widetilde{L}_{\mathrm{k-order}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} \|(\dot{x}_{t_i}^{(j)})^{\mathrm{true},i} - u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\ldots), x_{t_i}^i, t_i)\|^2$$

and

$$L_{\mathrm{k-order}}(\theta_1,\dots,\theta_k)=\sum_{j=1}^{k}\mathbb{E}[\|\dot{x}_t^{(j),\mathrm{true}}-u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\dots),x_t,t)\|^2].$$

Suppose each function class $\mathcal{F}_j$ has finite or at most polynomially growing complexity $\mathcal{C}(\mathcal{F}_j)$. Then for $\beta\in(0,1)$, with probability at least $1-\beta$, we have

$$|\widetilde{L}_{\mathrm{k-order}}-L_{\mathrm{k-order}}|\le O(\sum_{i=1}^{k}C(\mathcal{F}_i)+\ln(1/\beta))/N)^{1/2}.$$

Proof. Let $\mathcal{G}=\{\ell_\theta:\theta\in\Theta\}$ represent the complexity of $\mathcal{G}$ the Rademacher/VC dimension, as we defined in Definition C.5, C.7 and C.8 we calculate the empirical loss and population loss,

$$\widetilde{L}_{\mathrm{k-order}}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}\|(\dot{x}_{t_i}^{(j)})^{\mathrm{true},i}-u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\dots),x_{t_i}^i,t_i)\|^2$$

and

$$L_{\mathrm{k-order}}(\theta_1,\dots,\theta_k)=\sum_{j=1}^{k}\mathbb{E}[\|\dot{x}_t^{(j),\mathrm{true}}-u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\dots),x_t,t)\|^2].$$

let $\{x_i'\}_{i=1}^N$ be an i.i.d. sample from the same distribution as $\{x_i\}_{i=1}^N$, and let $\{\sigma_i\}_{i=1}^N$ be i.i.d. Rademacher random variables ($\sigma_i\in\{+1,-1\}$ with probability $1/2$ each). Then, for any $g\in\mathcal{G}$, we have

$$\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}g(x_i)-\mathbb{E}[g(x)]\|\le\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}(g(x_i))-g(x_i')\|+\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}g(x_i')-\mathbb{E}[g(x)]\| \tag{16}$$

We can upper bound the first term in Eq. (16),

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}(g(x_i))-g(x_i')\|\\
\le&\frac{2}{N}\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}\sum_{i=1}^{N}\sigma_i g(x_i)]\\
=&2\widetilde{\mathcal{R}}_N(\mathcal{G})\\
\le&2\cdot O(\sqrt{C(\mathcal{G})/N})\\
\le&O((\sum_{i=1}^{k}C(\mathcal{F}_i)/N)^{1/2})
\end{aligned}\tag{17}$$

where the first step follows from Lemma C.11, the second step comes from we define $\widetilde{\mathcal{R}}_N(G):=\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}}\frac{1}{N}\sum_{i=1}^{N}\sigma_i g(x_i)]$, the third step follows from Assumption 3.4, the forth step follows from the definition of $\mathcal{G}$ in k-th order case.

We can upper bound the second term in Eq. (16) by using Lemma C.12,

$$\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}g(x_i')-\mathbb{E}[g(x)]\|\le O(\sqrt{\ln(1/\beta)/N}) \tag{18}$$

Loading Eq. (17) and Eq. (18), we can obtain

$$\sup_{g\in\mathcal{G}}\|\frac{1}{N}\sum_{i=1}^{N}g(x_i)-\mathbb{E}[g(x)]\|\le O(((\sum_{i=1}^{k}C(\mathcal{F}_i)+\ln(1/\beta))/N)^{1/2})$$

□

D.6 DISCRETE PROPAGATION

In this section, we show the lemma about discrete propagation under noise quantifies how noise in the trajectory affects the error propagation in a discrete setting for k-th order flow.

Lemma D.10 (Discrete propagation under noise for k-th order)**.** *Under Assumptions D.2 and D.3, let*

$$e_l = \|x_l^{\text{est}} - x_l^{\text{true}}\|_{H^k(\Omega)}.$$

Then there is a constant $C_{\text{prop},k} > 0$ such that

$$e_{l+1} \le (1 + \Delta t C_{\text{prop},k}) e_l + C_{\text{prop},k} \delta\epsilon \Delta t.$$

Iterating from $l = 0$ to $l = L - 1$ (where $\Delta t = 1/L$) gives

$$e_L \le e_0 \exp(C_{\text{prop},k}) + \frac{\delta\epsilon}{C_{\text{prop},k}} (\exp(C_{\text{prop},k}) - 1).$$

Proof. By Assumptions D.2 and D.3, we have

$$\begin{aligned} x_{l+1}^{\text{est}} &= x_l^{\text{est}} + \Delta t u_{1,\theta_1}(x_l^{\text{est}}, t_l) + \frac{(\Delta t)^2}{2} u_{2,\theta_2}(u_{1,\theta_1}(\cdot), x_l^{\text{est}}, t_l) + \frac{(\Delta t)^3}{6} u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_l^{\text{est}}, t_l) \\ &\quad + \sum_{j=4}^{k} \frac{(\Delta t)^j}{j!} u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\dots u_{1,\theta_1}(x_l^{\text{est}}, t_l) \dots), x_l^{\text{est}}, t_l), \\ x_{l+1}^{\text{true}} &= x_l^{\text{true}} + \Delta t \dot{x}_l^{\text{true}} + \frac{(\Delta t)^2}{2} \ddot{x}_l^{\text{true}} + \frac{(\Delta t)^3}{6} \dddot{x}_l^{\text{true}} + \sum_{j=4}^{k} \frac{(\Delta t)^j}{j!} (\dot{x}_t^{\text{true}})^{(j)}. \end{aligned}$$

Subtracting the true update from the estimated one and taking the H^k-norm, the difference involves Lipschitz constants, the prior error e_l, and a noise term bounded by $\delta\epsilon$. One obtains

$$\|x_{l+1}^{\text{est}} - x_{l+1}^{\text{true}}\|_{H^k(\Omega)} \le (1 + \Delta t \cdot C_{\text{prop,k}}) \|x_l^{\text{est}} - x_l^{\text{true}}\|_{H^k(\Omega)} + C_{\text{prop,k}} \delta\epsilon \Delta t.$$

where $C_{\text{prop,k}}$ depends on Lipschitz constants of $u_{1,\theta_1}, u_{2,\theta_2}, \dots u_{k,\theta_k}$, and the boundedness of $(\dot{x}_t^{\text{true}})^{(k-1)}, (\dot{x}_t^{\text{true}})^k$.

Repeating this inequality from $l = 0$ to $l = L - 1$ and noting $\Delta t = 1/L$, by a discrete Gronwall argument we have

$$\begin{aligned} e_L &= \|x_L^{\text{est}} - x_L^{\text{true}}\|_{H^k(\Omega)} \\ &\le e_0 \exp(C_{\text{prop,k}}) + \sum_{l=0}^{L-1} (C_{\text{prop,k}} \delta\epsilon \Delta t \prod_{j=l+1}^{L-1} (1 + \Delta t C_{\text{prop,k}})), \\ &\le s e_0 \exp(C_{\text{prop,k}}) + \frac{\delta\epsilon}{C_{\text{prop,k}}} (\exp(C_{\text{prop,k}}) - 1). \end{aligned}$$

This completes the proof. □

D.7 MAIN RESULT FOR k-TH ORDER NOISE ROBUSTNESS

In this section, we formally state and prove our main result, leveraging the auxiliary lemmas to demonstrate the robustness of the learned trajectory against noise and the effects of finite sample sizes for k-th order flow.

Theorem D.11 (Noise robustness for k-th order flow matching)**.** *Suppose Assumptions D.1, D.2, D.3, and 3.3 hold. Let*

$$\widetilde{\theta} = (\widetilde{\theta}_1, \dots, \widetilde{\theta}_k)$$

be an approximately optimal solution minimizing the empirical k-th order loss in Definition D.6. Then, with probability at least $1 - \beta$, the final-time estimate x_L^{est} satisfies:

$$\|x_L^{\text{est}} - x_L^{\text{true}}\|_{H^k(\Omega)} \le C_1'' \exp(C_2'')(e_0 + \delta\epsilon) + C_3''((\sum_{i=1}^{k} \mathcal{C}(\mathcal{F}_i) + \ln(1/\beta))/N)^{1/2},$$

where $e_0 = \|x_0^{\text{est}} - x_0^{\text{true}}\|_{H^k(\Omega)}$ *and* C_1'', C_2'', C_3'' *depend on the Lipschitz constants, Sobolev embedding constants, and dimension* d*. The term* $e^{C_2''}$ *arises from the discrete Gronwall factor over* $[0, 1]$.

Proof. Let $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k)$ denote the approximate optimal solution for the estimated loss, use Lemma D.9, we have

$$|\widetilde{L}_{\text{k-order}} - L_{\text{k-order}}| \leq O(\sum_{i=1}^{k} C(\mathcal{F}_i) + \ln(1/\beta))/N)^{1/2}.$$

Therefore, under the true distribution, $\dot{x}_t^{\text{est}}$ and $\ddot{x}_t^{\text{est}}$ approximate $\dot{x}_t^{\text{true}}$ and $\ddot{x}_t^{\text{true}}$ well in an L^2 sense, the higher order field also hold.

As we defined $(\dot{x}_t^{\text{true}})^{(k)} \in H^k(\Omega)$,we then apply Lemma D.8, which leverages the Assumption D.1. If $L_{\text{k-order}}$ is small, there exist C_{regk} such that

$$\|(\dot{x}_t^{\text{est}})^{(k-1)} - (\dot{x}_t^{\text{true}})^{(k-1)}\|_{H^k(\Omega)} \leq C_{\text{regk}}(L_{\text{k-order}}^{1/2} + \|(\dot{x}_t^{\text{est}})^{(k-1)} - (\dot{x}_t^{\text{true}})^{(k-1)}\|_{L^2(\Omega)}).$$

We conclude that the learned fields are also close to the stronger $H^2(\Omega)$ norm. As we assumed in Assumption 3.3 and D.2, For or uniform time steps $\Delta t = 1/L$, the update for the estimate is

$$\begin{aligned} x_{l+1}^{\text{est}} &= x_l^{\text{est}} + \Delta t u_{1,\theta_1}(x_l^{\text{est}}, t_l) + \frac{(\Delta t)^2}{2} u_{2,\theta_2}(u_{1,\theta_1}(\cdot), x_l^{\text{est}}, t_l) + \frac{(\Delta t)^3}{6} u_{3,\theta_3}(u_{2,\theta_2}(\cdot), x_l^{\text{est}}, t_l) \\ &\quad + \sum_{j=4}^{k} \frac{(\Delta t)^j}{j!} u_{j,\theta_j}(u_{j-1,\theta_{j-1}}(\ldots u_{1,\theta_1}(x_l^{\text{est}}, t_l)\ldots), x_l^{\text{est}}, t_l), \\ x_{l+1}^{\text{true}} &= x_l^{\text{true}} + \Delta t \dot{x}_l^{\text{true}} + \frac{(\Delta t)^2}{2} \ddot{x}_l^{\text{true}} + \frac{(\Delta t)^3}{6} \dddot{x}_l^{\text{true}} + \sum_{j=4}^{k} \frac{(\Delta t)^j}{j!} (\dot{x}_t^{\text{true}})^{(j)}. \end{aligned}$$

Subtracting two updates and taking the H^k-norm, and invoking the Lipschitz condition in Assumption D.3, yields Lemma C.14, we have

$$e_{l+1} = \|x_{l+1}^{\text{est}} - x_{l+1}^{\text{true}}\|_{H^k(\Omega)} \leq (1 + \Delta t \cdot C_{\text{prop,k}}) e_l + C_{\text{prop,k}} \Delta t \cdot \delta \cdot \epsilon.$$

Here $C_{\text{prop,k}}$ depends on Lipschitz constants and bounds on $\dot{x}$, $\ddot{x}, \ldots, \dot{x}_t^{(k)}$. Iterating the above and a discrete Gronwell argument shows

$$e_L = \|x_L^{\text{est}} - x_L^{\text{true}}\|_{H^k(\Omega)} \leq e_0 \exp(C_{\text{prop,k}}) + \frac{\delta \cdot \epsilon}{C_{\text{prop,k}}} (\exp(C_{\text{prop,k}}) - 1).$$

Since $\exp(C_{\text{prop}})$ is just a constant factor, denote it by e^{C_2}. Combine all of these terms then yields the exact form of Theorem D.11:

$$\|x_L^{\text{est}} - x_L^{\text{true}}\|_{H^k(\Omega)} \leq C_1'' \exp(C_2'')(e_0 + \delta\epsilon) + C_3''((\sum_{i=1}^{k} \mathcal{C}(\mathcal{F}_i) + \ln(1/\beta))/N)^{1/2},$$

Thus, we complete the proof. □

E EMPIRICAL ABLATION STUDY

In Section E.1, we introduce the three datasets used in our experiments: the five-mode, 3-round spiral, and dot-hyperbola datasets. In Section E.2, we present results for the first-order loss applied to the datasets, and in Section E.3, we examine the effect of the second-order loss. Section E.4 extends this analysis by including the third-order loss.

E.1 THREE DATASET

We employ three datasets for our experiments: the five-mode dataset, the 3-round spiral dataset, and the dot-hyperbola Gaussian mixture distribution dataset, all with a variance of 0.3 for each Gaussian component. In the five-mode dataset, five source modes (**orange**) are positioned at a distance of $D_0 = 6$ from the origin, and five target modes (**pink**) are positioned at $D_0 = 13$, each mode containing 200 sampled points. For the 3-round spiral dataset, 600 points are drawn from Gaussian distributions, each with a variance of 0.3, for both the source and target distributions. Similarly, the dot-hyperbola dataset consists of 900 points sampled from Gaussian distributions with a variance of 0.3 for both the source and target.

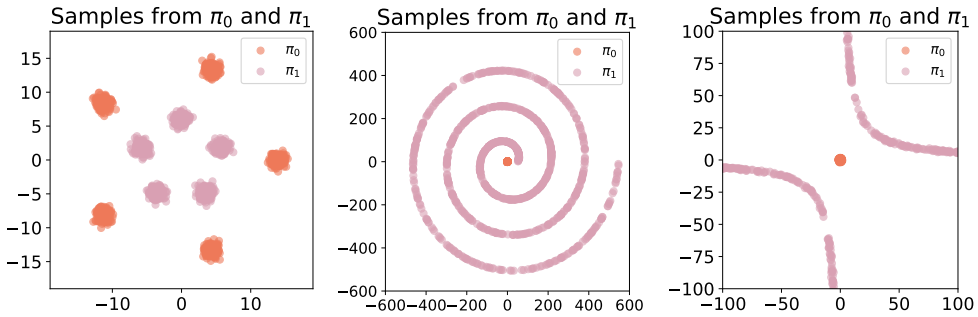


Figure 2: Gaussian mixture distributions visualized: five-mode dataset (**Left**), 3-round spiral dataset (**Middle**), and dot-hyperbola dataset (**Right**). The primary objective is for NRFlow to learn the transport trajectory from the source distribution π_0 (**orange**) to the target distribution π_1 (**pink**).

E.2 ONLY FIRST ORDER LOSS

The models are optimized by minimizing the sum of squared error (SSE). Both the source and target distributions are Gaussian. The target transport trajectory is modeled using the VP ODE framework from Liu et al. [2023], expressed as $x_t = \alpha_t x_0 + \beta_t x_1$. The parameters α_t and β_t are defined as $\alpha_t = \exp(-\frac{1}{4}a(1-t)^2 - \frac{1}{2}b(1-t))$ and $\beta_t = \sqrt{1-\alpha_t^2}$, with hyperparameters $a = 19.9$ and $b = 0.1$. In each of the five-mode, 3-round spiral, and dot-hyperbola datasets, 100 points are sampled from both the source and target distributions for each mode. The five-mode dataset training involves an ODE solver and Adam optimizer, using a 2-layer MLP with 100 hidden dimensions, a batch size of 800, a learning rate of 0.005, and 2000 training steps. For the 3-round spiral dataset, the training setup is similar, except with a batch size of 1000 and 1000 training steps. For the dot-hyperbola dataset, the batch size is increased to 1600 while maintaining the same learning rate and optimizer settings and 1000 training steps.

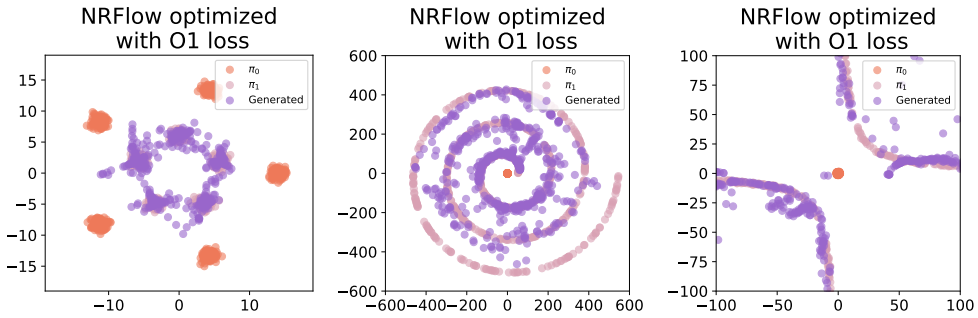


Figure 3: NRFlow generated distributions optimized by the first-order loss only: five-mode dataset (**Left**), 3-round spiral dataset (**Middle**), and dot-hyperbola dataset (**Right**). The source distribution π_0 (**orange**), the target distribution π_1 (**pink**), and the generated distribution (**purple**) are shown.

E.3 SECOND ORDER NRFLOW

The models are optimized by minimizing the sum of squared error (SSE). Both the source and target distributions are Gaussian. The target transport trajectory is modeled using the VP ODE framework from Liu et al. [2023], expressed as $x_t = \alpha_t x_0 + \beta_t x_1$. The parameters α_t and β_t are defined as $\alpha_t = \exp(-\frac{1}{4}a(1-t)^2 - \frac{1}{2}b(1-t))$ and $\beta_t = \sqrt{1-\alpha_t^2}$, with hyperparameters $a = 19.9$ and $b = 0.1$. In each of the five-mode, 3-round spiral, and dot-hyperbola datasets, 100 points are sampled from both the source and target distributions for each mode. The five-mode dataset training involves an ODE solver and Adam optimizer, using a 2-layer MLP with 100 hidden dimensions, a batch size of 800, a learning rate of 0.005, and 2000 training steps. For the 3-round spiral dataset, the training setup is similar, except with a batch size of 1000 and 1000 training steps. For the dot-hyperbola dataset, the batch size is increased to 1600 while maintaining the same learning rate and optimizer settings and 1000 training steps.

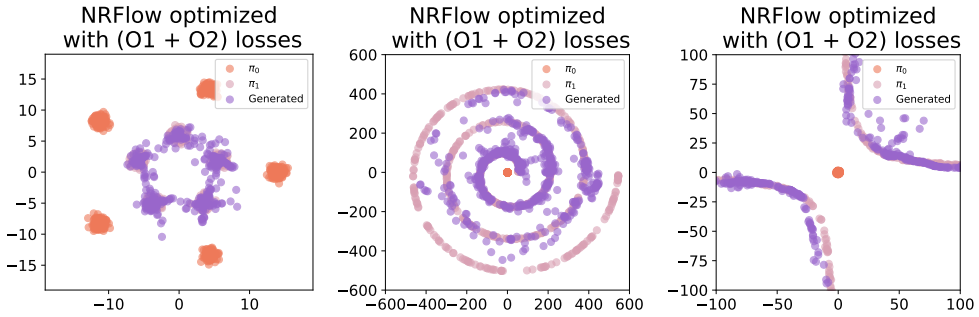


Figure 4: NRFlow generated distributions optimized by the first order and second order losses: five-mode dataset (**Left**), 3-round spiral dataset (**Middle**), and dot-hyperbola dataset (**Right**). The source distribution π_0 (**orange**), the target distribution π_1 (**pink**), and the generated distribution (**purple**) are shown.

E.4 THIRD ORDER NRFLOW

The models are optimized by minimizing the sum of squared error (SSE). Both the source and target distributions are Gaussian. The target transport trajectory is modeled using the VP ODE framework from Liu et al. [2023], expressed as $x_t = \alpha_t x_0 + \beta_t x_1$. The parameters α_t and β_t are defined as $\alpha_t = \exp(-\frac{1}{4}a(1-t)^2 - \frac{1}{2}b(1-t))$ and $\beta_t = \sqrt{1-\alpha_t^2}$, with hyperparameters $a = 19.9$ and $b = 0.1$. In each of the five-mode, 3-round spiral, and dot-hyperbola datasets, 100 points are sampled from both the source and target distributions for each mode. The five-mode dataset training involves an ODE solver and Adam optimizer, using a 2-layer MLP with 100 hidden dimensions, a batch size of 800, a learning rate of 0.005, and 2000 training steps. For the 3-round spiral dataset, the training setup is similar, except with a batch size of 1000 and 1000 training steps. For the dot-hyperbola dataset, the batch size is increased to 1600 while maintaining the same learning rate and optimizer settings and 1000 training steps.

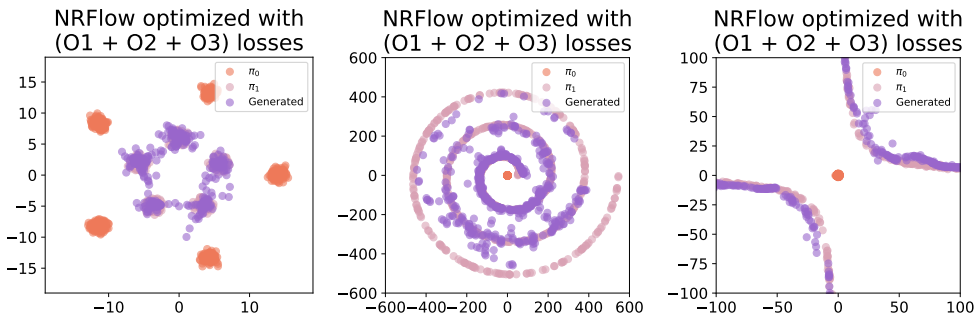


Figure 5: NRFlow generated distributions optimized by the first-order, second-order, and third-order losses: five-mode dataset (**Left**), 3-round spiral dataset (**Middle**), and dot-hyperbola dataset (**Right**). The source distribution π_0 (**orange**), the target distribution π_1 (**pink**), and the generated distribution (**purple**) are shown.