Escaping the Curse of Dimensionality in One-Step Flow-Based Generative Models

Konstantin Yakovlev
HSE University
kdyakovlev@hse.ru

Abstract

Flow and diffusion-based generative models achieve state-of-the-art quality, but require multiple forward passes at inference time, making them computationally costly. Consistency models overcome this by learning a flow map directly, enabling few-step generation and dramatically reducing the computation needed at test time. However, formal statistical guarantees for such one-step generators are scarce. Existing results either suffer from the curse of dimensionality or rely on impractical loss functions. We investigate whether one-step flow-based models can escape the curse of dimensionality under the assumption that the data distribution is generated by a low-dimensional surface perturbed with Gaussian noise. In this setup, we derive approximation and generalization error bounds for the flow map estimate that circumvent the curse of dimensionality. Our findings provide the first principled explanation for practical success of one-step flow-based generative models.

1 Introduction

Generative models based on flow [Lipman et al., 2023, Liu et al., 2023, Albergo et al., 2025] and diffusion [Ho et al., 2020, Song et al., 2021, Karras et al., 2022] have demonstrated exceptional performance across various domains, including images [Esser et al., 2024], audio [Lu et al., 2022], and video [Ho et al., 2022]. However, these models require a large number of evaluations of the learned model, presenting a significant computational challenge. A promising avenue is to estimate the flow map rather than the instantaneous velocity, which is accomplished within the framework of Consistency Models [Song et al., 2023, Song and Dhariwal, 2024]. This approach enables few-step generation, resulting in a reduction in sampling time by several orders of magnitude compared to traditional models. The recently developed methods for learning flow maps [Kim et al., 2024, Boffi et al., 2025a,b, Geng et al., 2025, Lu and Song, 2025, Frans et al., 2025, Zhang et al., 2026] effectively bridge the gap between few-step models and multi-step diffusion and flow-based models.

Despite the empirical success of these methods, theoretical guarantees regarding the statistical efficiency of distribution estimation using few-step generative models are scarce. Dou et al. [2024] made the first effort: under mild assumptions they proved convergence rates for a one-step generator, but the bounds suffer from the curse of dimensionality and rely on an impractical Wasserstein-loss consistency criterion. Li et al. [2025] obtained quantitative guarantees for distribution estimation of a one-step consistency model. However, their iterative training framework deviates from the framework of Song et al. [2023] and requires training $\mathcal{O}(D^{5/2}/\varepsilon)$ models from scratch, where D is the ambient dimension and ε is the target accuracy. This is highly impractical. Ding et al. [2024] studied statistical efficiency of flow-based one-step generative models and derived a convergence rate that suffers from the curse of dimensionality. Under the assumption that the data distribution is supported on a low-dimensional linear subspace, convolved with Gaussian noise, the bound was substantially improved. However, this assumption is not satisfied in practice.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Overall, the aforementioned works do not explain the success of flow map learning in practice. Note that in many practical scenarios the data distribution has an underlying low-dimensional structure [Bengio et al., 2013, Pope et al., 2021, Brown et al., 2023]. It has been recognized that generative models which exploit such intrinsic structure can escape the curse of dimensionality [Stephanovitch et al., 2024, Tang and Yang, 2024, Yakovlev and Puchkin, 2025, Chakraborty and Bartlett, 2025]. For these models, the exponent of n in the convergence rate is determined by the intrinsic dimension. This naturally raises the following research question:

Can one-step flow-based models escape the curse of dimensionality?

To answer this question, we consider the self-distillation framework introduced by Boffi et al. [2025a,b]. We focus on this framework because it admits a well-defined population objective and has proven effective for learning flow maps in practice.

Contributions. In this paper, we affirmatively answer the above question. Specifically, we investigate the statistical efficiency of self-distillation [Boffi et al., 2025b] under the noisy setting as in Yakovlev and Puchkin [2025]. Our main contributions are:

1. We establish that a flow map can be approximated without the curse of dimensionality, an insight that extends beyond the scope of this work.
2. We demonstrate that the flow map estimator obtained via self-distillation achieves a convergence upper bound of $\widetilde{\mathcal{O}}(n^{-\frac{2\beta}{5\beta+2d}})$, thereby avoiding the curse of dimensionality. Here, d denotes the intrinsic dimension, which can be substantially smaller than the ambient dimension D.
3. Finally, we show that, for a wide class of stochastic interpolant coefficients, the learned flow map recovers the target distribution without the curse of dimensionality in terms of the 2-Wasserstein distance.

Notation. Throughout the paper, we use $f \lesssim g$ and $g \gtrsim f$ to denote $f = \mathcal{O}(g)$. If both $f \lesssim g$ and $g \lesssim f$ hold, then we write $f \asymp g$. Furthermore, the notation $\widetilde{\mathcal{O}}(\cdot)$ is defined similarly to $\mathcal{O}(\cdot)$ except that the logarithmic factors are hidden. For any $R > 0$, the Euclidean ball in $\mathbb{R}^r$ centered at the origin is denoted by $\mathcal{B}(0, R)$. We denote the set of non-negative integers by $\mathbb{Z}_+$. For a multi-index $\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{Z}_+^r$ and a vector $v = (v_1, \dots, v_r) \in \mathbb{R}^r$, we define $v^{\mathbf{k}} = v_1^{k_1} v_2^{k_2} \dots v_r^{k_r}$ and $|\mathbf{k}| = k_1 + k_2 + \dots + k_r$. We denote the 2-Wasserstein distance between probability measures μ, ν by $W_2(\mu, \nu)$ and the distribution of a random vector X by $\mathrm{Law}(X)$. Lastly, we utilize the operators $a \vee b = \max(a, b)$, $a \wedge b := \min(a, b)$, and $\lfloor \beta \rfloor$ for the largest integer strictly less than $\beta > 0$.

1.1 Other Related Works

Statistical theory of diffusion and flow-based models. The estimation of the velocity field and the score function (namely, the gradient of the log-density) is a central topic in recent generative modeling literature [Oko et al., 2023, Wibisono et al., 2024, Zhang et al., 2024, Fukumizu et al., 2025]. While the manifold hypothesis is commonly employed to avoid the curse of dimensionality [Tang and Yang, 2024, Azangulov et al., 2024], such approaches rely on restrictive geometric or density assumptions. Yakovlev and Puchkin [2025] recently proposed a more general assumption that better suits real-world data. However, existing theory focuses mainly on instantaneous velocity learning. Our work extends this to the flow map learning, which is a more challenging problem.

Convergence of consistency models. A separate line of research focuses on the iteration complexity of consistency models [Lyu et al., 2024, Yang et al., 2025, Chen et al., 2025], typically assuming that the flow map has been accurately estimated. However, these works do not address the flow map estimation problem itself, leaving a critical gap in the theoretical understanding of statistical efficiency of such models.

2 Preliminaries and notations

Norms. For a vector v and a matrix A, we denote their Euclidean and operator norms by $\|v\|$ and $\|A\|$, respectively. The notations $\|\cdot\|_0$ and $\|\cdot\|_\infty$ refer to the number of non-zero entries and the

maximum absolute entry value of a vector or matrix. For a vector-valued function f defined on a set $\Omega \subseteq \mathbb{R}^r$ and a non-negative weighting function $p : \Omega \to \mathbb{R}$, we denote the L^∞ and weighted L^2 norms as

$$\|f\|_{L^\infty(\Omega)} = \sup_{x\in\Omega} \|f(x)\| \quad \text{and} \quad \|f\|_{L^2(\mathsf{p})} = \left\{ \int_\Omega \|f(x)\|^2 \mathsf{p}(x)\,\mathrm{d}x \right\}^{1/2}.$$

Finally, for a random variable ξ, we denote its ψ_1-norm by $\|\xi\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(|\xi|/t) \leqslant 2\}$.

Stochastic Interpolants. Let $X_1 \sim \mathsf{p}^*$, where p^* is the probability distribution on $\mathbb{R}^D$ we aim to sample from. The *stochastic interpolant* framework [Albergo et al., 2025] operates as follows. The stochastic interpolant is a stochastic process $(X_t)_{0\leqslant t\leqslant 1}$ defined by independently sampling $X_0 \sim \mathcal{N}(0, I_D)$, $X_1 \sim \mathsf{p}^*$ and setting

$$X_t = \alpha_t X_0 + \beta_t X_1. \tag{1}$$

Here, α_t and β_t are continuous on $[0, 1]$ and differentiable on $[0, 1)$, satisfying the boundary conditions $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$. Let p_t^* be the distribution of X_t. Under the regularity and integrability conditions of [Albergo et al., 2025, Theorem 6], the velocity field $b^*(y, t) = \mathbb{E}[\dot{X}_t \,|\, X_t = y]$ generates a deterministic flow between $\mathcal{N}(0, I_D)$ and p^*. Specifically, for

$$\dot{Z}_t = b^*(Z_t, t), \quad 0 \leqslant t \leqslant 1, \quad Z_0 \sim \mathcal{N}(0, I_D)$$

it holds that Z_t and X_t have the same distributions for every $t \in [0, 1]$. The challenge is that the velocity field b^* is unknown and should be estimated from data using flow matching objective. Specifically, given i.i.d. samples $Y_1, Y_2, \ldots, Y_n$ from p^*, the velocity field estimate $\widehat{b}(y, t)$ is defined as the empirical risk minimizer

$$\widehat{b} \in \operatorname*{argmin}_{b\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^n \ell_{FM}(b, Y_i), \quad \ell_{FM}(b, y) = \int_0^{1-t_0} \mathbb{E}[\|b(X_t, t) - \dot{X}_t\|^2 \,|X_1 = y]\,\mathrm{d}t, \tag{2}$$

where $\mathcal{B}$ is the hypothesis class and $t_0 \in [0, 1)$ is the early stopping time. We use this objective only when $\int_0^{1-t_0} \mathbb{E}\|\dot{X}_t\|^2\,\mathrm{d}t < \infty$. Early stopping is needed for some schedules, including the diffusion interpolant, but is not needed for the linear interpolant in our positive-noise setting. Its role in related estimation problems is discussed in Fukumizu et al. [2025], Oko et al. [2023], Yakovlev and Puchkin [2025]. The flow matching objective has the following property.

Proposition 2.1. *For every $b \in \mathcal{B}$ such that the maximum of $\mathbb{E}\|b(X_t, t)\|^2$ over $t \in [0, 1 - t_0]$ is finite, the following holds:*

$$\mathbb{E}[\ell_{FM}(b, X_1) - \ell_{FM}(b^*, X_1)] = \int_0^{1-t_0} \mathbb{E}\|b(X_t, t) - b^*(X_t, t)\|^2\,\mathrm{d}t. \tag{3}$$

The proof of Proposition 2.1 is deferred to Appendix A.1.

Self-Distillation. The self-distillation framework of Boffi et al. [2025b] accelerates sampling by directly learning a flow map that transports a simple base distribution to the target distribution in a single step. For an arbitrary $y \in \mathbb{R}^D$ and $0 \leqslant s \leqslant t \leqslant 1 - t_0$, consider the flow generated by the true velocity field $\dot{Z}_\tau = b^*(Z_\tau, \tau)$ with $s \leqslant \tau \leqslant t$ and $Z_s = y$. For every $0 \leqslant s < t \leqslant 1 - t_0$ and $y \in \mathbb{R}^D$, we define the actual flow map

$$\Psi^*_{s,t}(y) = y + \int_s^t b^*(Z_\tau, \tau)\,\mathrm{d}\tau = y + (t - s)v^*_{s,t}(y), \quad v^*_{s,t}(y) = \frac{1}{t-s}\int_s^t b^*(Z_\tau, \tau)\,\mathrm{d}\tau. \tag{4}$$

Following the notation in Geng et al. [2025], we refer to $v^*_{s,t}(y)$ as the *average velocity*. On the diagonal, define $\Psi^*_{t,t}(y) = y$ and $v^*_{t,t}(y) = b^*(y, t)$ by continuous extension, for all $y \in \mathbb{R}^D$ and $t \in [0, 1 - t_0]$. Moreover, Boffi et al. [2025b] demonstrate that $\Psi^*_{s,t}$ solves the Lagrangian equation

$$\partial_t \Psi^*_{s,t}(y) = b^*(\Psi^*_{s,t}(y), t), \quad \text{for } 0 \leqslant s < t \leqslant 1 - t_0 \text{ and } y \in \mathbb{R}^D. \tag{5}$$

Inspired by this property, we parametrize a candidate map as $\Psi_{s,t}(y) = y + (t-s)v_{s,t}(y)$ and define the self-distillation objective

$$\ell_{SD}(v, y) = \ell_{FM}(v, y) + \int_0^{1-t_0} \mathbb{E}\|\partial_t \Psi_{0,t}(X_0) - v_{t,t}(\Psi_{0,t}(X_0))\|^2 \, \mathrm{d}t. \tag{6}$$

Here, the first term is the flow matching loss defined in (2), where we identify the velocity candidate $b(y, t)$ with $v_{t,t}(y)$. The second term, which we denote by $R(v)$, penalizes deviations from the Lagrangian consistency condition (5). This penalty controls the discrepancy between the time derivative of the candidate trajectory and the instantaneous velocity. Consequently, the loss function in (6) is expressed as $\ell_{SD}(v, y) = \ell_{FM}(v, y) + R(v)$. Moreover, since v^* satisfies the Lagrangian equation (5), the consistency penalty vanishes, i.e., $R(v^*) = 0$. Given i.i.d. data points $Y_1, Y_2, \ldots, Y_n \sim \mathsf{p}^*$, the learned flow map $\widehat{\Psi}_{0,t}(y) = y + t\widehat{v}_{0,t}(y)$ is defined through empirical risk minimization

$$\widehat{v} \in \operatorname*{argmin}_{v \in \mathcal{V}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{SD}(v, Y_i) \right\}. \tag{7}$$

Here, $\mathcal{V}$ denotes the set of average velocity estimators. Finally, we emphasize that in the original formulation of Boffi et al. [2025b] the consistency term was integrated over both time variables s and t to enable few-step sampling. In this work, we show that integrating only over t is sufficient to establish one-step generation guarantees.

Neural Networks. In this paper we use the Gaussian error linear unit (GELU) activation [Hendrycks and Gimpel, 2016], defined as $\mathrm{GELU}(x) = x \cdot \Phi(x)$, where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution. We choose GELU because it is infinitely differentiable and has bounded derivatives. This allows us to invoke the framework of Yakovlev and Puchkin [2026] to construct function approximations with an explicitly controlled Lipschitz constant, which is crucial for our estimation-error analysis. This choice is also supported by the widespread use of GELU in modern neural network architectures [Brown et al., 2020, Peebles and Xie, 2023, Biderman et al., 2023].

For a shift vector $b = (b_1, \ldots, b_r) \in \mathbb{R}^r$ we use the shifted GELU, which maps a vector $x \in \mathbb{R}^r$ to $\mathbb{R}^r$ as $\mathrm{GELU}_b(x) = \big(\mathrm{GELU}(x_1 - b_1), \ldots, \mathrm{GELU}(x_r - b_r)\big)$. For $L \in \mathbb{N}$ and an architecture vector $W = (W_0, W_1, \ldots, W_L) \in \mathbb{N}^{L+1}$, a feedforward neural network of depth L and architecture W is a function $f : \mathbb{R}^{W_0} \to \mathbb{R}^{W_L}$ that is defined as the composition of linear maps and GELU nonlinearities:

$$f(x) = -b_L + A_L \circ \mathrm{GELU}_{b_{L-1}} \circ A_{L-1} \circ \mathrm{GELU}_{b_{L-2}} \circ \ldots \circ A_2 \circ \mathrm{GELU}_{b_1} \circ A_1 \circ x, \tag{8}$$

where $A_j \in \mathbb{R}^{W_j \times W_{j-1}}$ is a weight matrix and $b_j \in \mathbb{R}^{W_j}$ is a bias vector for every $j \in \{1, \ldots, L\}$. The maximum number of neurons in each layer is defined as $\|W\|_\infty$, which stands for the width of the network. We further restrict the class of neural networks f of the form (8) to those with at most $S \in \mathbb{N}$ non-zero parameters, with all weights and biases bounded by $B > 0$:

$$\mathsf{NN}(L, W, S, B) = \left\{ f : \sum_{j=1}^{L} \big(\|A_j\|_0 + \|b_j\|_0\big) \leqslant S, \max_{1 \leqslant j \leqslant L} \{\|A_j\|_\infty \vee \|b_j\|_\infty\} \leqslant B \right\}. \tag{9}$$

Smoothness classes. Let $\Omega \subseteq \mathbb{R}^r$. For a multi-index $\mathbf{k} \in \mathbb{Z}_+^r$, the partial derivative of a function $f : \Omega \to \mathbb{R}$ is denoted by

$$\partial^{\mathbf{k}} f(x) = \frac{\partial^{|\mathbf{k}|} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_r^{k_r}}, \quad x \in \Omega.$$

For $\beta > 0$, the *Hölder class* $\mathcal{H}^\beta(\Omega, \mathbb{R}, H)$ is the set of functions f such that

$$\max_{\substack{\mathbf{k} \in \mathbb{Z}_+^r \\ |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} \leqslant H \quad \text{and} \quad \max_{\substack{\mathbf{k} \in \mathbb{Z}_+^r \\ |\mathbf{k}| = \lfloor \beta \rfloor}} \sup_{\substack{x, y \in \Omega \\ x \neq y}} \frac{|\partial^{\mathbf{k}} f(x) - \partial^{\mathbf{k}} f(y)|}{1 \wedge \|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \leqslant H.$$

A vector-valued function $f : \Omega \to \mathbb{R}^m$ belongs to $\mathcal{H}^\beta(\Omega, \mathbb{R}^m, H)$ if every component function is in $\mathcal{H}^\beta(\Omega, \mathbb{R}, H)$.

Additionally, for $k \in \mathbb{Z}_+$, we denote by $W^{k,\infty}(\Omega)$ the Sobolev space of vector-valued functions whose derivatives $\partial^{\mathbf{k}} f$ are in $L^\infty(\Omega)$ for all $\mathbf{k} \in \mathbb{Z}_+^r$ with $|\mathbf{k}| \leqslant k$. For $j \in \{0, \dots, k\}$, the associated Sobolev seminorm is defined as

$$|f|_{W^{j,\infty}(\Omega)} = \max\left\{\|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} \,:\, \mathbf{k} \in \mathbb{Z}_+^r,\ |\mathbf{k}| = j\right\}, \quad f \in W^{k,\infty}(\Omega).$$

3 Results

In this section, we present the main findings of our paper. Before proceeding with upper bounds on the flow map approximation and estimation accuracy, we first detail the assumptions on the data distribution under which the bounds hold.

3.1 Assumptions

The key to sharp error estimates is the following structural assumption. As in Yakovlev and Puchkin [2025], Yakovlev et al. [2025], we model the target distribution p^* as a nonparametric Gaussian mixture, where the distribution of means is supported on a smooth low-dimensional surface.

Assumption 3.1 (relaxed manifold assumption, [Yakovlev and Puchkin, 2025])**.** *Let $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$, for some $H > 0$ and $d, D \in \mathbb{N}$, with $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$. We consider the model where the observed samples $Y_1, \dots, Y_n$ are independent copies of a random element $X_1 \in \mathbb{R}^D$, generated as follows:*

$$X_1 = g^*(U) + \sigma_{\mathrm{data}} Z,$$

where $U \sim \mathrm{Un}([0,1]^d)$ and $Z \sim \mathcal{N}(0, I_D)$ are independent and $0 < \sigma_{\min} \leqslant \sigma_{\mathrm{data}} < 1$.

Under Assumption 3.1, the target distribution is supported near a low-dimensional surface $\mathrm{Im}(g^*)$. Note that the effective dimension d can be substantially smaller than the ambient dimension D. Crucially, this assumption corresponds to scenarios where the data distribution exhibits low-dimensional structure and well-separated components, as supported by empirical studies [Khayatkhoei et al., 2018, Pope et al., 2021, Brown et al., 2023]. In addition, this assumption relaxes the standard manifold hypothesis employed in the analysis of generative models [Dahal et al., 2022, Tang and Yang, 2024, Azangulov et al., 2024] by eliminating the requirement of a density lower bound, a condition that is non-typical for real-world data (see the discussion in Zhang et al. [2024]). Compared with Yakovlev and Puchkin [2025], we further impose that the Gaussian noise level is strictly positive, $\sigma_{\mathrm{data}} \geqslant \sigma_{\min} > 0$. Positive $\sigma_{\min}$ allows for better spatial Lipschitz constant of the velocity field $b^*(y,t)$, which facilitates our analysis and allows us to employ a broad class of interpolant coefficients. We acknowledge that the assumption excludes scenarios involving noiseless data, and we leave this aspect for future research. Finally, note that Assumption 3.1 aligns with Assumption 2 in Ding et al. [2024], where the target distribution is a Gaussian convolution.

We now impose regularity conditions on the interpolant coefficients α_t and β_t.

Assumption 3.2. *The functions $\alpha_t, \beta_t : [0,1] \to \mathbb{R}$ satisfy*

(i) $\alpha_0 = \beta_1 = 1$ *and* $\alpha_1 = \beta_0 = 0$.

(ii) $\frac{1}{2}\sigma_{\mathrm{data}}^2 \leqslant \alpha_t^2 + \beta_t^2 \sigma_{\mathrm{data}}^2 \leqslant 1$ *and* $0 \leqslant \beta_t \leqslant 1$ *for all* $0 \leqslant t \leqslant 1$.

(iii) $\alpha \in C^2([0,1)) \cap C([0,1])$ *and* $\alpha^2, \beta \in C^3([0,1])$.

(iv) the function α_t is monotonically decreasing on the interval $[0,1]$.

Note that Assumption 3.2 covers a broad class of commonly used interpolants [Albergo et al., 2025]: the linear interpolant ($\alpha_t = 1 - t$ and $\beta_t = t$), the diffusion interpolant ($\alpha_t = \sqrt{1-t^2}$ and $\beta_t = t$), and the trigonometric interpolant ($\alpha_t = \cos(\pi t/2)$ and $\beta_t = \sin(\pi t/2)$). Our condition (iii) is stronger than the temporal regularity condition in Ding et al. [2024]. The additional derivative is used in the second-order discretization analysis.

3.2 Flow map approximation

In this section we introduce a class of neural networks that will be used to approximate the flow map. We then show that, within this class, the flow map can be approximated with a rate that does not suffer from the curse of dimensionality.

Before presenting the approximation result, we first derive an explicit representation of the true velocity field, which is stated in the following proposition.

Proposition 3.3. *Suppose that Assumptions 3.1 and 3.2 hold. For $t \in [0,1]$, set $\widetilde{\alpha}_t^2 = \alpha_t^2 + \beta_t^2 \sigma_{\mathrm{data}}^2$. Then, the true velocity field admits the representation*

$$b^*(y,t) = c_1(t, \sigma_{\mathrm{data}})y + c_2(t, \sigma_{\mathrm{data}}) f^*(y,t), \quad (y,t) \in \mathbb{R}^D \times [0,1],$$

where

$$c_1(t,\sigma) = \frac{\dot{\alpha}_t \alpha_t + \dot{\beta}_t \beta_t \sigma^2}{\alpha_t^2 + \beta_t^2 \sigma^2}, \quad c_2(t,\sigma) = \frac{\alpha_t(\dot{\beta}_t \alpha_t - \beta_t \dot{\alpha}_t)}{\alpha_t^2 + \beta_t^2 \sigma^2}, \quad \sigma \in (0,1],\ t \in [0,1].$$

Furthermore,

$$f^*(y,t) = \frac{\int\limits_{[0,1]^d} g^*(u) \exp\left(-\frac{\|y - \beta_t g^*(u)\|^2}{2\widetilde{\alpha}_t^2}\right) \mathrm{d}u}{\int\limits_{[0,1]^d} \exp\left(-\frac{\|y - \beta_t g^*(u)\|^2}{2\widetilde{\alpha}_t^2}\right) \mathrm{d}u}, \quad (y,t) \in \mathbb{R}^D \times [0,1]. \tag{10}$$

The proof of Proposition 3.3 is given in Appendix A.2. The proposition reveals that the true velocity field is a sum of a linear and a bounded term. This representation, already exploited in the analysis of diffusion models [Yakovlev and Puchkin, 2025], is the key insight that allows us to design the class of average velocities. Specifically, we approximate the bounded component of the velocity field using a cubic B-spline quasi-interpolant. The formal definition of this quasi-interpolant is detailed below.

Definition 3.4 (B-spline quasi-interpolant)**.** *For a fixed $t_0 \in [0,1)$, $N \in \mathbb{N}$, and a set of points $\{X_k\}_{k=0}^N \subseteq \mathbb{R}^D$, let the coefficients $\{\alpha_k\}_{k=-1}^{N+1}$ be defined as $\alpha_k = X_k$ for $k \in \{0,1,\dots,N\}$, with boundary values $\alpha_{-1} = 2X_0 - X_1$ and $\alpha_{N+1} = 2X_N - X_{N-1}$. The cubic B-spline quasi-interpolant $I : [0, 1-t_0] \to \mathbb{R}^D$ is defined by*

$$I(t; X_0, X_1, \dots, X_N) = \sum_{k=-1}^{N+1} \alpha_k B_3(h^{-1}t - (k-2)), \quad t \in [0, 1-t_0],$$

where B_3 is the cubic B-spline basis function supported on $[0,4]$ and $h = (1-t_0)/N$, with grid points $\tau_k = kh$.

Cubic B-splines are selected specifically for their temporal regularity, which is essential for an accurate approximation of the consistency penalty $R(v)$. We can now define the class of average velocities.

Definition 3.5 (average velocities class)**.** *Let $N \in \mathbb{N}$, $t_0 \in [0,1)$, and $L_f > 0$ be given. Suppose that $\{f_k : 0 \leqslant k \leqslant N\} \subseteq \mathsf{NN}(L, W, S, B)$ are L_f-Lipschitz functions satisfying*

$$(i) \quad \|f_k\|_{L^\infty(\mathbb{R}^D)} \lesssim \sqrt{D}, \qquad (ii) \quad \|\nabla f_k(y)\| \lesssim \sigma_{\min}^{-2}, \quad \textit{for all } \|y\| \leqslant R.$$

Set $h = (1-t_0)/N$ and $\tau_k = hk$ for $k \in \{0,1,\dots,N\}$. For any $y \in \mathbb{R}^D$ and $\sigma \in [\sigma_{\min}, 1]$, put

$$b_k(z) = c_1(\tau_k, \sigma) z + c_2(\tau_k, \sigma) f_k(z), \quad 0 \leqslant k \leqslant N.$$

Define the Heun iterates by

$$Y_{k+1} = Z_k + h b_k(Z_k), \quad Z_{k+1} = Z_k + \frac{h}{2}\left(b_k(Z_k) + b_{k+1}(Y_{k+1})\right), \quad Z_0 = y.$$

Here, the functions c_1 and c_2 are defined in Proposition 3.3. For all $(y,t) \in \mathbb{R}^D \times [0, 1-t_0]$ we define the velocity field

$$v_{t,t}(y) = c_1(t,\sigma) y + c_2(t,\sigma) f(y,t), \quad f(y,t) = I(t; f_0(y), f_1(y), \dots, f_N(y)).$$

Similarly, for all $t \in (0, 1 - t_0]$ and $y \in \mathbb{R}^D$, we define the average velocity as

$$v_{0,t}(y) = t^{-1} \left(I(t; Z_0, Z_1, \ldots, Z_N) - y \right).$$

Here, $I(t; \cdot)$ denotes the cubic B-spline quasi-interpolant from Definition 3.4. We denote the resulting class of average velocity candidates by $\mathcal{V}(L, W, S, B, N, R, L_f)$.

Having formally defined the class of average velocities, we now discuss the motivation behind our architectural choices.

The representation of the flow map via the average velocity in (4) motivates the use of residual neural networks [He et al., 2016] for $v_{0,t}$ and $v_{t,t}$ (Definition 3.5), rather than the standard feedforward architectures. Residual architectures have proven effective for approximating trajectories of ODEs [Baggenstos and Salimova, 2023] and possess favorable approximation complexity bounds [Müller, 2019, Oono and Suzuki, 2019, Liu et al., 2024]. Moreover, the iterative construction in Definition 3.5 explicitly implements Heun's method, a second-order Runge-Kutta discretization scheme for ODEs.

Although higher-order Runge-Kutta methods could potentially refine the approximation, their analysis is considerably more involved. Specifically, analyzing higher-order schemes requires determining not only the order of accuracy but also the precise dependence on the ambient dimension D, the stopping time t_0, and the noise parameter $\sigma_{\min}$. We find that a second-order method offers a good balance between numerical accuracy and theoretical tractability for our purposes. Similarly, we use cubic B-splines because they provide the temporal regularity needed to control the consistency penalty $R(v)$ while keeping the analysis tractable. Although B-splines are not, to our knowledge, explicitly used in existing architectures, we introduce them solely to guarantee the temporal smoothness of the flow-map estimator, which is essential for our estimation error analysis. Moreover, cubic B-splines can themselves be approximated by neural networks, so the approximation class in Definition 3.5 can be realized by a ResNet architecture. We keep the B-spline formulation because it makes the approximation arguments substantially more transparent.

It is important to note that the neural networks in Definition 3.5 are required to be Lipschitz continuous. This property is essential for the subsequent estimation analysis. In practice, this constraint can be enforced using spectral normalization [Miyato et al., 2018] or Lipschitz neural-network architectures [Anil et al., 2019, Qi et al., 2023]. This requirement is consistent with other recent theoretical studies, including the work of Ding et al. [2024], which also emphasizes the importance of Lipschitz continuity.

We now present one of our main results, which establishes that the class of average velocities, as defined in Definition 3.5, is sufficiently rich to approximate the target flow map without the curse of dimensionality.

Theorem 3.6 (approximation of the flow map)**.** *Assume Assumptions 3.1 and 3.2 hold. Fix $t_0 \in [0, 1/2]$. Then, for any $\varepsilon \in (0, 1/2]$ satisfying $\varepsilon^\beta \leqslant c\sigma_{\min}^4$ and*

$$D^{3/2} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(\varepsilon^\beta \vee \varepsilon) \left(\log(1/\varepsilon) + \log(D\sigma_{\min}^{-2}) \right) \lesssim 1,$$

there exists an average velocity candidate $\widetilde{v}$ as in Definition 3.5 with the parameters

$$N \asymp \varepsilon^{-\beta/2}\sigma_{\min}^{-2}, \quad L_f \asymp (\sigma_{\min}^{-2} D \log(1/\varepsilon))^{\mathcal{O}(1)}, \quad R \asymp \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} D \log(\varepsilon^{-1} D \sigma_{\min}^{-2})$$

such that

$$\mathbb{E}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] \lesssim (D \log(1/\varepsilon))^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\varepsilon^\beta.$$

Moreover, the architecture of $\widetilde{v}$ is as follows:

$$L \lesssim \sigma_{\min}^{-2} \log(D\sigma_{\min}^{-2} \log(1/\varepsilon)), \quad \|W\|_\infty \vee S \lesssim \varepsilon^{-d} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D \log(1/\varepsilon))^{\mathcal{O}(1)},$$
$$\log B \lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D \log(1/\varepsilon))^{\mathcal{O}(1)}$$

The hidden constants behind $\mathcal{O}(\cdot)$ may depend only on d, β, and H.

The proof of Theorem 3.6 is deferred to Appendix B. The proof uses the relation between the score function $s^*(y, t) = \nabla \log \mathsf{p}_t(y)$ and the velocity field $b^*(y, t)$. Specifically, $b^*(y, t)$ can be expressed as a linear combination of y and $s^*(y, t)$ (see, for instance, [Ding et al., 2024, Proposition 3.1] and

[Lipman et al., 2024, Section 10]). Building on this relation, we utilize the simultaneous approximation of the score function and its derivatives [Yakovlev and Puchkin, 2026]. This ensures that the resulting average velocity approximation is spatially Lipschitz, a property that is essential for the estimation analysis below.

Theorem 3.6 establishes that the average velocity can be approximated to accuracy $\widetilde{\mathcal{O}}(\varepsilon^{\beta})$ by the residual neural network architecture defined in Definition 3.5, using only $\widetilde{\mathcal{O}}(\varepsilon^{-d-\beta/2})$ non-zero parameters. Notably, the dependence on the ambient dimension D is polynomial, ensuring that the bound remains informative even in high-dimensional regimes where $D \gtrsim \log n$. Although the bounds scale exponentially with $\sigma_{\min}^{-2}$, relaxing this dependence to a polynomial one is an important direction for future work.

Under similar assumptions on the data distribution, Yakovlev and Puchkin [2025, 2026] derived approximation bounds for the score function $s^*(y,t) = \nabla \log \mathsf{p}_t(y)$ that avoid the curse of dimensionality. Specifically, achieving an accuracy of $\widetilde{\mathcal{O}}(\varepsilon^{2\beta})$ for the score function requires $\widetilde{\mathcal{O}}(\varepsilon^{-d})$ non-zero parameters, which is slightly lower than the complexity presented in Theorem 3.6. We attribute this gap to the fact that approximating the score function corresponds to estimating the instantaneous velocity, whereas estimating the average velocity is inherently more challenging, as it involves the analysis of discretization error in the underlying ODE.

Compared with Ding et al. [2024], we observe that our approach avoids the curse of dimensionality under more realistic assumptions. More specifically, by setting $\beta \geqslant 1$ and assuming $g^*(u) = Au$ for some $A \in \mathbb{R}^{D\times d}$ with orthonormal columns, we recover Assumption 3 of Ding et al. [2024], which describes the target distribution as the convolution of a distribution supported on a low-dimensional linear subspace with Gaussian noise. In their Lemma H.7, they achieved $\mathcal{O}(\varepsilon^2)$ approximation accuracy for their semi-group-penalized risk using $\widetilde{\mathcal{O}}(\varepsilon^{-d-2})$ parameters. In contrast, Theorem 3.6 provides similar approximation guarantees under more realistic assumptions by treating the low-dimensional structure of the data distribution more generally.

Finally, we expect that similar approximation guarantees can be established for other one-step flow-based objectives such as Mean Flow [Geng et al., 2025], shortcut loss [Frans et al., 2025], and Alpha Flow loss [Zhang et al., 2026]. Crucially, the framework in Theorem 3.6 is not specific to self-distillation and is therefore of independent interest beyond the present setting.

3.3 Flow map estimation

In this section, we establish the sample complexity of learning the flow map via the self-distillation objective given by (6). The following theorem constitutes one of the central results of this paper. It demonstrates that the flow map can be estimated without the curse of dimensionality.

Theorem 3.7. *Assume Assumptions 3.1 and 3.2 hold. Fix $t_0 \in [0, 1/2]$. Assume further that the sample size is sufficiently large in the sense that it satisfies*

$$D^{3/2} e^{\mathcal{O}(\sigma_{\min}^{-2})} \log(nD\sigma_{\min}^{-2}) \left(\frac{\lambda(t_0)}{n}\right)^{\frac{2(\beta\wedge 1)}{5\beta+2d}} \lesssim 1, \quad \left(\frac{\lambda(t_0)}{n}\right)^{\frac{2\beta}{5\beta+2d}} \lesssim \sigma_{\min}^4, \quad n \gtrsim \lambda(t_0),$$

where $\lambda(t_0) = |\alpha_t|_{W^{1,\infty}([0,1-t_0])} \vee e$. Then, with probability at least $(1-\delta)$, the empirical risk minimizer (7) over the class $\mathcal{V}(L, W, S, B, N, R, L_f)$ given by Definition 3.5 with the architecture

$$L \lesssim \sigma_{\min}^{-2} \log(D\sigma_{\min}^{-2} \log n), \quad \|W\|_\infty \vee S \lesssim \left(\frac{n}{\lambda(t_0)}\right)^{\frac{2d}{5\beta+2d}} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D\log n)^{\mathcal{O}(1)},$$

$$\log B \lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D\log n)^{\mathcal{O}(1)}$$

and the parameters $L_f \asymp (\sigma_{\min}^{-2} D \log n)^{\mathcal{O}(1)}$, $N \asymp (n/\lambda(t_0))^{\beta/(5\beta+2d)} \sigma_{\min}^{-2}$, and $R \asymp \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} D \log(nD\sigma_{\min}^{-2})$ satisfies the inequality

$$\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] \lesssim (D\, e^{\sigma_{\min}^{-4}} \log n)^{\mathcal{O}(1)} \lambda(t_0) \left(\frac{\lambda(t_0)}{n}\right)^{\frac{2\beta}{5\beta+2d}} \log(\lambda(t_0)) \log(4/\delta).$$

Here, the hidden constant may depend on d, β, and H only.

The proof of Theorem 3.7 is provided in Appendix C. The proof uses an ϵ-net argument in conjunction with the Bernstein-type inequality for unbounded random variables [Lecué and Mitchell, 2012]. Furthermore, we establish that the loss landscape is well-behaved in the sense that it satisfies the Bernstein condition, which enables sharp convergence rates.

Theorem 3.7 establishes an excess risk rate of $\widetilde{\mathcal{O}}(n^{-2\beta/(5\beta+2d)})$, with a polynomial dependence on the ambient dimension D and $\lambda(t_0)$. Under similar data distribution assumptions, this behaviour is consistent with results from the score estimation literature [Chen et al., 2023, Yakovlev and Puchkin, 2025]. To the best of our knowledge, Theorem 3.7 is the first result to provide convergence rates for flow map estimation that do not suffer from the curse of dimensionality under reasonable assumptions on the low-dimensional structure of the data distribution. The rate exhibits exponential dependence on $\sigma_{\min}^{-4}$, consistent with Theorem 3.6. Whether this bound can be improved remains an important open problem for future work.

Finally, we observe that the rates in the score estimation literature [Yakovlev and Puchkin, 2025, Yakovlev et al., 2025], under similar assumptions on the data distribution, scale as $\widetilde{\mathcal{O}}(n^{-2\beta/(2\beta+d)})$, which decays faster than the rate in Theorem 3.7. As discussed following Theorem 3.6, this discrepancy arises because flow map estimation is a more challenging task, requiring a more involved approximation error analysis.

3.4 Distribution estimation

In this section, we analyze the estimation of the target distribution with respect to the 2-Wasserstein distance. The following theorem demonstrates that the one-step generator, learned via the self-distillation objective (6), recovers the target distribution without suffering from the curse of dimensionality.

Theorem 3.8 (distribution estimation)**.** *Let $\mathcal{V}(L, W, S, B, N, R, L_f)$ be the class defined in Theorem 3.7, and let $\widehat{\Psi}_{0,t}(y) = y + t \cdot \widehat{v}_{0,t}(y)$ be a flow map, where $\widehat{v}$ is the empirical risk minimizer from (7). Define $\widehat{\mathsf{p}}_{1-t_0} = \mathrm{Law}(\widehat{\Psi}_{0,1-t_0}(X_0))$. Then, with probability at least $(1-\delta)$, it holds that*

$$W_2\left(\widehat{\mathsf{p}}_{1-t_0}, \mathsf{p}^*\right) \lesssim D^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\min}^{-4})\} \sqrt{\lambda(t_0)} \left(\frac{\lambda(t_0)}{n}\right)^{\frac{\beta}{5\beta+2d}} \mathscr{L}(t_0, n, \delta) + \sqrt{D}(\alpha_{1-t_0} \vee (1-\beta_{1-t_0})),$$

where $\mathscr{L}(t_0, n, \delta) = (\log n)^{\mathcal{O}(1)} \sqrt{\log(\lambda(t_0)) \log(4/\delta)}$. The hidden constant behind $\mathcal{O}(\cdot)$ depends only on d, β, and H.

The proof of Theorem 3.8 is moved to Appendix D. The primary technical challenge is that the standard analysis based on the Alekseev-Gröbner lemma [Benton et al., 2024, Fukumizu et al., 2025] requires the velocity field estimate to be globally Lipschitz, with the W_2 distribution error scaling exponentially with this Lipschitz constant. A direct application of these results would not yield a consistent estimate, as the global Lipschitz constant guaranteed by Theorem 3.7 grows at least logarithmically with the sample size n. However, the local Lipschitz constant provided in Theorem 3.7 is well-behaved. By extending the standard analysis, we demonstrate that local Lipschitzness is sufficient to obtain a sharp convergence rate.

Theorem 3.8 has several important implications for different choices of the interpolant coefficients α_t and β_t, as detailed in the following remark.

Remark 3.9. *Under the assumptions of Theorem 3.8, consider the following two cases:*

1. ***Linear Interpolant:*** *For the linear interpolant ($\alpha_t = 1 - t$ and $\beta_t = t$), we have $\lambda(t_0) \asymp 1$ for all $t_0 \in [0, 1)$. In this case, for $t_0 = 0$ and all sufficiently large n, the 2-Wasserstein distance scales as $W_2(\widehat{\mathsf{p}}_1, \mathsf{p}^*) = \widetilde{\mathcal{O}}(n^{-\beta/(5\beta+2d)})$.*

2. ***Diffusion Interpolant:*** *For the diffusion interpolant ($\alpha_t = \sqrt{1-t^2}$ and $\beta_t = t$), we have $\lambda(t_0) \lesssim t_0^{-1/2}$ and $\alpha_{1-t_0} \vee (1-\beta_{1-t_0}) \lesssim \sqrt{t_0}$. Therefore, setting $t_0 = n^{-4\beta/(17\beta+6d)}$ for all sufficiently large n, we obtain a convergence rate of $W_2(\widehat{\mathsf{p}}_{1-t_0}, \mathsf{p}^*) = \widetilde{\mathcal{O}}(n^{-2\beta/(17\beta+6d)})$.*

In both cases the dependence on the ambient dimension D is polynomial, which is crucial in high-dimensional settings.

Building on Theorem 3.8 and Remark 3.9, we compare our findings with the results presented in Ding et al. [2024]. In Corollary 3.16 of their work (convergence of linear interpolant under low-dimensional hypothesis), the authors obtain a rate for the squared W_2 distance of $\widetilde{\mathcal{O}}(n^{-4/(3d+9)})$. Consequently, if $\beta > 4d/(3d-1)$, our approach achieves a faster rate of convergence. Note that such a choice of β is admissible in the linear case, where g^* is infinitely smooth. Furthermore, our guarantees hold for a significantly broader class of data distributions, extending beyond linear subspaces. To the best of our knowledge, Theorem 3.8 is the first such result in the literature on one-step flow-based and diffusion-based generative models. Whether these distribution estimation bounds are optimal remains an important open question for future research.

Finally, we discuss the exponential dependence on $\sigma_{\min}^{-4}$ in Theorems 3.7 and 3.8. To our knowledge, this dependence can be mitigated under substantially stronger assumptions on the data distribution. For example, if the target distribution p^* is weakly log-concave, the result of [Stéphanovitch, 2026, Theorem 3] suggests that the factor $\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}$ can be replaced by a constant. However, that result does not apply under Assumption 3.1, and extending it to our setting remains an open problem. More broadly, obtaining simultaneously polynomial dependence on the ambient dimension D, optimal statistical rates in the sample size n, and only mild dependence on $\sigma_{\min}$ appears to be a fundamental theoretical challenge. Our analysis establishes favourable dependence on D and n. We view this as a meaningful step toward understanding the statistical properties of one-step flow-based models.

4 Conclusion

In this work, we developed a theoretical framework for the estimation of flow maps in one-step generative models. We derived approximation and estimation bounds that avoid the curse of dimensionality, assuming that the target distribution is the convolution of a distribution supported on a low-dimensional surface with Gaussian noise. These findings provide a first step toward explaining the empirical success of one-step flow-based generative models. Future work will focus on further refining the derived bounds, particularly by exploring the use of higher-order B-splines and more sophisticated Runge-Kutta integration schemes.

Acknowledgments and Disclosure of Funding

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University №139-15-2025-009.

References

M. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL `http://jmlr.org/papers/v26/23-1605.html`.

C. Anil, J. Lucas, and R. Grosse. Sorting out Lipschitz function approximation. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/anil19a.html`.

I. Azangulov, G. Deligiannidis, and J. Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

J. Baggenstos and D. Salimova. Approximation properties of residual neural networks for Kolmogorov PDEs. *Discrete and Continuous Dynamical Systems - B*, 28(5):3193–3215, 2023. ISSN 1531-3492. doi: 10.3934/dcdsb.2022210. URL `https://www.aimsciences.org/article/id/6359d0876aa93c57bccf00a9`.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

J. Benton, G. Deligiannidis, and A. Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=uqQPyWFDhY.

S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

N. M. Boffi, M. S. Albergo, and E. Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=cqDH0e6ak2.

N. M. Boffi, M. S. Albergo, and E. Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

B. C. Brown, A. L. Caterini, B. L. Ross, J. C. Cresswell, and G. Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2023.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

S. Chakraborty and P. L. Bartlett. On the statistical properties of generative adversarial models for low intrinsic data dimension. *Journal of Machine Learning Research*, 26(111):1–57, 2025.

M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023.

Y. Chen, Y. Zhang, O. Oertell, and W. Sun. Convergence of consistency model with multistep sampling under general data assumptions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vsJsR3ieCx.

B. Dahal, A. Havrilla, M. Chen, T. Zhao, and W. Liao. On deep generative models for approximation and estimation of distributions on manifolds. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10615–10628. Curran Associates, Inc., 2022.

Z. Ding, C. Duan, Y. Jiao, R. Li, J. Z. Yang, and P. Zhang. Characteristic learning for provable one step generation. Preprint, ArXiv:2405.05512, 2024.

Z. Dou, M. Chen, M. Wang, and Z. Yang. Theory of consistency diffusion models: Distribution estimation meets fast sampling. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=pAPykbqUHf.

P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

K. Frans, D. Hafner, S. Levine, and P. Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

K. Fukumizu, T. Suzuki, N. Isobe, K. Oko, and M. Koyama. Flow matching achieves almost minimax optimal convergence. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2OMyAFjiJJ.

Z. Geng, M. Deng, X. Bai, J. Z. Kolter, and K. He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uWj4s7rMnR.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

M. Khayatkhoei, M. K. Singh, and A. Elgammal. Disconnected manifold learning for generative adversarial networks. *Advances in Neural Information Processing Systems*, 31, 2018.

D. Kim, C.-H. Lai, W.-H. Liao, N. Murata, Y. Takida, T. Uesaka, Y. He, Y. Mitsufuji, and S. Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ymjI8feDTD.

G. Lecué and C. Mitchell. Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, 6(none):1803 – 1837, 2012. doi: 10.1214/12-EJS730. URL https://doi.org/10.1214/12-EJS730.

G. Li, Z. Huang, and Y. Wei. Towards a mathematical theory for consistency training in diffusion models. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=SVkYxDYnJK.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Y. Lipman, M. Havasi, P. Holderrieth, N. Shaul, M. Le, B. Karrer, R. T. Chen, D. Lopez-Paz, H. Ben-Hamu, and I. Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

C. Liu, E. Liang, and M. Chen. Characterizing ResNet's universal approximation capability. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 31477–31515. PMLR, 21–27 Jul 2024.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

C. Lu and Y. Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LyJi5ugyJx.

Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7402–7406. IEEE, 2022.

J. Lyu, Z. Chen, and S. Feng. Sampling is as easy as keeping the consistency: convergence guarantee for consistency models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ZPiEIhQpos.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

J. Müller. On the space-time expressivity of resnets. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2019. URL https://openreview.net/forum?id=dYH-JNPNGI.

K. Oko, S. Akiyama, and T. Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

K. Oono and T. Suzuki. Approximation and non-parametric estimation of ResNet-type convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4922–4931. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/oono19a.html.

W. Peebles and S. Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182. IEEE, 2023.

G. M. Phillips. *Interpolation and approximation by polynomials*, volume 14. Springer Science & Business Media, 2003.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

X. Qi, J. Wang, Y. Chen, Y. Shi, and L. Zhang. LipsFormer: Introducing Lipschitz Continuity to Vision Transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cHf1DcCwcH3.

Y. Song and P. Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

A. Stéphanovitch. Lipschitz regularity in flow matching and diffusion models: sharp sampling rates and functional inequalities. Preprint, ArXiv:2604.06065, 2026.

A. Stephanovitch, E. Aamari, and C. Levrard. Wasserstein generative adversarial networks are minimax optimal distribution estimators. *The Annals of Statistics*, 52(5):2167–2193, 2024.

R. Tang and Y. Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR, 2024.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

A. Wibisono, Y. Wu, and K. Y. Yang. Optimal score estimation via empirical bayes smoothing. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4958–4991. PMLR, 2024.

K. Yakovlev and N. Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5824–5891. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/yakovlev25a.html.

K. Yakovlev and N. Puchkin. Simultaneous approximation of the score function and its derivatives by deep neural networks. *Journal of Complexity*, page 102070, 2026.

K. Yakovlev, A. Markovich, and N. Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density Hessian estimation. Preprint. ArXiv:2512.24378, 2025.

R. Yang, B. Jiang, C. Chen, and S. Li. Improved discretization complexity analysis of consistency models: Variance exploding forward process and decay discretization scheme. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=CZTcRSxkfe.

H. Zhang, A. Siarohin, W. Menapace, M. Vasilkovsky, S. Tulyakov, Q. Qu, and I. Skorokhodov. Alphaflow: Understanding and improving meanflow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=adacb4JTIv.

K. Zhang, H. Yin, F. Liang, and J. Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Forty-first International Conference on Machine Learning*, 2024.

A Proofs of supporting propositions

This section collects the proofs of Propositions 2.1 and 3.3.

A.1 Proof of Proposition 2.1

From the definition of the flow matching loss defined in (2) we obtain that

$$\mathbb{E}[\ell_{FM}(b, X_1) - \ell_{FM}(b^*, X_1)] = \int_0^{1-t_0} \mathbb{E}\|b(X_t, t) - b^*(X_t, t)\|^2 \, \mathrm{d}t$$
$$+ 2 \int_0^{1-t_0} \mathbb{E}\langle b(X_t, t) - b^*(X_t, t), b^*(X_t, t) - \dot{X}_t \rangle \, \mathrm{d}t.$$

Taking into account that $b^*(y, t) = \mathbb{E}[\dot{X}_t \,|\, X_t = y]$ and using the tower property of the conditional expectation, we conclude that

$$\mathbb{E}\langle b(X_t, t) - b^*(X_t, t), b^*(X_t, t) - \dot{X}_t \rangle = \mathbb{E}\left[\langle b(X_t, t) - b^*(X_t, t), b^*(X_t, t) - \mathbb{E}[\dot{X}_t \,|\, X_t] \rangle\right] = 0.$$

Consequently, the claim readily follows.

□

A.2 Proof of Proposition 3.3

Recall that $b^*(y, t) = \mathbb{E}[\dot{X}_t \,|\, X_t = y]$ for $t < 1$. Therefore, it holds that

$$b^*(y, t) = \mathbb{E}[\mathbb{E}[\dot{\alpha}_t X_0 + \dot{\beta}_t (g^*(u) + \sigma_{\mathrm{data}} Z) \,|\, X_t = y, U = u] \,|\, X_t = y].$$

Next, observe that for a fixed $U = u$, it holds that

$$(X_0, Z, X_t)^\top \,|\, (U = u) \sim \mathcal{N}\left(\begin{pmatrix} 0_D \\ 0_D \\ \beta_t g^*(u) \end{pmatrix}, \begin{pmatrix} I_D & 0_{D\times D} & \alpha_t I_D \\ 0_{D\times D} & I_D & \beta_t \sigma_{\mathrm{data}} I_D \\ \alpha_t I_D & \beta_t \sigma_{\mathrm{data}} I_D & \widetilde{\alpha}_t^2 I_D \end{pmatrix} \right).$$

Therefore, using the conditional law of a Gaussian vector, we deduce that

$$\mathbb{E}[X_0 \,|\, X_t = y,\, U = u] = \frac{\alpha_t(y - \beta_t g^*(u))}{\widetilde{\alpha}_t^2}, \quad \mathbb{E}[Z \,|\, X_t = y, U = u] = \frac{\beta_t \sigma_{\mathrm{data}}(y - \beta_t g^*(u))}{\widetilde{\alpha}_t^2}.$$

Hence, we arrive at

$$b^*(y,t) = \mathbb{E}\left[\frac{(\alpha_t\dot{\alpha}_t + \beta_t\dot{\beta}_t\sigma_{\mathrm{data}}^2)y}{\widetilde{\alpha}_t^2} + \left(\dot{\beta}_t - \frac{\beta_t(\alpha_t\dot{\alpha}_t + \beta_t\dot{\beta}_t\sigma_{\mathrm{data}}^2)}{\widetilde{\alpha}_t^2}\right) g^*(U) \,\middle|\, X_t = y\right]. \quad (11)$$

We next notice that $X_t|(U=u) \sim \mathcal{N}(\beta_t g^*(u), \widetilde{\alpha}_t^2 I_D)$ and, hence, $\mathbb{E}[g^*(U) \,|\, X_t = y] = f^*(y,t)$, where $f^*(y,t)$ is defined in (10). Substituting this identity into (11) yields the desired result.

□

B Proof of Theorem 3.6

The proof constructs $v \in \mathcal{V}(L, W, S, B, N, R, L_f)$ by controlling both terms of $\ell_{SD}(v,\cdot) - \ell_{SD}(v^*,\cdot) = \ell_{FM}(v,\cdot) - \ell_{FM}(v^*,\cdot) + R(v)$. The flow matching term requires a uniform in time approximation of the velocity field b^*, while $R(v)$ also requires temporal smoothness and an accurate trajectory. An approximation for a fixed time alone is therefore insufficient. We first approximate $b^*(\cdot, \tau_k)$ at each grid time by GELU networks, then interpolate them with cubic B-splines to obtain a uniform in time bound and the temporal regularity needed for $R(v)$. Next, Heun's method and a second B-spline quasi-interpolation yield the average velocity and the flow map, with controlled L^2 and derivative errors. Combining these estimates yields the desired approximation error bound.

For clarity, we split the proof of Theorem 3.6 into several steps.

Approximation of the velocity field. We begin by introducing a uniform grid $\tau_k = kh$ with $k \in \{0, \dots, N\}$, where the step size $h = (1 - t_0)/N$. Here, $N \in \mathbb{N}$ will be specified later in the proof. Our first objective is to construct an approximation of the velocity field at each grid point $\tau_k \in [0, 1 - t_0]$, a result established in the following lemma.

Lemma B.1 (approximation of the velocity field)**.** *Suppose Assumptions 3.1 and 3.2 hold, and fix $t \in [0, 1 - t_0]$. Assume further that $\varepsilon \in (0, 1/2]$ satisfies the condition*

$$D^{3/2} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(\varepsilon^\beta \vee \varepsilon)\left(\log(1/\varepsilon) + \log(D\sigma_{\min}^{-2})\right) \lesssim 1. \quad (12)$$

Then, there exists a GELU neural network $\bar{f} \in \mathsf{NN}(L, W, S, B)$ such that the corresponding velocity field approximation, defined by

$$\bar{b}(y) = c_1(t, \sigma_{\mathrm{data}})y + c_2(t, \sigma_{\mathrm{data}})\bar{f}(y), \quad y \in \mathbb{R}^D,$$

satisfies the following properties:

$$(i) \quad \max_{0 \leqslant s \leqslant 1 - t_0} \|\bar{f} - f^*(\cdot, t)\|^2_{L^2(\mathsf{p}_s^*)} \vee \|\bar{b} - b^*(\cdot, t)\|^2_{L^2(\mathsf{p}_s^*)} \lesssim \left(De^{\sigma_{\min}^{-2}} \log(1/\varepsilon)\right)^{\mathcal{O}(1)} \varepsilon^{2\beta},$$

$$(ii) \quad \sup_{y \in \mathcal{K}_t} \|\nabla \bar{f}(y)\| \lesssim \sigma_{\mathrm{data}}^{-2},\ |\bar{f}|_{L^\infty(\mathbb{R}^D)} \lesssim \sqrt{D},\ \max_{1 \leqslant l \leqslant D} |\bar{f}_l|_{W^{1,\infty}(\mathbb{R}^D)} \lesssim \left(\sigma_{\min}^{-2} D \log(1/\varepsilon)\right)^{\mathcal{O}(1)}.$$

Here, $\mathcal{K}_t$ is defined as

$$\mathcal{K}_t = \left\{ y \in \mathbb{R}^D \,:\, \min_{u \in [0,1]^d} \|y - \beta_t g^*(u)\| \leqslant R \right\},$$

with $R \asymp \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} D \left(\log(1/\varepsilon) + \log(D\sigma_{\min}^{-2})\right)$. Moreover, the architecture of $\bar{f}$ is given by

$$L \lesssim \sigma_{\min}^{-2} \log(D\sigma_{\min}^{-2} \log(1/\varepsilon)), \quad \|W\|_\infty \vee S \lesssim \varepsilon^{-d} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D \log(1/\varepsilon))^{\mathcal{O}(1)},$$
$$\log B \lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D \log(1/\varepsilon))^{\mathcal{O}(1)}$$

The proof of Lemma B.1 is provided in Appendix B.1 and builds on Yakovlev and Puchkin [2026]. We choose a substantially larger radius R than in Yakovlev and Puchkin [2026] because our analysis requires the velocity field approximation to be Lipschitz on a ball whose radius grows as

$\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}$. Now, we analyze the derivatives of the constructed velocity field approximation. Let

$$\bar{b}(y,\tau_k) = c_1(\tau_k,\sigma_{\text{data}})y + c_2(\tau_k,\sigma_{\text{data}})\bar{f}(y,\tau_k)$$

be the velocity field approximation provided by Lemma B.1 for each $k \in \{0,1,\dots,N\}$ and $\varepsilon \in (0,1)$ satisfying (12). Apply Lemma B.1 with a tube radius at least twice the class radius below. Introducing $\widetilde{f}_k(y) = \bar{f}(y,\tau_k)$ and invoking claim (ii), we can set

$$R \asymp \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} D \left(\log(1/\varepsilon) + \log(D\sigma_{\min}^{-2})\right) \tag{13}$$

in Definition 3.5. Therefore, for every $k \in \{0,\dots,N\}$, the function $\widetilde{f}_k$ is Lipschitz on $\mathcal{B}(0,R)$. The Lipschitz constant scales as $\sigma_{\text{data}}^{-2}$. Furthermore, $\widetilde{f}_k$ is globally L_f-Lipschitz, where L_f is given by

$$L_f \asymp (\sigma_{\min}^{-2} D \log(1/\varepsilon))^{\mathcal{O}(1)}.$$

Next, observe that, by Assumption 3.2, the functions c_1 and c_2, defined in Proposition 3.3, satisfy

$$|c_j(\cdot,\sigma_{\text{data}})|_{W^{k,\infty}([0,1-t_0])} \lesssim \sigma_{\text{data}}^{-2-2k}, \quad j \in \{1,2\},\ k \in \{0,1,2\}. \tag{14}$$

Consequently, for all $y \in \mathbb{R}^D$, it follows that

$$\begin{aligned}\|\bar{b}(y,\tau_k)\| &\leqslant |c_1(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|y\| + |c_2(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|\widetilde{f}_k(y)\| \\ &\lesssim \sigma_{\text{data}}^{-2}\|y\| + \sigma_{\text{data}}^{-2}\sqrt{D}.\end{aligned}$$

Furthermore, for all $y \in \mathcal{K}_{\tau_k}$, we have that

$$\|\nabla\bar{b}(y,\tau_k)\| \leqslant |c_1(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])} + |c_2(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|\nabla\widetilde{f}_k(y)\| \lesssim \sigma_{\text{data}}^{-4}.$$

We now establish bounds on the derivatives of f^*, as defined in (10), which are provided in the following lemma.

Lemma B.2. *Suppose Assumptions 3.1 and 3.2 hold, and let f^* be defined as in* (10)*. Then, for all $y \in \mathbb{R}^D$ and $t \in [0, 1-t_0]$, the following bounds hold:*

$$\begin{aligned}
&(i) \quad \|f^*(y,t)\| \leqslant 1, \quad \|\nabla f^*(y,t)\| \leqslant 2\sigma_{\text{data}}^{-2}, \\
&(ii) \quad \|\partial_t f^*(y,t)\| \leqslant 2\|y\| \cdot |\beta_t/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])} + |\beta_t^2/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])}, \\
&(iii) \quad \|\nabla^2 f^*(y,t)\| \leqslant 32\sigma_{\text{data}}^{-4}, \\
&(iv) \quad \|\nabla\partial_t f^*(y,t)\| \leqslant 2|\beta_t/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])} \\
&\qquad + 16\sigma_{\text{data}}^{-2}\left(\|y\|\,|\beta_t/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])} + |\beta_t^2/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])}\right), \\
&(v) \quad \|\partial_{tt}^2 f^*(y,t)\| \leqslant 2|\beta_t/\widetilde{\alpha}_t^2|_{W^{2,\infty}([0,1-t_0])}\|y\| + |\beta_t^2/\widetilde{\alpha}_t^2|_{W^{2,\infty}([0,1-t_0])} \\
&\qquad + 8\left(\|y\| \cdot |\beta_t/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])} + \frac{1}{2}|\beta_t^2/\widetilde{\alpha}_t^2|_{W^{1,\infty}([0,1-t_0])}\right)^2.
\end{aligned}$$

The proof of Lemma B.2 is postponed to Appendix B.2. By Lemma B.2, we have that, for all $y \in \mathbb{R}^D$, the true velocity field $b^*(y,t)$ satisfies

$$\begin{aligned}\|b^*(y,t)\| &\lesssim |c_1(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|y\| + |c_2(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|f^*(y,t)\| \\ &\lesssim \sigma_{\text{data}}^{-2}(1+\|y\|).\end{aligned}$$

Furthermore, we have that

$$\begin{aligned}\|\nabla b^*(y,t)\| &\leqslant |c_1(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])} + |c_2(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|\nabla f^*(y,t)\| \\ &\lesssim \sigma_{\text{data}}^{-4}\end{aligned} \tag{15}$$

and

$$\|\nabla^2 b^*(y,t)\| \leqslant |c_2(\cdot,\sigma_{\text{data}})|_{W^{0,\infty}([0,1-t_0])}\|\nabla^2 f^*(y,t)\| \lesssim \sigma_{\text{data}}^{-6}.$$

Next, Assumption 3.2 ensures that

$$|\beta_t^j/\widetilde{\alpha}_t^2|_{W^{k,\infty}([0,1-t_0])} \lesssim \sigma_{\text{data}}^{-2k-2}, \quad \text{for all } j \in \{1,2\}, \quad k \in \{0,1,2\}. \tag{16}$$

Therefore, we have that

$$\begin{aligned}\|\partial_t b^*(y,t)\| &\leqslant |c_1(\cdot,\sigma_{\mathrm{data}})|_{W^{1,\infty}([0,1-t_0])}\|y\| + |c_2(\cdot,\sigma_{\mathrm{data}})|_{W^{1,\infty}([0,1-t_0])} \\ &\quad + |c_2(\cdot,\sigma_{\mathrm{data}})|_{W^{0,\infty}([0,1-t_0])}\|\partial_t f^*(y,t)\| \\ &\lesssim \sigma_{\mathrm{data}}^{-6}(1+\|y\|).\end{aligned} \tag{17}$$

In addition, it holds that

$$\begin{aligned}\|\nabla\partial_t b^*(y,t)\| &\leqslant |c_1(\cdot,\sigma_{\mathrm{data}})|_{W^{1,\infty}([0,1-t_0])} + |c_2(\cdot,\sigma_{\mathrm{data}})|_{W^{1,\infty}([0,1-t_0])}\|\nabla f^*(y,t)\| \\ &\quad + |c_2(\cdot,\sigma_{\mathrm{data}})|_{W^{0,\infty}([0,1-t_0])}\|\partial_t\nabla f^*(y,t)\| \\ &\lesssim \sigma_{\mathrm{data}}^{-8}(1+\|y\|).\end{aligned}$$

Finally, we find that

$$\begin{aligned}\|\partial_{tt}^2 b^*(y,t)\| &\lesssim |c_1|_{W^{2,\infty}}\|y\| + |c_2|_{W^{2,\infty}} + |c_2|_{W^{1,\infty}}\|\partial_t f^*(y,t)\| + |c_2|_{W^{0,\infty}}\|\partial_{tt}^2 f^*\| \\ &\lesssim \sigma_{\mathrm{data}}^{-10}(1+\|y\|)^2.\end{aligned}$$

Uniform in time approximation of the velocity field. We now aim to extend the approximation of b^* in such a way that it holds uniformly for all $t \in [0, 1-t_0]$. For this purpose, we employ cubic B-spline quasi-interpolants, whose approximation properties are detailed in the following lemma.

Lemma B.3 (approximation properties of B-splines)**.** *Let $f \in C^2([0,1-t_0])$, where $t_0 \in [0,1)$ is arbitrary. For some $K \in \mathbb{N}$ consider a uniform grid $\tau_k = kh$ for all $k \in \{0,1,\dots,K\}$, where $h = (1-t_0)/K$. Define the B-spline quasi-interpolant as*

$$\widetilde{f}(t) = \sum_{k=-1}^{K+1} \alpha_k B_3(h^{-1}t - (k-2)), \quad t \in \mathbb{R},$$

where $\alpha_k = f(\tau_k)$ for every $0 \leqslant k \leqslant K$, and the boundary coefficients are $\alpha_{-1} = 2f(0) - f(\tau_1)$ and $\alpha_{K+1} = 2f(\tau_K) - f(\tau_{K-1})$. Then it holds that

$$\begin{aligned}&(i) \quad |\widetilde{f} - f|_{W^{0,\infty}([0,1-t_0])} \leqslant 4|f|_{W^{2,\infty}([0,1-t_0])}h^2, \\ &(ii) \quad |\widetilde{f} - f|_{W^{1,\infty}([0,1-t_0])} \leqslant 8|f|_{W^{2,\infty}([0,1-t_0])}h.\end{aligned}$$

We move the proof of Lemma B.3 to Appendix B.5. Recall that the quasi-interpolant $\widetilde{f}(t)$ in Lemma B.3 coincides with $I(t; f(0), f(\tau_1), \dots, f(\tau_K))$, where $I(t;\cdot)$ is defined in Definition 3.4. Introduce

$$\widetilde{f}(y,t) = \sum_{k=-1}^{N+1} \alpha_k B_3(h^{-1}t - (k-2)), \quad (t,y) \in [0,1-t_0] \times \mathbb{R}^D,$$

where $\alpha_k = \widetilde{f}_k(y)$ for every $k \in \{0,1,\dots,N\}$, and the boundary coefficients are $\alpha_{-1} = 2\widetilde{f}_0(y) - \widetilde{f}_1(y)$ and $\alpha_{N+1} = 2\widetilde{f}_N(y) - \widetilde{f}_{N-1}(y)$. In the following analysis, we utilize several key properties of B-splines, which are summarized in the following lemma.

Lemma B.4 (properties of uniform B-splines)**.** *Let $\{B_n : n \in \mathbb{Z}_+\}$ be the family of uniform B-splines, defined recursively by*

$$B_n(x) = \frac{x}{n}B_{n-1}(x) + \frac{n+1-x}{n}B_{n-1}(x-1), \quad n \in \mathbb{N}, \quad x \in \mathbb{R},$$

where $B_0(x) = \mathbb{1}[0 < x \leqslant 1]$. Fix an arbitrary $K \in \mathbb{N}$ and $b > 0$, with $h = b/K$. Then, the following properties hold for all $x \in [0,b]$ and $y \in \mathbb{R}$:

1. $0 \leqslant B_n(y) \leqslant 1$;
2. $\mathrm{supp}(B_n) = [0, n+1]$;
3. $\sum_{k=-1}^{K+1} B_n(h^{-1}x - (k-2)) = 1$ *for* $n \in \{2,3\}$;
4. $\sum_{k=-1}^{K+1}(kh - x)B_3(h^{-1}x - (k-2)) = 0$;

5. $B_n'(y) = B_{n-1}(y) - B_{n-1}(y-1)$ *for* $n \geqslant 2$ *at every* y*, and for* $n = 1$ *almost everywhere.*

The proof of Lemma B.4 is deferred to Appendix B.4. Consequently, defining

$$\widetilde{b}(y,t) = c_1(t,\sigma_{\mathrm{data}})y + c_2(t,\sigma_{\mathrm{data}})\widetilde{f}(y,t),$$

it follows from (14) and Lemmata B.1 and B.4 that, for all $y \in \mathbb{R}^D$,

$$\begin{aligned}\|\widetilde{b}(y,t)\| &\lesssim \sigma_{\mathrm{data}}^{-2}\|y\| + \sigma_{\mathrm{data}}^{-2} \max_{-1\leqslant k\leqslant N+1} \|\alpha_k\| \sum_{k=-1}^{N+1} B_3(h^{-1}t - (k-2)) \\ &\lesssim \sigma_{\mathrm{data}}^{-2}\|y\| + \sigma_{\mathrm{data}}^{-2}\sqrt{D}.\end{aligned}$$

Similarly, for all $y \in \mathbb{R}^D$ and $t \in [0, 1-t_0]$, we have that

$$\|\nabla\widetilde{f}(y,t)\| \lesssim \max_{0\leqslant k\leqslant N} \|\nabla\widetilde{f}_k(y)\| \sum_{k=-1}^{N+1} B_3(h^{-1}t - (k-2)) \lesssim \left(\sigma_{\min}^{-2} D\log(1/\varepsilon)\right)^{\mathcal{O}(1)} \tag{18}$$

and, hence,

$$\|\nabla\widetilde{b}(y,t)\| \lesssim \sigma_{\mathrm{data}}^{-2} + \sigma_{\mathrm{data}}^{-2}\cdot\left(\sigma_{\min}^{-2} D\log(1/\varepsilon)\right)^{\mathcal{O}(1)} \lesssim \left(\sigma_{\min}^{-2} D\log(1/\varepsilon)\right)^{\mathcal{O}(1)}. \tag{19}$$

We now establish the velocity field approximation error. Specifically, for every $t \in [0, 1-t_0]$ and every $y \in \mathbb{R}^D$, it holds that

$$\begin{aligned}\|\widetilde{b}(y,t) - b^*(y,t)\| \lesssim |c_2(t,\sigma_{\mathrm{data}})| &\sum_{k=-1}^{N+1} \|\alpha_k - \alpha_k^*\| B_3(h^{-1}t - (k-2)) \\ &+ \sigma_{\mathrm{data}}^{-2} \left\| \sum_{k=-1}^{N+1} \alpha_k^* B_3(h^{-1}t - (k-2)) - f^*(y,t) \right\|,\end{aligned}$$

where the coefficients α_k^* correspond to f^*, and are defined analogously to α_k. The direct denoiser bound in Lemma B.1(i) gives, for every $0 \leqslant k \leqslant N$ and every t,

$$|c_2(t,\sigma_{\mathrm{data}})|\,\|\widetilde{f}_k - f^*(\cdot,\tau_k)\|_{L^2(\mathsf{p}_t^*)} \lesssim (D\exp\{\sigma_{\min}^{-2}\}\log(1/\varepsilon))^{\mathcal{O}(1)}\varepsilon^\beta.$$

The triangle inequality gives the same bound, up to a fixed factor, for the two extrapolated boundary coefficients. Combining this result with the B-spline properties detailed in Lemmata B.3 and B.4, it follows that

$$\begin{aligned}\|\widetilde{b}(\cdot,t) - b^*(\cdot,t)\|_{L^2(\mathsf{p}_t^*)} \lesssim (D\exp\{\sigma_{\min}^{-2}\}\log(1/\varepsilon))^{\mathcal{O}(1)}\varepsilon^\beta \\ + \sigma_{\mathrm{data}}^{-2}\mathbb{E}^{1/2}|f^*(X_t,\cdot)|^2_{W^{2,\infty}([0,1-t_0])}h^2.\end{aligned}$$

Finally, invoking Lemma B.2 and (16), we arrive at the final bound

$$\|\widetilde{b}(\cdot,t) - b^*(\cdot,t)\|_{L^2(\mathsf{p}_t^*)} \lesssim \left(D\exp\{\sigma_{\min}^{-2}\}\log(1/\varepsilon)\right)^{\mathcal{O}(1)}(\varepsilon^\beta + h^2), \quad t \in [0,1-t_0]. \tag{20}$$

ODE trajectory approximation. We employ Heun's method to approximate the ODE trajectory on a uniform grid. The following lemma quantifies the discretization error.

Lemma B.5 (Heun's Method error bound)**.** *Let* $N \in \mathbb{N}$ *and* $t_0 \in [0,1)$ *be arbitrary, and let* $\tau_k = kh$ *for all* $0 \leqslant k \leqslant N$*, where* $h = (1-t_0)/N$*. Let* f *and* f^* *be mappings from* $\mathbb{R}^D \times \mathbb{R} \to \mathbb{R}^D$ *such that for all* $y \in \mathbb{R}^D$ *and* $t \in [0, 1-t_0]$ *(and* τ_k *where applicable), the following hold:*

(i) $\|\nabla f(y,\tau_k)\| \vee \|\nabla f^*(y,t)\| \leqslant M_y$ *for all* $y \in \mathcal{B}(0,R)$ *for some* $R > 0$ *and for all* $t \in [0, 1-t_0]$.

(ii) $\|f^*(y,t)\| \leqslant M(1+\|y\|)$, $\|f(y,\tau_k)\| \leqslant M\|y\| + M_f$, $\|\partial_t f^*(y,t)\| \leqslant M_t(1+\|y\|)$.

(iii) $\|\nabla^2 f^*(y,t)\| \leqslant M_{yy}$, $\|\nabla\partial_t f^*(y,t)\| \leqslant M_{ty}(1+\|y\|)$, $\|\partial_{tt}^2 f^*(y,t)\| \leqslant M_{tt}(1+\|y\|)^2$.

Let also $\{Z(t) : 0 \leqslant t \leqslant 1 - t_0\}$ *be the solution to an ODE* $\dot{Z}(t) = f^*(Z(t), t)$ *with* $Z(0) = y$. *For* $0 \leqslant k \leqslant N - 1$, *consider the Heun discretization scheme:*

$$Y_{k+1} = Z_k + hf(Z_k, \tau_k), \quad Z_{k+1} = Z_k + \frac{h}{2}(f(Z_k, \tau_k) + f(Y_{k+1}, \tau_{k+1})), \quad Z_0 = y.$$

Assume further that h *is sufficiently small such that* $h(M_y \vee M) \leqslant 1$. *Then, for all* $y \in \mathbb{R}^D$ *satisfying*

$$(1 + \|y\|)e^M \vee \left(2\exp\{3M/2\}\left(\|y\| + \frac{3M_f}{2}\right) + M_f h\right) \leqslant R,$$

it holds that

$$\begin{aligned}
&\max_{0\leqslant k\leqslant N} \|Z_k - Z(\tau_k)\| \\
&\leqslant 4e^{(3M_y/2+2M)(1-t_0)}h^2(M_y(M_yM + M_t) + M_{yy}M^2 + 2M_{ty}M + M_{tt})(1 + \|y\|)^2 \\
&\quad + 4e^{3M_y(1-t_0)/2}h\sum_{k=0}^{N} \|f(Z(\tau_k), \tau_k) - f^*(Z(\tau_k), \tau_k)\|.
\end{aligned}$$

Furthermore, if we only assume the growth condition specified in (ii)*:*

$$\|f^*(y, t)\| \leqslant M(1+\|y\|),\ \|f(y, \tau_k)\| \leqslant M\|y\|+M_f, \quad (y, t) \in \mathbb{R}^D \times [0, 1-t_0],\ k \in \{0, \ldots, N\},$$

and also $hM \leqslant 1$, *then, for all* $y \in \mathbb{R}^D$ *and* $0 \leqslant t \leqslant 1 - t_0$, *the following holds:*

$$\max_{0\leqslant k\leqslant N} \|Z_k\| \leqslant \exp\{3M/2\}\left(\|y\| + \frac{3M_f}{2}\right), \quad \|Z(t)\| \leqslant (1 + \|y\|)e^{Mt}.$$

The proof of Lemma B.5 is deferred to Appendix B.3. We now apply Lemma B.1 at each grid point $\tau_k = kh$ for $k \in \{0, \ldots, N\}$, where N will be specified later in the proof. With the enlarged tube radius chosen above, $\mathcal{B}(0, R) \subseteq \mathcal{K}_{\tau_k}$ for all $0 \leqslant k \leqslant N$. Therefore, we can invoke Lemma B.5 using the velocity field approximations $\{\bar{b}(\cdot, \tau_k) : 0 \leqslant k \leqslant N\}$ and the true velocity field $b^*(y, t)$. The constants obtained from our construction are $M_y \asymp \sigma_{\text{data}}^{-4}$, $M \asymp \sigma_{\text{data}}^{-2}$, $M_f \asymp \sigma_{\text{data}}^{-2}\sqrt{D}$, $M_t \asymp \sigma_{\text{data}}^{-6}$, $M_{yy} \asymp \sigma_{\text{data}}^{-6}$, $M_{ty} \asymp \sigma_{\text{data}}^{-8}$, and $M_{tt} \asymp \sigma_{\text{data}}^{-10}$. Choose the constant in the exponent defining R larger than the constants in the growth bounds below. With a sufficiently small constant in the initial radius condition, this ensures that all sequences under consideration remain inside the ball where the local Lipschitz bounds hold. Therefore, we conclude that, for all $y \in \mathbb{R}^D$ satisfying

$$\|y\| \lesssim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}\left(\log(1/\varepsilon) + \log(D\sigma_{\text{data}}^{-2})\right), \tag{21}$$

it holds that

$$\begin{aligned}
&\max_{0\leqslant k\leqslant N} \|Z_k - Z(\tau_k)\| \\
&\lesssim \exp\{\mathcal{O}(\sigma_{\text{data}}^{-4})\}(1 + \|y\|)^2h^2 + \exp\{\mathcal{O}(\sigma_{\text{data}}^{-4})\}h\sum_{k=0}^{N} \|\bar{b}(Z(\tau_k), \tau_k) - b^*(Z(\tau_k), \tau_k)\|.
\end{aligned} \tag{22}$$

Here, $h \lesssim \sigma_{\text{data}}^4$ and $\dot{Z}(t) = b^*(Z(t), t)$ with $Z_0 = y$. Furthermore, for all $y \in \mathbb{R}^D$, it follows that

$$\max_{0\leqslant k\leqslant N} \|Z_k\| \lesssim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}(1 + \|y\|). \tag{23}$$

Interpolation of the ODE trajectory. In the previous step, we constructed an approximation of $Z(\tau_k)$ for all $\tau_k \in [0, 1-t_0]$ with $k \in \{0, \ldots, N\}$. We now extend the approximation to the interval $[0, 1 - t_0]$. To achieve this, we use B-splines. Specifically, consider the approximation

$$\widetilde{Z}(t) = \sum_{k=-1}^{N+1} \alpha_k(y)B_3(h^{-1}t - (k - 2)),$$

where the coefficients are given by $\alpha_k(y) = Z_k$ for all $0 \leqslant k \leqslant N$ and $\alpha_{-1}(y) = 2y - Z_1$, $\alpha_{N+1}(y) = 2Z_N - Z_{N-1}$. Similarly, we define $\alpha_k^* = Z(\tau_k)$ for every $0 \leqslant k \leqslant N$ and $\alpha_{-1}^*(y) =$

$2y - Z(\tau_1)$, $\alpha^*_{N+1}(y) = 2Z(\tau_N) - Z(\tau_{N-1})$. Hence, by Lemma B.3, for all $t \in [0, 1-t_0]$, it holds that

$$\begin{aligned}\|\widetilde{Z}(t) - Z(t)\| &\leqslant \sum_{k=-1}^{N+1} \|\alpha_k(y) - \alpha_k^*(y)\| B_3(h^{-1}t - (k-2)) \\ &\quad + \left\| \sum_{k=-1}^{N+1} \alpha_k^*(y) B_3(h^{-1}t - (k-2)) - Z(t) \right\| \\ &\lesssim \max_{0 \leqslant k \leqslant N} \|Z_k - Z(\tau_k)\| + |Z|_{W^{2,\infty}} h^2,\end{aligned} \tag{24}$$

where the last inequality uses partition of unity property of B-splines outlined in Lemma B.4. Observe that, according to Lemma B.5 and the bounds (15) and (17), it holds that

$$|Z|_{W^{2,\infty}([0,1-t_0])} \lesssim \exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}(1 + \|y\|), \quad \text{for all } y \in \mathbb{R}^D. \tag{25}$$

Therefore, from (22) we conclude that, for all y satisfying (21), it follows that

$$\begin{aligned}|\widetilde{Z} - Z|_{W^{0,\infty}([0,1-t_0])} &\lesssim \exp\{\mathcal{O}(\sigma_{\text{data}}^{-4})\}(1 + \|y\|)^2 h^2 + \exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}(1 + \|y\|) h^2 \\ &\quad + \exp\{\mathcal{O}(\sigma_{\text{data}}^{-4})\} h \sum_{k=0}^{N} \|\bar{b}(Z(\tau_k), \tau_k) - b^*(Z(\tau_k), \tau_k)\|.\end{aligned} \tag{26}$$

Furthermore, from Lemma B.5 in conjunction with (23), (24), and (25), we find that

$$|\widetilde{Z} - Z|_{W^{0,\infty}([0,1-t_0])} \lesssim \sqrt{D} \exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}(1 + \|y\|), \quad \text{for all } y \in \mathbb{R}^D. \tag{27}$$

In addition, we have that $\widetilde{Z}(t)$ also approximates the derivative of $Z(t)$. Specifically, applying Lemmata B.3 and B.4, we deduce that, for every $t \in [0, 1-t_0]$,

$$\begin{aligned}\|\partial_t \widetilde{Z}(t) - \partial_t Z(t)\| &\leqslant h^{-1} \left\| \sum_{k=-1}^{N+1} (\alpha_k(y) - \alpha_k^*(y))(B_2(h^{-1}t - (k-2)) - B_2(h^{-1}t - (k-1))) \right\| \\ &\quad + \left\| \partial_t \sum_{k=-1}^{N+1} \alpha_k^*(y) B_3(h^{-1}t - (k-2)) - \partial_t Z(t) \right\| \\ &\lesssim h^{-1} \max_{0 \leqslant k \leqslant N} \|Z_k - Z(\tau_k)\| + |Z|_{W^{2,\infty}([0,1-t_0])} h.\end{aligned} \tag{28}$$

Therefore, from the derived bound, (22), and (25), we find that, for every y satisfying (21), the following holds:

$$\begin{aligned}|\widetilde{Z} - Z|_{W^{1,\infty}([0,1-t_0])} &\lesssim \exp\{\mathcal{O}(\sigma_{\text{data}}^{-4})\} \sum_{k=0}^{N} \|\bar{b}(Z(\tau_k), \tau_k) - b^*(Z(\tau_k), \tau_k)\| \\ &\quad + \exp\{\mathcal{O}(\sigma_{\text{data}}^{-4})\}(1 + \|y\|)^2 h + \exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}(1 + \|y\|) h.\end{aligned} \tag{29}$$

Finally, from the bounds in (23), (25), (28), and Lemma B.5, we conclude that

$$|\widetilde{Z} - Z|_{W^{1,\infty}([0,1-t_0])} \lesssim h^{-1} \sqrt{D} \exp\{\mathcal{O}(\sigma_{\text{data}}^{-2})\}(1 + \|y\|), \quad \text{for all } y \in \mathbb{R}^D. \tag{30}$$

Flow approximation in the L^2-norm. Having derived bounds in Sobolev seminorms, we now derive the approximation guarantees in the L^2-norm. Let $\widetilde{R}$ denote the radius defined by the right-hand side of (21). We introduce the following approximations:

$$\begin{aligned}\widetilde{v}_{t,t}(y) &= \widetilde{b}(y, t), \quad (t, y) \in [0, 1-t_0] \times \mathbb{R}^D, \\ \widetilde{v}_{0,t}(y) &= \frac{\widetilde{Z}(t) - y}{t}, \quad (t, y) \in (0, 1-t_0] \times \mathbb{R}^D.\end{aligned}$$

Recall that $\widetilde{Z}$ depends on the initial point y. Let $\widetilde{\Psi}_{0,t}(y) = y + t\,\widetilde{v}_{0,t}(y) = \widetilde{Z}(t)$. From (26) and (27) we deduce that, for all $t \in [0, 1-t_0]$,

$$
\begin{aligned}
&\|\widetilde{\Psi}_{0,t} - \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} \\
&\leqslant \left(\int_{\mathcal{B}(0,\widetilde{R})} \|\widetilde{\Psi}_{0,t}(y) - \Psi^*_{0,t}(y)\|^2 \mathsf{p}^*_0(y) \mathrm{d}y \right)^{1/2} + \left(\int_{\mathbb{R}^D \setminus \mathcal{B}(0,\widetilde{R})} \|\widetilde{\Psi}_{0,t}(y) - \Psi^*_{0,t}(y)\|^2 \mathsf{p}^*_0(y) \mathrm{d}y \right)^{1/2} \\
&\lesssim \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\}(1+\widetilde{R})^2 h^2 + \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\} h \sum_{k=0}^{N} \|\bar{b}(\cdot,\tau_k) - b^*(\cdot,\tau_k)\|_{L^2(\mathsf{p}^*_{\tau_k})} \\
&\quad + \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-2})\} \left(\int_{\mathbb{R}^D \setminus \mathcal{B}(0,\widetilde{R})} (1+\|y\|)^2 \, \mathsf{p}^*_0(y) \, \mathrm{d}y \right)^{1/2} .
\end{aligned}
$$

By Lemma B.1, we have that

$$
\begin{aligned}
\|\widetilde{\Psi}_{0,t} - \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} &\lesssim \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\}(1+\widetilde{R})^2 h^2 + \exp\{\mathcal{O}(\sigma_{\mathrm{min}}^{-4})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)} \varepsilon^{\beta} \\
&\quad + \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-2})\} \left(\int_{\mathbb{R}^D \setminus \mathcal{B}(0,\widetilde{R})} (1+\|y\|)^2 \, \mathsf{p}^*_0(y) \, \mathrm{d}y \right)^{1/2} .
\end{aligned}
$$

Applying the concentration bound for Gaussian random vector (see Lemma E.1) together with the Hölder inequality and using (21), we arrive at

$$
\begin{aligned}
\|\widetilde{\Psi}_{0,t} - \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} &\lesssim \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\}(1+\widetilde{R})^2 h^2 + \exp\{\mathcal{O}(\sigma_{\mathrm{min}}^{-4})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)} \varepsilon^{\beta} \\
&\quad + D^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-2})\} \varepsilon^{\beta} \\
&\lesssim (D \log(1/\varepsilon))^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\mathrm{min}}^{-4})\} (\varepsilon^{\beta} + h^2).
\end{aligned} \tag{31}
$$

Similarly, we obtain the bound for the derivatives w.r.t. t. Formally, from (29) and (30) we have that

$$
\begin{aligned}
&\|\partial_t \widetilde{\Psi}_{0,t} - \partial_t \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} \lesssim \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\}(1+\widetilde{R})^2 h \\
&+ \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\} \sum_{k=0}^{N} \|\bar{b}(\cdot,\tau_k) - b^*(\cdot,\tau_k)\|_{L^2(\mathsf{p}^*_{\tau_k})} \\
&+ h^{-1} \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-2})\} \left(\int_{\mathbb{R}^D \setminus \mathcal{B}(0,\widetilde{R})} (1+\|y\|)^2 \mathsf{p}^*_0(y) \, \mathrm{d}y \right)^{1/2} .
\end{aligned}
$$

Now let $h \asymp \sigma_{\mathrm{min}}^2 \varepsilon^{\mathcal{O}(1)}$, where the exact dependence on ε will be specified later. By applying Lemma B.1, the Gaussian concentration bound from Lemma E.1, and the Hölder inequality, we conclude that

$$
\begin{aligned}
\|\partial_t \widetilde{\Psi}_{0,t} - \partial_t \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} &\lesssim \sqrt{D} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-4})\}(1+\widetilde{R})^2 h + h^{-1} \exp\{\mathcal{O}(\sigma_{\mathrm{min}}^{-4})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)} \varepsilon^{\beta} \\
&\quad + D^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\mathrm{data}}^{-2})\} \varepsilon^{\beta}.
\end{aligned}
$$

Thus, choose $N = \lceil C \sigma_{\mathrm{min}}^{-2} \varepsilon^{-\beta/2} \rceil$. Since $1/2 \leqslant 1 - t_0 \leqslant 1$, we have $h \asymp \sigma_{\mathrm{min}}^2 \varepsilon^{\beta/2}$; the smallness condition $\varepsilon^{\beta} \leqslant c \sigma_{\mathrm{min}}^4$ ensures $h(M_y \vee M) \leqslant 1$. This gives

$$
\|\partial_t \widetilde{\Psi}_{0,t} - \partial_t \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} \lesssim (D \log(1/\varepsilon))^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\mathrm{min}}^{-4})\} \varepsilon^{\beta/2}. \tag{32}
$$

From (31) we obtain

$$
\|\widetilde{\Psi}_{0,t} - \Psi^*_{0,t}\|_{L^2(\mathsf{p}^*_0)} \lesssim (D \log(1/\varepsilon))^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\mathrm{min}}^{-4})\} \varepsilon^{\beta}. \tag{33}
$$

Finally, the choice of h implies that $N \asymp \sigma_{\mathrm{min}}^{-2} \varepsilon^{-\beta/2}$ in Definition 3.5.

Approximating the loss function. From Proposition 2.1 and the fact that $R(v^*) = 0$ we find that

$$\begin{aligned}\mathbb{E}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] &= \mathbb{E}[\ell_{FM}(\widetilde{v}, X_1) - \ell_{FM}(v^*, X_1)] + R(\widetilde{v}) \\ &= \int_0^{1-t_0} \mathbb{E}\|\widetilde{v}_{t,t}(X_t) - b^*(X_t, t)\|^2 \,\mathrm{d}t + \int_0^{1-t_0} \mathbb{E}\|\partial_t \widetilde{\Psi}_{0,t}(X_0) - \widetilde{v}_{t,t}(\widetilde{\Psi}_{0,t}(X_0))\|^2 \,\mathrm{d}t.\end{aligned}$$

Taking into account that $\widetilde{v}_{t,t}$ is globally Lipschitz, as established by (19), we deduce that the last term on the right-hand side of the above expression is bounded by

$$\begin{aligned}&\int_0^{1-t_0} \mathbb{E}\|\partial_t \widetilde{\Psi}_{0,t}(X_0) - \widetilde{v}_{t,t}(\widetilde{\Psi}_{0,t}(X_0))\|^2 \,\mathrm{d}t \lesssim \int_0^{1-t_0} \mathbb{E}\|\partial_t \widetilde{\Psi}_{0,t}(X_0) - \partial_t \Psi^*_{0,t}(X_0)\|^2 \,\mathrm{d}t \\ &+ \left(D\sigma_{\mathrm{data}}^{-2} \log(1/\varepsilon)\right)^{\mathcal{O}(1)} \int_0^{1-t_0} \mathbb{E}\|\widetilde{\Psi}_{0,t}(X_0) - \Psi^*_{0,t}(X_0)\|^2 \,\mathrm{d}t + \int_0^{1-t_0} \mathbb{E}\|\widetilde{v}_{t,t}(X_t) - b^*(X_t, t)\|^2 \,\mathrm{d}t.\end{aligned}$$

By applying the bounds (20), (32), (33), and considering the choice of h, we conclude that

$$\mathbb{E}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] \lesssim (D \log(1/\varepsilon))^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\varepsilon^{\beta}.$$

The proof is now complete.

□

B.1 Proof of Lemma B.1

Notice that $X_t \sim \mathsf{p}_t^*$ satisfies Assumption 3.1 in Yakovlev and Puchkin [2026] with variance parameter $\widetilde{\alpha}_t^2 = \alpha_t^2 + \beta_t^2 \sigma_{\mathrm{data}}^2$. Specifically, the distribution of X_t coincides with that of $\beta_t g^*(U) + \widetilde{\alpha}_t Z$, where $U \sim \mathrm{Un}([0,1]^d)$ and $Z \sim \mathcal{N}(0, I_D)$ are independent. Furthermore, $0 \leqslant \beta_t \leqslant 1$ and $\sigma_{\mathrm{data}}^2/2 \leqslant \widetilde{\alpha}_t^2 \leqslant 1$ by Assumption 3.2.

Set the derivative order $m = 1$ in the construction of [Yakovlev and Puchkin, 2026, Theorem 3.2]. Let g° be the piecewise Taylor approximation of g^* from Lemma C.1 of that work. Then

$$\delta_g := \|g^\circ - g^*\|_{L^\infty([0,1]^d)} \lesssim \sqrt{D}\varepsilon^{\beta}, \quad \|g^\circ\|_{L^\infty([0,1]^d)} \leqslant 2,$$

provided that ε is sufficiently small. Define

$$f^\circ(y) = \frac{\int_{[0,1]^d} g^\circ(u) \exp\left(-\frac{\|y - \beta_t g^\circ(u)\|^2}{2\widetilde{\alpha}_t^2}\right) \mathrm{d}u}{\int_{[0,1]^d} \exp\left(-\frac{\|y - \beta_t g^\circ(u)\|^2}{2\widetilde{\alpha}_t^2}\right) \mathrm{d}u}.$$

We emphasize that the numerator contains g°, not $\beta_t g^\circ$. To obtain this function, apply the construction in Lemmata 4.3 and 4.4 of Yakovlev and Puchkin [2026] with $\beta_t g^\circ$. Choose

$$R \asymp \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} D \left(\log(1/\varepsilon) + \log(D\sigma_{\min}^{-2})\right), \tag{34}$$

and set $\mathcal{K}_\infty = [-R_\infty, R_\infty]^D$, where $R_\infty = (R+2) \vee (5/2)$. In particular, $\mathcal{K}_t \subseteq \mathcal{B}(0, R+1) \subseteq \mathcal{K}_\infty$. The proof of Theorem 3.2 and Lemma 4.4 of Yakovlev and Puchkin [2026] yield a GELU network $\bar{f} \in \mathsf{NN}(L, W, S, B)$ with architecture

$$\begin{aligned}L &\lesssim \log(DR\sigma_{\min}^{-2} \log(1/\varepsilon)) \lesssim \sigma_{\min}^{-2} \log(D\sigma_{\min}^{-2} \log(1/\varepsilon)), \\ \|W\|_\infty \vee S &\lesssim \varepsilon^{-d} \left(DR\sigma_{\min}^{-2} \log(1/\varepsilon) \log(DR\sigma_{\min}^{-2})\right)^{\mathcal{O}(1)} \lesssim \varepsilon^{-d} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)}, \\ \log B &\lesssim \left(R\sigma_{\min}^{-2} \log(1/\varepsilon) \log(DR\sigma_{\min}^{-2})\right)^{\mathcal{O}(1)} \lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)}.\end{aligned}$$

Moreover, for every $1 \leqslant l \leqslant D$,

$$\begin{aligned}&\max_{j \in \{0,1\}} |\bar{f}_l - f_l^\circ|_{W^{j,\infty}(\mathcal{K}_\infty)} \leqslant \varepsilon^{\beta}, \\ &\|\bar{f}_l\|_{L^\infty(\mathbb{R}^D)} \lesssim 1, \\ &|\bar{f}_l|_{W^{1,\infty}(\mathbb{R}^D)} \lesssim \sigma_{\mathrm{data}}^{-2} \left(D \log(1/\varepsilon) \log(R\sigma_{\mathrm{data}}^{-2})\right)^{\mathcal{O}(1)}.\end{aligned} \tag{35}$$

Here the first inequality controls both the function and its first derivatives. Differentiating the normalized integral defining f° implies that

$$\nabla f^\circ(y) = \frac{\beta_t}{\widetilde{\alpha}_t^2} \operatorname{Cov}_{y,t}(g^\circ(U)),$$

where the covariance is taken under the probability measure on $[0,1]^d$ whose density is proportional to $\exp(-\|y - \beta_t g^\circ(u)\|^2/(2\widetilde{\alpha}_t^2))$. Consequently, $\|\nabla f^\circ(y)\| \lesssim \sigma_{\mathrm{data}}^{-2}$ for all $y \in \mathbb{R}^D$. Taking into account that $\mathcal{K}_t \subseteq \mathcal{K}_\infty$, we obtain

$$\begin{aligned}\sup_{y\in\mathcal{K}_t} \|\nabla \bar{f}(y)\| &\leqslant D \max_{1\leqslant l\leqslant D} |\bar{f}_l - f_l^\circ|_{W^{1,\infty}(\mathcal{K}_t)} + \sup_{y\in\mathcal{K}_t} \|\nabla f^\circ(y)\| \\ &\lesssim D\varepsilon^\beta + \sigma_{\mathrm{data}}^{-2} \lesssim \sigma_{\mathrm{data}}^{-2}.\end{aligned}$$

The last inequality follows from (12). Furthermore, (35) yields $\|\bar{f}\|_{L^\infty(\mathbb{R}^D)} \lesssim \sqrt{D}$. Combining (34) and (35) also gives

$$\max_{1\leqslant l\leqslant D} |\bar{f}_l|_{W^{1,\infty}(\mathbb{R}^D)} \lesssim \big(\sigma_{\min}^{-2} D \log(1/\varepsilon)\big)^{\mathcal{O}(1)}.$$

This establishes claim (ii).

We now prove claim (i). For $\|y\| \leqslant R+1$, the logarithms of the Gaussian weights corresponding to g° and g^* satisfy

$$\sup_{u\in[0,1]^d} \left| \frac{\|y - \beta_t g^\circ(u)\|^2 - \|y - \beta_t g^*(u)\|^2}{2\widetilde{\alpha}_t^2} \right| \lesssim \sigma_{\mathrm{data}}^{-2}(R+3)\delta_g.$$

Let $\eta = C\sigma_{\mathrm{data}}^{-2}(R+3)\delta_g$. If $\eta \leqslant 1$, the ratio of the corresponding normalized densities lies between $e^{-2\eta}$ and $e^{2\eta}$. Subtracting the normalized integrals therefore gives

$$\sup_{\|y\|\leqslant R+1} \|f^\circ(y) - f^*(y,t)\| \lesssim \delta_g + \eta \lesssim \sqrt{D}\, \sigma_{\mathrm{data}}^{-2}(R+3)\varepsilon^\beta.$$

Together with (35), this implies

$$\sup_{\|y\|\leqslant R+1} \|\bar{f}(y) - f^*(y,t)\| \lesssim \sqrt{D}\, \sigma_{\mathrm{data}}^{-2}(R+3)\varepsilon^\beta.$$

For every $s \in [0, 1-t_0]$, the distribution of X_s coincides with that of $\beta_s g^*(U) + \widetilde{\alpha}_s Z$, with $\|\beta_s g^*(U)\| \leqslant 1$ and $\widetilde{\alpha}_s \leqslant 1$. Hence, by Lemma E.1 and a sufficiently large constant in (34),

$$\sup_{0\leqslant s\leqslant 1-t_0} \mathbb{P}(\|X_s\| > R+1) \leqslant \mathbb{P}(\|Z\| > R) \lesssim \varepsilon^{2\beta}.$$

Since $\bar{f}$ and $f^*(\cdot,t)$ are globally bounded by $C\sqrt{D}$, splitting the integral into $\mathcal{B}(0,R+1)$ and its complement yields

$$\begin{aligned}\sup_{0\leqslant s\leqslant 1-t_0} \|\bar{f} - f^*(\cdot,t)\|^2_{L^2(\mathsf{p}_s^*)} &\lesssim D\sigma_{\mathrm{data}}^{-4}(R+3)^2\varepsilon^{2\beta} + D\varepsilon^{2\beta} \\ &\lesssim \big(D\exp(\sigma_{\min}^{-2})\log(1/\varepsilon)\big)^{\mathcal{O}(1)}\varepsilon^{2\beta}.\end{aligned}$$

Finally,

$$\bar{b}(y) - b^*(y,t) = c_2(t,\sigma_{\mathrm{data}})\big(\bar{f}(y) - f^*(y,t)\big),$$

and $|c_2(t,\sigma_{\mathrm{data}})| \lesssim \sigma_{\mathrm{data}}^{-2}$. This proves (i).

It remains to check the smallness requirements. The assumptions of [Yakovlev and Puchkin, 2026, Lemma 4.3] are ensured by

$$\sqrt{D}(R+3)(\varepsilon^\beta \vee \varepsilon) \lesssim \sigma_{\mathrm{data}}^2.$$

Substituting (34), this follows from (12). The proof is complete.

□

B.2 Proof of Lemma B.2

Fix arbitrary $y \in \mathbb{R}^D$ and $t \in [0, 1 - t_0]$. It follows immediately from (10) that $\|f^*(y,t)\| \leqslant 1$. For brevity, write

$$a(t) = \frac{\beta_t}{\widetilde{\alpha}_t^2}, \quad q(t) = \frac{\beta_t^2}{\widetilde{\alpha}_t^2},$$

and introduce

$$w(u) = \exp\left(a(t) y^\top g^*(u) - \frac{q(t)}{2}\|g^*(u)\|^2\right), \quad J = \int\limits_{[0,1]^d} w(u)\,\mathrm{d}u.$$

The dependence of w and J on (y,t) is suppressed. Then

$$f^*(y,t) = \frac{1}{J}\int\limits_{[0,1]^d} g^*(u) w(u)\,\mathrm{d}u. \tag{36}$$

For products of integration variables, let us write $g_j = g^*(u_j)$ and $w_j = w(u_j)$. Taking the gradient with respect to y yields

$$\nabla f^*(y,t) = \frac{a(t)}{J^2}\int\limits_{([0,1]^d)^2} P_y(u_1,u_2) w_1 w_2\,\mathrm{d}u_1\mathrm{d}u_2,$$

where $P_y(u_1,u_2) = g_1 \otimes (g_1 - g_2)$. To obtain the constant in (i), observe that

$$\frac{1}{J^2}\int\limits_{([0,1]^d)^2} P_y(u_1,u_2) w_1 w_2\,\mathrm{d}u_1\mathrm{d}u_2 = \frac{1}{J}\int\limits_{[0,1]^d} g^*(u)\otimes g^*(u)\, w(u)\,\mathrm{d}u - f^*(y,t)\otimes f^*(y,t).$$

This is a covariance matrix under the probability density w/J. Its operator norm is at most one, since $\|g^*(u)\| \leqslant 1$. Consequently, $\|\nabla f^*(y,t)\| \leqslant a(t) \leqslant 2\sigma_{\mathrm{data}}^{-2}$, where the last inequality follows from $\beta_t \leqslant 1$ and $\widetilde{\alpha}_t^2 \geqslant \sigma_{\mathrm{data}}^2/2$. This proves (i).

Taking another derivative with respect to y, we obtain

$$\nabla^2 f^*(y,t) = \frac{a(t)^2}{J^3}\int\limits_{([0,1]^d)^3} P_{yy}(u_1,u_2,u_3) w_1 w_2 w_3\,\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3,$$

where $P_{yy}(u_1,u_2,u_3) = P_y(u_1,u_2)\otimes(g_1+g_2-2g_3)$. Using $\|P_y\| \leqslant 2$ and $\|g_1+g_2-2g_3\| \leqslant 4$, we find that

$$\|\nabla^2 f^*(y,t)\| \leqslant 8a(t)^2 \leqslant 32\sigma_{\mathrm{data}}^{-4},$$

which proves (iii).

By differentiating (36) with respect to t, we obtain

$$\partial_t f^*(y,t) = \frac{\displaystyle\int\limits_{([0,1]^d)^2} P_t(u_1,u_2) w_1 w_2\,\mathrm{d}u_1\mathrm{d}u_2}{J^2}, \tag{37}$$

where

$$P_t(u_1,u_2) = a'(t) g_1 y^\top (g_1 - g_2) - \frac{q'(t)}{2} g_1\big(\|g_1\|^2 - \|g_2\|^2\big). \tag{38}$$

Set

$$Q(y,t) = \|y\|\,|a'(t)| + \frac{1}{2}|q'(t)|.$$

Since $\|g_j\| \leqslant 1$, we have $\|P_t\| \leqslant 2Q(y,t)$ and therefore

$$\|\partial_t f^*(y,t)\| \leqslant 2Q(y,t) = 2\|y\|\,|a'(t)| + |q'(t)|.$$

Hence, claim (ii) readily follows.

We next differentiate (37) with respect to y:

$$\nabla\partial_t f^*(y,t) = \frac{1}{J^3}\int\limits_{([0,1]^d)^3} P_{y,t}(u_1,u_2,u_3)w_1w_2w_3\,\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3,$$

where

$$P_{y,t}(u_1,u_2,u_3) = a(t)P_t(u_1,u_2)\otimes(g_1+g_2-2g_3) + \nabla_y P_t(u_1,u_2).$$

In particular,

$$\nabla_y P_t(u_1,u_2) = a'(t)g_1\otimes(g_1-g_2), \quad \|\nabla_y P_t(u_1,u_2)\| \leqslant 2|a'(t)|.$$

Hence,

$$\begin{aligned}\|\nabla\partial_t f^*(y,t)\| &\leqslant 4a(t)\|P_t\|_{L^\infty(([0,1]^d)^2)} + 2|a'(t)| \\ &\leqslant 8a(t)Q(y,t) + 2|a'(t)| \\ &\leqslant 16\sigma_{\mathrm{data}}^{-2}\left(\|y\|\,|a'(t)| + \frac{1}{2}|q'(t)|\right) + 2|a'(t)|.\end{aligned}$$

Therefore, claim (iv) readily follows.

Finally, define

$$\phi_j = a'(t)y^\top g_j - \frac{q'(t)}{2}\|g_j\|^2, \quad |\phi_j| \leqslant Q(y,t).$$

Differentiating (37) once more in t yields

$$\partial_{tt} f^*(y,t) = \frac{1}{J^3}\int\limits_{([0,1]^d)^3} P_{tt}(u_1,u_2,u_3)w_1w_2w_3\,\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3,$$

where

$$P_{tt}(u_1,u_2,u_3) = \partial_t P_t(u_1,u_2) + P_t(u_1,u_2)(\phi_1+\phi_2-2\phi_3).$$

From (38) we find that

$$\|\partial_t P_t(u_1,u_2)\| \leqslant 2\|y\|\,|a''(t)| + |q''(t)|.$$

Furthermore, $|\phi_1+\phi_2-2\phi_3| \leqslant 4Q(y,t)$. Consequently,

$$\begin{aligned}\|\partial_{tt} f^*(y,t)\| &\leqslant 2\|y\|\,|a''(t)| + |q''(t)| + 4Q(y,t)\|P_t\|_{L^\infty(([0,1]^d)^2)} \\ &\leqslant 2\|y\|\,|a''(t)| + |q''(t)| + 8\left(\|y\|\,|a'(t)| + \frac{1}{2}|q'(t)|\right)^2.\end{aligned}$$

Claim (v) follows immediately. The proof is complete.

□

B.3 Proof of Lemma B.5

For clarity, we divide the proof into several steps.

Bounding $\|Z_k\|$. Let us first establish the last claim of the lemma. For all $0 \leqslant k \leqslant N-1$, it holds that

$$\begin{aligned}\|Z_{k+1}\| &\leqslant \|Z_k\| + \frac{h}{2}\left(2M_f + M\|Z_k\| + M\|Y_{k+1}\|\right) \\ &\leqslant \|Z_k\| + \frac{h}{2}\left(2M_f + M\|Z_k\| + M(\|Z_k\| + h(M_f + M\|Z_k\|))\right) \\ &= \left(1 + Mh + \frac{M^2h^2}{2}\right)\|Z_k\| + M_f h + \frac{MM_f h^2}{2}.\end{aligned}$$

Using the assumption that $Mh \leqslant 1$ and applying the discrete Grönwall lemma, we arrive at

$$\max_{0\leqslant k\leqslant N}\|Z_k\| \leqslant \left(1+\frac{3Mh}{2}\right)^N\|y\| + \frac{3M_f}{2}\left(1+\frac{3Mh}{2}\right)^N \leqslant \exp\{3M/2\}\left(\|y\| + \frac{3M_f}{2}\right).$$

In addition, we have that, for all $0 \leqslant k \leqslant N-1$,

$$\begin{aligned}\|Y_{k+1}\| &\leqslant \|Z_k\| + h\|f(Z_k,\tau_k)\| \leqslant (1+Mh)\|Z_k\| + M_f h \\ &\leqslant 2\exp\{3M/2\}\left(\|y\| + \frac{3M_f}{2}\right) + M_f h.\end{aligned}$$

We also note that, for all $0 \leqslant t \leqslant 1-t_0$ and $y \in \mathbb{R}^D$,

$$\|Z(t)\| \leqslant \|y\| + \int_0^t \|f^*(Z(s),s)\|\,\mathrm{d}s \leqslant \|y\| + M\int_0^t \mathrm{d}s + M\int_0^t \|Z(s)\|\,\mathrm{d}s.$$

Hence, the Grönwall lemma yields that

$$\|Z(t)\| \leqslant (1+\|y\|)e^{Mt}, \quad \text{for all } y \in \mathbb{R}^D,\ 0 \leqslant t \leqslant 1-t_0, \tag{39}$$

and the last claim follows.

ODE solver error bound. Consider $y \in \mathbb{R}^D$ satisfying the inequality

$$(1+\|y\|)e^M \vee \left(2\exp\{3M/2\}\left(\|y\| + \frac{3M_f}{2}\right) + M_f h\right) \leqslant R.$$

Therefore, for all such y we have that $Z(t) \in \mathcal{B}(0,R)$ for all $0 \leqslant t \leqslant 1-t_0$, $Z_k \in \mathcal{B}(0,R)$ for all $0 \leqslant k \leqslant N$, and $Y_k \in \mathcal{B}(0,R)$ for every $1 \leqslant k \leqslant N$. We now proceed with the approximation error bound for the ODE solver. For every $0 \leqslant k \leqslant N-1$ we have that

$$\begin{aligned}\|Z_{k+1} - Z(\tau_{k+1})\| &\leqslant \|Z_k - Z(\tau_k)\| + \frac{h}{2}\left(M_y\|Z_k - Z(\tau_k)\| + \|f(Z(\tau_k),\tau_k) - f^*(Z(\tau_k),\tau_k)\|\right) \\ &\quad + \frac{h}{2}\left(M_y\|Y_{k+1} - Z(\tau_{k+1})\| + \|f(Z(\tau_{k+1}),\tau_{k+1}) - f^*(Z(\tau_{k+1}),\tau_{k+1})\|\right) \\ &\quad + \left\|\int_{\tau_k}^{\tau_{k+1}} f^*(Z(\tau),\tau)\mathrm{d}\tau - \frac{h}{2}(f^*(Z(\tau_k),\tau_k) + f^*(Z(\tau_{k+1}),\tau_{k+1}))\right\|. \end{aligned} \tag{40}$$

Next, notice that the function

$$g : u \mapsto \int_{\tau_k}^{\tau_k+u} f^*(Z(\tau),\tau)\mathrm{d}\tau - \frac{u}{2}(f^*(Z(\tau_k),\tau_k) + f^*(Z(\tau_k+u),\tau_k+u))$$

satisfies $g(0) = 0$ and $g'(0) = 0$. Furthermore, its second derivative is

$$\begin{aligned}\frac{\mathrm{d}^2 g(u)}{\mathrm{d}u^2} &= \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{1}{2}(f^*(Z(\tau_k+u),\tau_k+u) - f^*(Z(\tau_k),\tau_k)) - \frac{u}{2}\frac{\mathrm{d}f^*(Z(\tau_k+u),\tau_k+u)}{\mathrm{d}u}\right] \\ &= -\frac{u}{2}\frac{\mathrm{d}^2 f^*(Z(\tau_k+u),\tau_k+u)}{\mathrm{d}u^2}.\end{aligned}$$

Therefore, by Taylor's theorem, we have that

$$\begin{aligned}&\left\|\int_{\tau_k}^{\tau_{k+1}} f^*(Z(\tau),\tau)\mathrm{d}\tau - \frac{h}{2}(f^*(Z(\tau_k),\tau_k) + f^*(Z(\tau_{k+1}),\tau_{k+1}))\right\| \\ &\leqslant \frac{h^3}{4}\max_{0\leqslant u\leqslant h}\left\|\frac{\mathrm{d}^2 f^*(Z(\tau_k+u),\tau_k+u)}{\mathrm{d}u^2}\right\|.\end{aligned}$$

We now observe that the second derivative satisfies

$$\begin{aligned}\left\|\frac{\mathrm{d}^2 f^*(Z(\tau_k+u),\tau_k+u)}{\mathrm{d}u^2}\right\| &= \left\|\nabla^2 f^*[f^*,f^*] + 2\nabla\partial_t f^*\, f^* + \partial_{tt} f^* + \nabla f^*(\nabla f^*\, f^* + \partial_t f^*)\right\| \\ &\leqslant (M_y(M_y M + M_t) + M_{yy}M^2 + 2M_{ty}M + M_{tt})(1+\|Z(\tau_k+u)\|)^2,\end{aligned}$$

where we abbreviate f^* for $f^*(Z(\tau_k+u),\tau_k+u)$. Hence, we obtain from (39) that

$$\begin{aligned}&\max_{0\leqslant u\leqslant h}\left\|\frac{\mathrm{d}^2 f^*(Z(\tau_k+u),\tau_k+u)}{\mathrm{d}u^2}\right\|\\&\leqslant 4(M_y(M_yM+M_t)+M_{yy}M^2+2M_{ty}M+M_{tt})(1+\|y\|)^2e^{2M(1-t_0)}.\end{aligned}$$

Thus, it follows that

$$\begin{aligned}&\left\|\int_{\tau_k}^{\tau_{k+1}} f^*(Z(\tau),\tau)\mathrm{d}\tau-\frac{h}{2}(f^*(Z(\tau_k),\tau_k)+f^*(Z(\tau_{k+1}),\tau_{k+1}))\right\| \\&\leqslant h^3(M_y(M_yM+M_t)+M_{yy}M^2+2M_{ty}M+M_{tt})(1+\|y\|)^2e^{2M(1-t_0)}.\end{aligned} \tag{41}$$

Following this, we bound the error corresponding to the auxiliary sequence. For all $0\leqslant k\leqslant N-1$, we obtain

$$\begin{aligned}\|Y_{k+1}-Z(\tau_{k+1})\|&\leqslant \|Z_k-Z(\tau_k)\|+M_yh\|Z_k-Z(\tau_k)\|+h\|f^*(Z(\tau_k),\tau_k)-f(Z(\tau_k),\tau_k)\|\\&\quad+\left\|\int_{\tau_k}^{\tau_{k+1}}[f^*(Z(\tau),\tau)-f^*(Z(\tau_k),\tau_k)]\,\mathrm{d}\tau\right\|.\end{aligned} \tag{42}$$

The last term is evaluated as

$$\begin{aligned}&\left\|\int_{\tau_k}^{\tau_{k+1}}[f^*(Z(\tau),\tau)-f^*(Z(\tau_k),\tau_k)]\,\mathrm{d}\tau\right\|\\&\leqslant\int_{\tau_k}^{\tau_{k+1}}\int_{\tau_k}^{\tau}\left\|\frac{\mathrm{d}f^*(Z(u),u)}{\mathrm{d}u}\right\|\mathrm{d}u\,\mathrm{d}\tau\\&\leqslant\frac{h^2}{2}\max_{u\in[\tau_k,\tau_{k+1}]}\|\nabla f^*(Z(u),u)f^*(Z(u),u)+\partial_t f^*(Z(u),u)\|\\&\leqslant\frac{h^2}{2}(M_yM+M_t)\left(1+\max_{u\in[\tau_k,\tau_{k+1}]}\|Z(u)\|\right).\end{aligned}$$

Using the bound (39), we arrive at

$$\left\|\int_{\tau_k}^{\tau_{k+1}}[f^*(Z(\tau),\tau)-f^*(Z(\tau_k),\tau_k)]\,\mathrm{d}\tau\right\|\leqslant h^2(M_yM+M_t)e^{M(1-t_0)}(1+\|y\|).$$

Therefore, combining this bound with (40), (41), and (42), we conclude that

$$\begin{aligned}&\|Z_{k+1}-Z(\tau_{k+1})\|\\&\leqslant\left(1+\frac{3}{2}hM_y\right)\|Z_k-Z(\tau_k)\|+h^3M_y(M_yM+M_t)(1+\|y\|)e^{M(1-t_0)}\\&\quad+h(\|f(Z(\tau_k),\tau_k)-f^*(Z(\tau_k),\tau_k)\|+\|f(Z(\tau_{k+1}),\tau_{k+1})-f^*(Z(\tau_{k+1}),\tau_{k+1})\|)\\&\quad+h^3(M_y(M_yM+M_t)+M_{yy}M^2+2M_{ty}M+M_{tt})(1+\|y\|)^2e^{2M(1-t_0)},\end{aligned}$$

where we used $hM_y\leqslant 1$. Thus, applying the discrete Grönwall lemma yields

$$\begin{aligned}&\max_{0\leqslant k\leqslant N}\|Z_k-Z(\tau_k)\|\\&\leqslant 4e^{3M_y(1-t_0)/2}h^2(M_y(M_yM+M_t)+M_{yy}M^2+2M_{ty}M+M_{tt})(1+\|y\|)^2e^{2M(1-t_0)}\\&\quad+4e^{3M_y(1-t_0)/2}h\sum_{k=0}^{N}\|f(Z(\tau_k),\tau_k)-f^*(Z(\tau_k),\tau_k)\|.\end{aligned}$$

The desired claim now follows immediately.

□

B.4 Proof of Lemma B.4

Note that properties 1, 2, 3, and 5 follow directly from standard results in Phillips [2003]:

- Property 1 is established by Problem 6.2.3;
- Property 2 follows from Theorem 6.2.1;
- Property 3 follows from Theorem 6.2.5;
- Property 5 follows from Theorem 6.2.2;

We now prove property 4. First, observe that $x \mapsto B_3(x+2)$ is symmetric about zero. Consequently, at $x = 0$,

$$\sum_{k=-1}^{K+1} kh\, B_3(-(k-2)) = \sum_{k\in\mathbb{Z}} kh\, B_3(-(k-2)) = 0.$$

To extend this to all $x \in [0, b]$, it suffices to show that, for all such x,

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\sum_{k=-1}^{K+1}(kh-x)B_3(h^{-1}x-(k-2))\right] = 0.$$

By property 5, we have that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\sum_{k=-1}^{K+1}(kh-x)B_3(h^{-1}x-(k-2))\right] = -\sum_{k=-1}^{K+1} B_3(h^{-1}x-(k-2))$$
$$+ h^{-1}\sum_{k=-1}^{K+1}(kh-x)\left(B_2(h^{-1}x-(k-2)) - B_2(h^{-1}x-(k-1))\right).$$

Applying property 3 and rearranging the terms, we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\sum_{k=-1}^{K+1}(kh-x)B_3(h^{-1}x-(k-2))\right] = h^{-1}(-h-x)B_2(h^{-1}x+3)$$
$$- h^{-1}(Kh+h-x)B_2(h^{-1}x-K) - B_2(h^{-1}x+3).$$

Using property 2, we deduce that the derivative is zero on $[0, b]$, proving property 4. The proof is finished.

□

B.5 Proof of Lemma B.3

We prove claims (i) and (ii) separately.

Proving claim (i)**.** First, assume that $f \in C^2([-h, 1-t_0+h])$ and set $\alpha_{-1} = f(-h)$, $\alpha_{K+1} = f(1-t_0+h)$. Let $\bar{f}$ be the quasi-interpolant with these coefficients to avoid confusions. We will reduce the original problem to this setting later in the proof. Using the Taylor expansion around x, we obtain

$$f(y) = f(x) + f'(x)(y-x) + \int_x^y (y-t)f''(t)\,\mathrm{d}t.$$

Hence, substituting this into the definition of $\bar{f}$, we deduce that, for all $x \in [0, 1-t_0]$,

$$\bar{f}(x) = \sum_{k=-1}^{K+1}\left(f(x) + f'(x)(kh-x) + \int_x^{kh}(kh-t)f''(t)\,\mathrm{d}t\right)B_3(h^{-1}x-(k-2)).$$

Using the fact that $\{B_3(h^{-1}x-(k-2))(x) : -1 \leqslant k \leqslant K+1\}$ sum to one and

$$\sum_{k=-1}^{K+1}(kh-x)B_3(h^{-1}x-(k-2))=0$$

for $0 \leqslant x \leqslant 1-t_0$ (see Lemma B.4), we conclude that

$$\begin{aligned}|\bar{f}(x)-f(x)| &\leqslant \sum_{k=-1}^{K+1}\left|\int_x^{kh}(kh-t)f''(t)\,\mathrm{d}t\right| B_3(h^{-1}x-(k-2))\\ &\leqslant |f|_{W^{2,\infty}([-h,1-t_0+h])}\sum_{k=-1}^{K+1}\frac{(kh-x)^2}{2}B_3(h^{-1}x-(k-2)).\end{aligned}$$

Notice that $B_3(h^{-1}x-(k-2))$ is non-zero if and only if $(k-2)h \leqslant x \leqslant (k+2)h$ and, hence,

$$\begin{aligned}|\bar{f}(x)-f(x)| &\leqslant |f|_{W^{2,\infty}([-h,1-t_0+h])}\sum_{k=-1}^{K+1}\frac{(2h)^2}{2}B_3(h^{-1}x-(k-2))\\ &= 2h^2|f|_{W^{2,\infty}([-h,1-t_0+h])} \qquad (43)\end{aligned}$$

holds for all $x \in [0,1-t_0]$. We now consider the case where the function f is defined on $[0,1-t_0]$. Extend f to the interval $[-h,1-t_0]$ by parabolic extrapolation:

$$f(x)=f(0)+f'(0)x+f''(0)x^2/2, \quad x\in[-h,0],$$

and, similarly,

$$f(x)=f(1-t_0)+f'(1-t_0)(x-1+t_0)+f''(1-t_0)(x-1+t_0)^2/2, \quad x\in[1-t_0,1-t_0+h].$$

Hence, the resulting extension belongs to $C^2([-h,1-t_0+h])$. Taking into account that $0 \leqslant B_3(x) \leqslant 1$ for all $x \in \mathbb{R}$, we deduce that

$$|\widetilde{f}(x)-\bar{f}(x)| \leqslant |f(-h)-(2f(0)-f(\tau_1))|+|f(1-t_0+h)-(2f(\tau_K)-f(\tau_{K-1}))|.$$

We now elaborate on the first term on the right-hand side of the above bound. From the Taylor theorem it follows that

$$\begin{aligned}&|f(-h)-2f(0)+f(\tau_1)|\\ &=\left|f(0)+f'(0)(-h)+\int_0^{-h}(-h-t)f''(t)\,\mathrm{d}t-2f(0)+f(0)+f'(0)h+\int_0^{h}(h-t)f''(t)\,\mathrm{d}t\right|\\ &\leqslant |f|_{W^{2,\infty}([-h,1-t_0+h])}h^2. \qquad (44)\end{aligned}$$

The same bound also holds for the right boundary term. Therefore, this bound together with (43) implies that

$$|\widetilde{f}(x)-f(x)| \leqslant |\widetilde{f}(x)-\bar{f}(x)|+|\bar{f}(x)-f(x)| \leqslant 4|f|_{W^{2,\infty}([-h,1-t_0+h])}h^2, \quad 0\leqslant x\leqslant 1-t_0.$$

Finally, observe that the extension satisfies

$$|f|_{W^{2,\infty}([-h,1-t_0+h])}=|f|_{W^{2,\infty}([0,1-t_0])},$$

and the claim (i) follows.

Proving claim (ii)**.** As in the proof of claim (i), we first assume that the function f is defined on $[-h,1-t_0+h]$. By Lemma B.4, for all $x \in [0,1-t_0]$, we have that

$$\begin{aligned}\bar{f}'(x) &= h^{-1}\sum_{k=-1}^{K+1}f(kh)B_3'(h^{-1}x-(k-2))\\ &= h^{-1}\sum_{k=-1}^{K+1}f(kh)(B_2(h^{-1}x-(k-2))-B_2(h^{-1}x-(k-1)))\\ &= h^{-1}\sum_{k=0}^{K+1}[f(kh)-f((k-1)h)]B_2(h^{-1}x-(k-2)),\end{aligned}$$

where the last equality follows from the observation that $B_2(h^{-1}x + 3) = 0$ for all $x \geqslant 0$ and, similarly, $B_2(h^{-1}x - K) = 0$ for all $x \leqslant Kh$. The Taylor theorem implies that

$$\begin{aligned}
&\bar{f}'(x) - f'(x) \\
&= \sum_{k=0}^{K+1} \left[f'((k-1)h) - f'(x) + h^{-1} \int_{(k-1)h}^{kh} f''(t)(kh - t)\mathrm{d}t \right] B_2(h^{-1}x - (k-2)).
\end{aligned}$$

This yields the error bound

$$|\bar{f}'(x) - f'(x)| \leqslant \sum_{k=0}^{K+1} \left| \int_{x}^{(k-1)h} f''(t)\,\mathrm{d}t \right| B_2(h^{-1}x - (k-2)) + \frac{h}{2}|f|_{W^{2,\infty}([-h,1-t_0+h])}.$$

Since $B_2(h^{-1}x - (k-2))$ is non-zero only if $(k-2)h \leqslant x \leqslant (k+1)h$, then

$$\begin{aligned}
|\bar{f}'(x) - f'(x)| &\leqslant \sum_{k=0}^{K+1} (3h)|f|_{W^{2,\infty}([-h,1-t_0+h])} B_2(h^{-1}x - (k-2)) + \frac{h}{2}|f|_{W^{2,\infty}([-h,1-t_0+h])} \\
&= \frac{7h}{2}|f|_{W^{2,\infty}([-h,1-t_0+h])}, \qquad (45)
\end{aligned}$$

for all $x \in [0, 1 - t_0]$. Therefore, from (44) and Lemma B.4 we deduce that

$$\begin{aligned}
&|\widetilde{f}'(x) - \bar{f}'(x)| \\
&\leqslant h^{-1} \max_{y \in \mathbb{R}} |B_3'(y)| \cdot (|f(-h) - (2f(0) - f(\tau_1))| + |f(1 - t_0 + h) - (2f(\tau_K) - f(\tau_{K-1}))|) \\
&\leqslant 4h|f|_{W^{2,\infty}([-h,1-t_0+h])}.
\end{aligned}$$

Hence, claim (ii) follows immediately from (45), the above bound, and the fact that the extension of f has the same $W^{2,\infty}$-seminorm.

□

C Proof of Theorem 3.7

We divide the proof into several steps for clarity.

Verifying the Bernstein condition. We begin by establishing that the excess loss function class fulfills Bernstein's condition. Fix an arbitrary $v \in \mathcal{V}(L, W, S, B, N, R, L_f)$ (see Definition 3.5). From (6) we deduce that

$$\mathrm{Var}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] = \mathrm{Var}[\ell_{FM}(v, X_1) - \ell_{FM}(v^*, X_1)]. \qquad (46)$$

Therefore, since $R(v)$ is independent of X_1, it suffices to verify the Bernstein condition for the flow matching objective. Note that this is similar to the case of the denoising score matching objective, which was previously done by Yakovlev and Puchkin [2025]. The key technical step is formalized in the following lemma.

Lemma C.1. *Assume Assumption 3.2 holds, and let $f, \breve{f} : \mathbb{R}^D \times [0, 1 - t_0] \to \mathbb{R}^D$ be such that*

$$\|f\|_{L^\infty(\mathbb{R}^D \times [0,1-t_0])} \vee \|\breve{f}\|_{L^\infty(\mathbb{R}^D \times [0,1-t_0])} \leqslant M$$

for some $M > 0$. For all $(y, t) \in \mathbb{R}^D \times [0, 1 - t_0]$, consider the corresponding velocity fields

$$b(y, t) = c_1(t, \sigma)y + c_2(t, \sigma)f(y, t), \quad \breve{b}(y, t) = c_1(t, \breve{\sigma})y + c_2(t, \breve{\sigma})\breve{f}(y, t),$$

where $\sigma, \breve{\sigma} \in [\sigma_{\min}, 1]$, and the functions c_1 and c_2 are defined as in Proposition 3.3. Then, for every $y \in \mathbb{R}^D$, it follows that

$$\begin{aligned}
&(\ell_{FM}(b, y) - \ell_{FM}(\breve{b}, y))^2 \\
&\lesssim \left((\|y\|^2 + D + M^2)\sigma_{\min}^{-4} + D|\dot{\alpha}_t|^2_{W^{0,\infty}([0,1-t_0])} \right) \int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y} \|b(X_t, t) - \breve{b}(X_t, t)\|^2 \,\mathrm{d}t.
\end{aligned}$$

Furthermore, we have that

$$\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y}\|b(X_t,t)-\breve{b}(X_t,t)\|^2\,\mathrm{d}t \lesssim (\|y\|^2+D+M^2)\sigma_{\min}^{-4}.$$

and

$$\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y}\|\breve{b}(X_t,t)-\dot{X}_t\|^2\,\mathrm{d}t \lesssim (\|y\|^2+D+M^2)\sigma_{\min}^{-4}+D|\dot{\alpha}_t|^2_{W^{0,\infty}([0,1-t_0])}.$$

Here, the hidden constants in the inequalities are absolute.

The proof of Lemma C.1 can be found in Appendix C.1. With the aid of Lemma C.1, we show that the Bernstein condition holds for the excess loss class

$$\mathcal{L}=\{\ell_{SD}(v,\cdot)-\ell_{SD}(v^*,\cdot)\,:\,v\in\mathcal{V}(L,W,S,B,N,R,L_f)\}. \tag{47}$$

Notice that, according to Definition 3.5, we can set $M\asymp\sqrt{D}$ in Lemma C.1. Now fix some $\varkappa\geqslant 2$, which will be determined later in the proof. Then, in view of (46), Lemma C.1, and the Hölder inequality, we obtain that

$$\begin{aligned}\mathrm{Var}[\ell_{SD}(v,X_1)-\ell_{SD}(v^*,X_1)] &\lesssim \left[\int_0^{1-t_0}\|v_{t,t}-b^*(\cdot,t)\|^2_{L^2(\mathsf{p}_t^*)}\,\mathrm{d}t\right]^{1-1/\varkappa}\\ &\cdot\mathbb{E}^{1/\varkappa}\left[\left((\|X_1\|^2+D)\sigma_{\min}^{-4}+D\lambda^2(t_0)\right)^{\varkappa}\int_0^{1-t_0}\mathbb{E}_{X_t|X_1}\|v_{t,t}(X_t)-b^*(X_t,t)\|^2\,\mathrm{d}t\right].\end{aligned}$$

Here we introduced

$$\lambda(t_0)=|\alpha_t|_{W^{1,\infty}([0,1-t_0])}\vee e. \tag{48}$$

By the second part of Lemma C.1, we have that

$$\begin{aligned}&\mathrm{Var}[\ell_{SD}(v,X_1)-\ell_{SD}(v^*,X_1)]\\ &\lesssim\left[\int_0^{1-t_0}\|v_{t,t}-b^*(\cdot,t)\|^2_{L^2(\mathsf{p}_t^*)}\,\mathrm{d}t\right]^{1-1/\varkappa}\left[\mathbb{E}\left((\|X_1\|^2+D)\sigma_{\min}^{-4}+D\lambda^2(t_0)\right)^{\varkappa+1}\right]^{1/\varkappa}.\end{aligned}$$

By Assumption 3.1, it follows that

$$\left\|\|X_1\|^2\right\|_{\psi_1}\lesssim\left\|\|g^*(U)\|^2\right\|_{\psi_1}+\sigma_{\mathrm{data}}^2\left\|\|Z\|^2\right\|_{\psi_1}\lesssim 1+\sigma_{\mathrm{data}}^2 D. \tag{49}$$

Therefore, applying [Vershynin, 2018, Proposition 2.7.1], we obtain that

$$\begin{aligned}&\left[\mathbb{E}\left((\|X_1\|^2+D)\sigma_{\min}^{-4}+D\lambda^2(t_0)\right)^{\varkappa+1}\right]^{1/\varkappa}\\ &\lesssim(\varkappa+1)^{1+1/\varkappa}\left\{\left(\left\|\|X_1\|^2\right\|_{\psi_1}+D\right)\sigma_{\min}^{-4}+D\lambda^2(t_0)\right\}^{1+1/\varkappa}\\ &\lesssim\varkappa\left(D(\sigma_{\min}^{-4}+\lambda^2(t_0))\right)^{1+1/\varkappa}.\end{aligned}$$

Thus, we deduce that, for all $v\in\mathcal{V}(L,W,S,B,N,R,L_f)$,

$$\begin{aligned}&\mathrm{Var}[\ell_{SD}(v,X_1)-\ell_{SD}(v^*,X_1)]\\ &\lesssim\varkappa\left(D(\sigma_{\min}^{-4}+\lambda^2(t_0))\right)^{1+1/\varkappa}\left[\int_0^{1-t_0}\|v_{t,t}-b^*(\cdot,t)\|^2_{L^2(\mathsf{p}_t^*)}\,\mathrm{d}t\right]^{1-1/\varkappa}.\end{aligned}$$

Applying Proposition 2.1 and using the fact that $R(v)$ is non-negative, we conclude that

$$\begin{aligned}&\mathrm{Var}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\\ &\lesssim \varkappa \left(D(\sigma_{\min}^{-4} + \lambda^2(t_0))\right)^{1+1/\varkappa} \{\mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\}^{1-1/\varkappa}. \end{aligned} \tag{50}$$

Consequently, the Bernstein condition is satisfied.

Large deviation bound. Fix an arbitrary $v \in \mathcal{V}(L, W, S, B, N, R, L_f)$ and denote

$$\widehat{\mathbb{E}}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] = \frac{1}{n}\sum_{i=1}^{n} \left(\ell_{SD}(v, Y_i) - \ell_{SD}(v^*, Y_i)\right).$$

Recall that $Y_1, \ldots, Y_n$ are independent copies of X_1. We now provide a high-probability upper bound on

$$\left|\widehat{\mathbb{E}}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] - \mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\right|.$$

To achieve this, we invoke the Bernstein inequality for unbounded random variables [Lecué and Mitchell, 2012, Proposition 5.2]. This inequality states that, for any $\delta \in (0, 1)$, with probability at least $(1 - \delta/2)$,

$$\begin{aligned}&\left|\widehat{\mathbb{E}}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] - \mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\right|\\ &\lesssim \sqrt{\frac{\mathrm{Var}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\log(4/\delta)}{n}}\\ &+ \frac{\|\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)\|_{\psi_1} \log n \log(4/\delta)}{n}.\end{aligned}$$

We now derive an upper bound on the ψ_1-norm of the excess loss $\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)$. This bound is provided by the following lemma.

Lemma C.2. *Assume Assumptions 3.1 and 3.2 hold. Assume also $h \lesssim \sigma_{\min}^4$. Then, it follows that*

$$\sup_{v \in \mathcal{V}(L,W,S,B,N,R,L_f)} \|\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)\|_{\psi_1} \lesssim D\left(h^{-2}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\} + \lambda^2(t_0)\right).$$

The proof of Lemma C.2 is deferred to Appendix C.2. Therefore, setting

$$\varkappa \asymp \log n + \log \lambda(t_0) \tag{51}$$

and applying this lemma together with (50) yields, for each fixed candidate $v \in \mathcal{V}$,

$$\begin{aligned}&\left|\widehat{\mathbb{E}}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] - \mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\right|\\ &\lesssim \sqrt{\frac{C_b\{\mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\}^{1-1/\varkappa}\log(4/\delta)}{n}} + \frac{C_b \log(4/\delta)}{n},\end{aligned} \tag{52}$$

with probability at least $(1 - \delta/2)$. Here, we introduced

$$C_b = D^2 h^{-2}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\lambda^2(t_0)\log(\lambda(t_0))\log n. \tag{53}$$

Covering number evaluation. We now derive a uniform Bernstein-type bound for the excess loss class, as defined in (47). To achieve this, we need to evaluate the covering number of this class, which is established by the following lemma.

Lemma C.3. *Under Assumptions 3.1 and 3.2, fix $\varepsilon \in (0, 1)$. Let the parameters R and L_f in the class $\mathcal{V}(L, W, S, B, N, R, L_f)$ (see Definition 3.5) be given as in Theorem 3.6, and let the parameter $h \lesssim \sigma_{\min}^4$. Then, for every $\delta \in (0, 1)$, there exists a set $\mathcal{V}_\varepsilon \subseteq \mathcal{V}(L, W, S, B, N, R, L_f)$ such that*

$$\begin{aligned}&\sup_{v \in \mathcal{V}} \inf_{v' \in \mathcal{V}_\varepsilon} \left\{|\mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v', X_1)]| + |\widehat{\mathbb{E}}[\ell_{SD}(v, X_1) - \ell_{SD}(v', X_1)]|\right\}\\ &\lesssim h^{-2}\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}(D\log(1/\varepsilon))^{\mathcal{O}(1)}\log(n/\delta)\lambda(t_0)\varepsilon^{2\beta}\end{aligned}$$

holds with probability at least $(1 - \delta)$. Moreover, it holds that

$$\log|\mathcal{V}_\varepsilon| \lesssim SL\sigma_{\min}^{-2}h^{-1}\log(\varepsilon^{-1}L(B \vee 1)(\|W\|_\infty + 1)D) + SLh^{-1}\log(e \vee \log(n/\delta)).$$

The proof of Lemma C.3 is given in Appendix C.3. Define

$$\tau = h^{-2} \exp\{\mathcal{O}(\sigma_{\min}^{-4})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)} \log(n/\delta) \lambda(t_0) \varepsilon^{2\beta}. \tag{54}$$

Assume $h \lesssim \sigma_{\min}^4$ throughout the remainder of the proof. Then, with the aid of this lemma and the union bound, we obtain from (52) that, with probability at least $(1-\delta)$, the following inequality holds for all $v \in \mathcal{V}(L, W, S, B, N, R, L_f)$ simultaneously:

$$\begin{aligned}
&\left| \widehat{\mathbb{E}}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] - \mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)] \right| \\
&\lesssim \tau + \sqrt{\frac{C_b \{\tau + \mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\}^{1-1/\varkappa} \log(4|\mathcal{V}_\varepsilon|/\delta)}{n}} + \frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \\
&\lesssim \tau + \sqrt{\frac{C_b \{\mathbb{E}[\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)]\}^{1-1/\varkappa} \log(4|\mathcal{V}_\varepsilon|/\delta)}{n}} + \frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \\
&\quad + \left(\frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \right)^{\varkappa/(\varkappa+1)},
\end{aligned}$$

where the last inequality uses the Young inequality.

Deriving the final bound. In this step, we aim to establish a high-probability upper bound on the expected excess loss

$$\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)],$$

where X_1 is independent from the sample $Y_1, \ldots, Y_n$. Since $\widehat{v}$ minimizes the empirical risk $\widehat{\mathbb{E}}_{X_1} \ell_{SD}(v, X_1)$ with respect to $v \in \mathcal{V}(L, W, S, B, N, R, L_f)$, we deduce that

$$\begin{aligned}
\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] &\leqslant \mathbb{E}_{X_1}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] \\
&+ \left| \widehat{\mathbb{E}}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] - \mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] \right| \\
&+ \left| \widehat{\mathbb{E}}_{X_1}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] - \mathbb{E}_{X_1}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] \right|,
\end{aligned}$$

where $\widetilde{v} \in \mathcal{V}(L, W, S, B, N, R, L_f)$ is arbitrary. By Young's inequality, with probability at least $(1-\delta)$,

$$\begin{aligned}
&\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] \lesssim \mathbb{E}_{X_1}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] + \tau + \frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \\
&+ \sqrt{\frac{C_b \{\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)]\}^{1-1/\varkappa} \log(4|\mathcal{V}_\varepsilon|/\delta)}{n}} + \left(\frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \right)^{\varkappa/(\varkappa+1)}.
\end{aligned}$$

Using the fact that $x \leqslant a\sqrt{x^{1-1/\varkappa}} + b$ implies $x \leqslant 4a^{2\varkappa/(1+\varkappa)} + 2b$ for all non-negative a, b, and x, we deduce that

$$\begin{aligned}
\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] &\lesssim \mathbb{E}_{X_1}[\ell_{SD}(\widetilde{v}, X_1) - \ell_{SD}(v^*, X_1)] + \tau \\
&+ \left(\frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \right)^{\varkappa/(\varkappa+1)} + \frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n},
\end{aligned}$$

with probability at least $(1-\delta)$. We now utilize the approximation result provided by Theorem 3.6. Fix an arbitrary $\varepsilon \in (0,1)$ satisfying the conditions of Theorem 3.6, and let $\widetilde{v}$ be the corresponding approximation. Then, taking into account that $h^{-2} \asymp \varepsilon^{-\beta} \sigma_{\min}^{-4}$, we obtain from (54) that

$$\begin{aligned}
\tau &= h^{-2} \exp\{\mathcal{O}(\sigma_{\min}^{-4})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)} \log(n/\delta) \lambda(t_0) \varepsilon^{2\beta} \\
&\lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-4})\} (D \log(1/\varepsilon))^{\mathcal{O}(1)} \log(n/\delta) \lambda(t_0) \varepsilon^{\beta}.
\end{aligned}$$

Consequently, with probability at least $(1-\delta)$, it holds that

$$\begin{aligned}
&\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)] \lesssim (D \log(1/\varepsilon))^{\mathcal{O}(1)} \exp\{\mathcal{O}(\sigma_{\min}^{-4})\} \log(n/\delta) \lambda(t_0) \varepsilon^{\beta} \\
&+ \left(\frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n} \right)^{\varkappa/(\varkappa+1)} + \frac{C_b \log(4|\mathcal{V}_\varepsilon|/\delta)}{n}.
\end{aligned}$$

Now it remains to bound $\log(|\mathcal{V}_\varepsilon|)$. We must ensure that the approximation $\widetilde{v}$ belongs to $\mathcal{V}(L, W, S, B, N, R, L_f)$. Therefore, it suffices to take the architecture and the parameters L_f and R as specified in Theorem 3.6. Thus, by Lemma C.3, we have that

$$\begin{aligned}\log|\mathcal{V}_\varepsilon| &\lesssim \varepsilon^{-d} h^{-1} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D\log(1/\varepsilon))^{\mathcal{O}(1)}\log(n/\delta)\\ &\lesssim \varepsilon^{-d-\beta/2}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D\log(1/\varepsilon))^{\mathcal{O}(1)}\log(n/\delta).\end{aligned}$$

Since $\varkappa \gtrsim \log n$ due to (51), we deduce that, with probability at least $(1-\delta)$,

$$\begin{aligned}&\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)]\\ &\lesssim (D\log(1/\varepsilon))^{\mathcal{O}(1)}\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\log(n/\delta)\left\{\lambda(t_0)\varepsilon^\beta + \frac{(C_b\vee 1)\varepsilon^{-d-\beta/2}}{n}\right\}.\end{aligned}$$

From (51) and (53) we conclude that

$$\begin{aligned}&\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)]\\ &\lesssim (D\log(1/\varepsilon))^{\mathcal{O}(1)}\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\lambda(t_0)\left\{\varepsilon^\beta + \frac{\lambda(t_0)\varepsilon^{-d-3\beta/2}}{n}\right\}\log(\lambda(t_0))\log(4/\delta)\log^2 n,\end{aligned}$$

with probability at least $(1-\delta)$. Thus, setting $\varepsilon = (\lambda(t_0)/n)^{1/(d+5\beta/2)}$ yields that, with probability at least $(1-\delta)$,

$$\begin{aligned}&\mathbb{E}_{X_1}[\ell_{SD}(\widehat{v}, X_1) - \ell_{SD}(v^*, X_1)]\\ &\lesssim (D\log n)^{\mathcal{O}(1)}\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\lambda(t_0)\left(\frac{\lambda(t_0)}{n}\right)^{\frac{2\beta}{5\beta+2d}}\log(\lambda(t_0))\log(4/\delta).\end{aligned}$$

It now remains to ensure that this choice of ε satisfies the constraints of Theorem 3.6. This requires the sample size n to be sufficiently large, specifically

$$D^{3/2}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\log(nD\sigma_{\min}^{-2})\left(\frac{\lambda(t_0)}{n}\right)^{\frac{2(\beta\wedge 1)}{5\beta+2d}} \lesssim 1, \quad \left(\frac{\lambda(t_0)}{n}\right)^{\frac{2\beta}{5\beta+2d}} \lesssim \sigma_{\min}^4, \quad n \gtrsim \lambda(t_0).$$

Furthermore, Theorem 3.6 suggests that the architecture of the class $\mathcal{V}(L, W, S, B, N, R, L_f)$ is the following:

$$\begin{aligned}&L \lesssim \sigma_{\min}^{-2}\log(D\sigma_{\min}^{-2}\log n), \quad \|W\|_\infty \vee S \lesssim \left(\frac{n}{\lambda(t_0)}\right)^{\frac{2d}{5\beta+2d}}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D\log n)^{\mathcal{O}(1)},\\ &\log B \lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(D\log n)^{\mathcal{O}(1)}.\end{aligned}$$

Finally, the theorem also implies that

$$L_f \asymp (\sigma_{\min}^{-2} D\log n)^{\mathcal{O}(1)}, \quad N \asymp \left(\frac{n}{\lambda(t_0)}\right)^{\frac{\beta}{5\beta+2d}}\sigma_{\min}^{-2}, \quad R \asymp \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}D\log(nD\sigma_{\min}^{-2})$$

in Definition 3.5. The explicit smallness condition on ε^β and the sufficiently large constant in N ensure $h \leqslant c\sigma_{\min}^4$, as required. The proof is now complete.

□

C.1 Proof of Lemma C.1

We begin by decomposing the loss difference

$$\begin{aligned}\ell_{FM}(b, y) - \ell_{FM}(\breve{b}, y) = &\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y}\|b(X_t, t) - \breve{b}(X_t, t)\|^2\,\mathrm{d}t\\ &+ 2\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\langle b(X_t, t) - \breve{b}(X_t, t), \breve{b}(X_t, t) - \dot{X}_t\rangle\,\mathrm{d}t.\end{aligned}$$

Applying the Cauchy-Schwarz inequality yields

$$(\ell_{FM}(b,y)-\ell_{FM}(\breve{b},y))^2 \leqslant 2\left(\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\|b(X_t,t)-\breve{b}(X_t,t)\|^2\,\mathrm{d}t\right)^2 \tag{55}$$
$$+8\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\|b(X_t,t)-\breve{b}(X_t,t)\|^2\,\mathrm{d}t\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\|\breve{b}(X_t,t)-\dot{X}_t\|^2\,\mathrm{d}t.$$

Following this, by Proposition 3.3 and Assumption 3.2, we have that

$$\max_{\sigma_{\min}\leqslant\sigma\leqslant 1}\left(|c_1(\cdot,\sigma)|_{W^{0,\infty}([0,1-t_0])}\vee|c_2(\cdot,\sigma)|_{W^{0,\infty}([0,1-t_0])}\right)\lesssim\sigma_{\min}^{-2}.$$

Hence, it follows that

$$\begin{aligned}&\|b(X_t,t)-\breve{b}(X_t,t)\|^2\\&\leqslant 3\|X_t\|^2(c_1(t,\sigma)-c_1(t,\breve{\sigma}))^2+3c_2^2(t,\sigma)\|f(X_t,t)\|^2+3c_2^2(t,\breve{\sigma})\|\breve{f}(X_t,t)\|^2\\&\lesssim(\|X_t\|^2+M^2)\sigma_{\min}^{-4}.\end{aligned}$$

Using the moment bound

$$\mathbb{E}_{X_t|X_1=y}\|X_t\|^2\leqslant 2\|\beta_t y\|^2+2\mathbb{E}_{X_t|X_1=y}\|X_t-\beta_t y\|^2=2\beta_t^2\|y\|^2+2D\alpha_t^2,$$

we conclude that

$$\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\|b(X_t,t)-\breve{b}(X_t,t)\|^2\,\mathrm{d}t\lesssim(\|y\|^2+D+M^2)\sigma_{\min}^{-4}, \tag{56}$$

thereby establishing the second claim of the lemma. We next observe that

$$\mathbb{E}_{X_t|X_1=y}\|\dot{X}_t\|^2\leqslant 2\dot{\beta}_t^2\|y\|^2+2\dot{\alpha}_t^2 D.$$

Therefore, it follows that

$$\begin{aligned}&\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\|\breve{b}(X_t,t)-\dot{X}_t\|^2\,\mathrm{d}t\\&\leqslant 3\int_0^{1-t_0}\left(c_1^2(t,\breve{\sigma})\mathbb{E}_{X_t|X_1=y}\|X_t\|^2+c_2^2(t,\breve{\sigma})M^2+\mathbb{E}_{X_t|X_1=y}\|\dot{X}_t\|^2\right)\mathrm{d}t\\&\lesssim(\|y\|^2+D+M^2)\sigma_{\min}^{-4}+D|\dot{\alpha}_t|^2_{W^{0,\infty}([0,1-t_0])}.\end{aligned}$$

Consequently, the third claim is established. Combining this with (55) and (56), we conclude that

$$\begin{aligned}&(\ell_{FM}(b,y)-\ell_{FM}(\breve{b},y))^2\\&\lesssim\left((\|y\|^2+D+M^2)\sigma_{\min}^{-4}+D|\dot{\alpha}_t|^2_{W^{0,\infty}([0,1-t_0])}\right)\int_0^{1-t_0}\mathbb{E}_{X_t|X_1=y}\|b(X_t,t)-\breve{b}(X_t,t)\|^2\,\mathrm{d}t.\end{aligned}$$

The proof is finished.

□

C.2 Proof of Lemma C.2

For brevity, we suppress the full notation $\mathcal{V}(L,W,S,B,N,R,L_f)$ and simply write $\mathcal{V}$. It holds that

$$\sup_{v\in\mathcal{V}}\|\ell_{SD}(v,X_1)-\ell_{SD}(v^*,X_1)\|_{\psi_1}\lesssim\sup_{v\in\mathcal{V}}\|\ell_{FM}(v,X_1)-\ell_{FM}(v^*,X_1)\|_{\psi_1}+\sup_{v\in\mathcal{V}}R(v).$$

We proceed by analyzing each term individually. Applying the third claim of Lemma C.1 with $M \asymp \sqrt{D}$, we obtain that

$$\begin{aligned}\sup_{v\in\mathcal{V}} \|\ell_{FM}(v, X_1) - \ell_{FM}(v^*, X_1)\|_{\psi_1} &\leqslant \sup_{v\in\mathcal{V}} \|\ell_{FM}(v, X_1)\|_{\psi_1} + \sup_{v\in\mathcal{V}} \|\ell_{FM}(v^*, X_1)\|_{\psi_1} \\ &\lesssim \left(\left\|\|X_1\|^2\right\|_{\psi_1} + D\right)\sigma_{\min}^{-4} + D\lambda^2(t_0). \end{aligned} \tag{57}$$

Using (49), we deduce that

$$\sup_{v\in\mathcal{V}} \|\ell_{FM}(v, X_1) - \ell_{FM}(v^*, X_1)\|_{\psi_1} \lesssim D(\sigma_{\min}^{-4} + \lambda^2(t_0)). \tag{58}$$

We now analyze the consistency penalty term. Let $\Psi_{0,t}(y) = y + t\, v_{0,t}(y)$ be as in Definition 3.5. Then it follows that

$$\begin{aligned} R(v) &\lesssim \int_0^{1-t_0} \mathbb{E}\|\partial_t \Psi_{0,t}(X_0)\|^2 \,\mathrm{d}t + \int_0^{1-t_0} \mathbb{E}\|v_{t,t}(\Psi_{0,t}(X_0))\|^2 \,\mathrm{d}t \\ &\lesssim \int_0^{1-t_0} h^{-2}\mathbb{E} \max_{0\leqslant k\leqslant N} \|Z_k\|^2 \,\mathrm{d}t + \int_0^{1-t_0} (c_1(t,\sigma)^2 \mathbb{E}\|\Psi_{0,t}(X_0)\|^2 + c_2(t,\sigma)^2 D)\,\mathrm{d}t. \end{aligned}$$

In view of Proposition 3.3, we have that

$$|c_j(\cdot,\sigma)|_{W^{0,\infty}([0,1-t_0])} \lesssim \sigma_{\min}^{-2}, \quad j \in \{1,2\}.$$

This implies that

$$R(v) \lesssim \int_0^{1-t_0} \left((h^{-2} + \sigma_{\min}^{-4})\mathbb{E} \max_{0\leqslant k\leqslant N} \|Z_k\|^2 + \sigma_{\min}^{-4} D\right) \mathrm{d}t.$$

Applying the second claim of Lemma B.5 with $M \asymp \sigma_{\min}^{-2}$ and $M_f = \sigma_{\min}^{-2}\sqrt{D}$, we conclude that

$$\sup_{v\in\mathcal{V}} R(v) \lesssim Dh^{-2}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}.$$

Combining this with (57) and (58), we obtain that

$$\sup_{v\in\mathcal{V}} \|\ell_{SD}(v, X_1) - \ell_{SD}(v^*, X_1)\|_{\psi_1} \lesssim D\left(h^{-2}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\} + \lambda^2(t_0)\right).$$

This completes the proof.

□

C.3 Proof of Lemma C.3

Let $\mathcal{B}(0, \widetilde{R}) \subseteq \mathbb{R}^D$, where $\widetilde{R} > 0$ will be chosen later in the proof, and consider average velocity field candidates $v^{(1)}, v^{(2)} \in \mathcal{V}(L, W, S, B, N, R, L_f)$ with corresponding

$$f^{(j)}(y,t), \quad \{f_k^{(j)}(y,t) : 0 \leqslant k \leqslant N\}, \quad \sigma_{\min} \leqslant \sigma^{(j)} \leqslant 1, \quad j \in \{1,2\}$$

as in Definition 3.5. Then, we have that

$$|\ell_{SD}(v^{(1)}, y) - \ell_{SD}(v^{(2)}, y)| \leqslant |\ell_{FM}(v^{(1)}, y) - \ell_{FM}(v^{(2)}, y)| + |R(v^{(1)}) - R(v^{(2)})|. \tag{59}$$

We proceed by examining each term individually.

The distance between flow matching losses. Applying Lemma C.1 with $M \asymp \sqrt{D}$, we obtain that for all $y \in \mathbb{R}^D$,

$$\begin{aligned} &|\ell_{FM}(v^{(1)}, y) - \ell_{FM}(v^{(2)}, y)| \\ &\lesssim \left(\sigma_{\min}^{-2}(\|y\| + \sqrt{D}) + \sqrt{D}\lambda(t_0)\right) \left(\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y}\|v_{t,t}^{(1)}(X_t) - v_{t,t}^{(2)}(X_t)\|^2 \,\mathrm{d}t\right)^{1/2}. \end{aligned} \tag{60}$$

Furthermore, Lemma C.1 implies that, for all $y \in \mathbb{R}^D$,

$$|\ell_{FM}(v^{(1)}, y) - \ell_{FM}(v^{(2)}, y)| \lesssim \left(\sigma_{\min}^{-2}(\|y\| + \sqrt{D}) + \sqrt{D}\lambda(t_0)\right) \sigma_{\min}^{-2}(\|y\| + \sqrt{D}). \tag{61}$$

We now note that

$$\begin{aligned}
&\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y} \|v_{t,t}^{(1)}(X_t) - v_{t,t}^{(2)}(X_t)\|^2 \, \mathrm{d}t \\
&\lesssim \int_0^{1-t_0} (c_2(t, \sigma^{(1)}) - c_2(t, \sigma^{(2)}))^2 \|f^{(1)}\|^2_{L^\infty(\mathbb{R}^D \times [0, 1-t_0])} \, \mathrm{d}t \\
&+ \int_0^{1-t_0} (c_1(t, \sigma^{(1)}) - c_1(t, \sigma^{(2)}))^2 \mathbb{E}_{X_t|X_1=y} \|X_t\|^2 \, \mathrm{d}t \\
&+ \int_0^{1-t_0} c_2^2(t, \sigma^{(2)}) \mathbb{E}_{X_t|X_1=y} \|f^{(1)}(X_t, t) - f^{(2)}(X_t, t)\|^2 \, \mathrm{d}t.
\end{aligned}$$

Next, from Proposition 3.3 we find that, for all $0 \leqslant t \leqslant 1 - t_0$,

$$|c_j(t, \sigma^{(1)}) - c_j(t, \sigma^{(2)})| \lesssim \sigma_{\min}^{-4} |\sigma^{(1)} - \sigma^{(2)}|, \quad |c_j(t, \sigma^{(2)})| \lesssim \sigma_{\min}^{-2}, \quad j \in \{1, 2\}. \tag{62}$$

Leveraging the bound $\mathbb{E}_{X_t|X_1=y} \|X_t\|^2 \lesssim \alpha_t^2 D + \beta_t^2 \|y\|^2$, we obtain

$$\begin{aligned}
&\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y} \|v_{t,t}^{(1)}(X_t) - v_{t,t}^{(2)}(X_t)\|^2 \, \mathrm{d}t \lesssim (D + \|y\|^2) \sigma_{\min}^{-8} (\sigma^{(1)} - \sigma^{(2)})^2 \\
&\quad + \sigma_{\min}^{-4} \|f^{(1)} - f^{(2)}\|^2_{L^\infty(\mathcal{B}(0, \widetilde{R}) \times [0, 1-t_0])} + \sigma_{\min}^{-4} D \max_{0 \leqslant t \leqslant 1-t_0} \mathbb{P}(\|X_t\| \geqslant \widetilde{R} \,|\, X_1 = y).
\end{aligned}$$

The Gaussian concentration bound in Lemma E.1 implies that choosing the radius $\widetilde{R}$ satisfying

$$\widetilde{R} \asymp \sqrt{\log(1/\varepsilon)} + \sqrt{D} \tag{63}$$

ensures that

$$\begin{aligned}
&\int_0^{1-t_0} \mathbb{E}_{X_t|X_1=y} \|v_{t,t}^{(1)}(X_t) - v_{t,t}^{(2)}(X_t)\|^2 \, \mathrm{d}t \lesssim (D + \|y\|^2) \sigma_{\min}^{-8} (\sigma^{(1)} - \sigma^{(2)})^2 \\
&\quad + \sigma_{\min}^{-4} \|f^{(1)} - f^{(2)}\|^2_{L^\infty(\mathcal{B}(0, \widetilde{R} + \|y\|) \times [0, 1-t_0])} + \sigma_{\min}^{-4} D \varepsilon^{4\beta}.
\end{aligned}$$

Substituting this into (60), we conclude that, for all $y \in \mathbb{R}^D$,

$$\begin{aligned}
&|\ell_{FM}(v^{(1)}, y) - \ell_{FM}(v^{(2)}, y)| \lesssim \left(\sigma_{\min}^{-2}(\|y\| + \sqrt{D}) + \sqrt{D}\lambda(t_0)\right) \\
&\cdot \left(\sigma_{\min}^{-4}(\sqrt{D} + \|y\|)|\sigma^{(1)} - \sigma^{(2)}| + \sigma_{\min}^{-2} \|f^{(1)} - f^{(2)}\|_{L^\infty(\mathcal{B}(0, \widetilde{R} + \|y\|) \times [0, 1-t_0])} + \sigma_{\min}^{-2} \sqrt{D} \varepsilon^{2\beta}\right).
\end{aligned}$$

By applying Gaussian concentration again, we deduce that setting

$$R_{\text{data}} \asymp 1 + \sigma_{\text{data}} \sqrt{\log(n/\delta) + \log(1/\varepsilon)} + \sigma_{\text{data}} \sqrt{D} \tag{64}$$

ensures that, with probability at least $(1 - \delta)$, the following inequality holds:

$$\begin{aligned}
&\|\ell_{FM}(v^{(1)}, \cdot) - \ell_{FM}(v^{(2)}, \cdot)\|_{L^\infty(\mathcal{B}(0, R_{\text{data}}))} \lesssim \sigma_{\min}^{-6} \left(\sqrt{\log(n\varepsilon^{-1}/\delta)} + \sqrt{D}\lambda(t_0)\right) \qquad (65) \\
&\cdot \left((\sqrt{\log(n\varepsilon^{-1}/\delta)} + \sqrt{D})|\sigma^{(1)} - \sigma^{(2)}| + \|f^{(1)} - f^{(2)}\|_{L^\infty(\mathcal{B}(0, \widetilde{R} + R_{\text{data}}) \times [0, 1-t_0])} + \sqrt{D} \varepsilon^{2\beta}\right).
\end{aligned}$$

By Assumption 3.1, we have the bound $\mathbb{E}^{1/2}(\|X_1\| + \sqrt{D})^4 \lesssim D$. Hence, using (61), we obtain that

$$\begin{aligned}
&|\mathbb{E}\ell_{FM}(v^{(1)}, X_1) - \mathbb{E}\ell_{FM}(v^{(2)}, X_1)| \\
&\lesssim \|\ell_{FM}(v^{(1)}, \cdot) - \ell_{FM}(v^{(2)}, \cdot)\|_{L^\infty(\mathcal{B}(0,R_{\text{data}}))} + \varepsilon^{2\beta}\sqrt{\mathbb{E}(\ell_{FM}(v^{(1)}, X_1) - \ell_{FM}(v^{(2)}, X_1))^2} \\
&\lesssim \|\ell_{FM}(v^{(1)}, \cdot) - \ell_{FM}(v^{(2)}, \cdot)\|_{L^\infty(\mathcal{B}(0,R_{\text{data}}))} + \varepsilon^{2\beta}\sigma_{\min}^{-2} D(\sigma_{\min}^{-2} + \sqrt{D}\lambda(t_0)). \qquad (66)
\end{aligned}$$

The distance between consistency penalties. We now analyze the proximity between the penalty functions $R(v^{(1)})$ and $R(v^{(2)})$. Introduce $\Psi_{0,t}^{(j)}(y) = y + t\, v_{0,t}^{(j)}(y)$ as in Definition 3.5. By Cauchy-Schwarz, we have that

$$\begin{aligned}
&|R(v^{(1)}) - R(v^{(2)})| \\
&\leqslant \left(\int_0^{1-t_0} \mathbb{E}\|\partial_t[\Psi_{0,t}^{(1)}(X_0) - \Psi_{0,t}^{(2)}(X_0)] - v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) + v_{t,t}^{(2)}(\Psi_{0,t}^{(2)}(X_0))\|^2 \,\mathrm{d}t \right)^{1/2} \qquad (67) \\
&\quad \cdot \left(\int_0^{1-t_0} \mathbb{E}\|\partial_t[\Psi_{0,t}^{(1)}(X_0) + \Psi_{0,t}^{(2)}(X_0)] - v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) - v_{t,t}^{(2)}(\Psi_{0,t}^{(2)}(X_0))\|^2 \,\mathrm{d}t \right)^{1/2}.
\end{aligned}$$

By Lemma B.4, we have that

$$\int_0^{1-t_0} \mathbb{E}\|\partial_t[\Psi_{0,t}^{(1)}(X_0) - \Psi_{0,t}^{(2)}(X_0)]\|^2 \,\mathrm{d}t \lesssim h^{-2} \int_0^{1-t_0} \mathbb{E} \max_{0\leqslant k\leqslant N} \|Z_k^{(1)} - Z_k^{(2)}\|^2 \,\mathrm{d}t, \qquad (68)$$

where $Z_k^{(j)}$ is defined in Definition 3.5 with $\{f_k^{(j)} : 0 \leqslant k \leqslant N\}$ and $Z_0^{(j)} = X_0$ for $j \in \{1, 2\}$. In addition, it holds that

$$\begin{aligned}
&\int_0^{1-t_0} \mathbb{E}\|v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) - v_{t,t}^{(2)}(\Psi_{0,t}^{(2)}(X_0))\|^2 \,\mathrm{d}t \lesssim \int_0^{1-t_0} \mathbb{E}\|v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) - v_{t,t}^{(2)}(\Psi_{0,t}^{(1)}(X_0))\|^2 \,\mathrm{d}t \\
&\quad + \sigma_{\min}^{-4}(1 + L_f^2) \int_0^{1-t_0} \mathbb{E}\|\Psi_{0,t}^{(1)}(X_0) - \Psi_{0,t}^{(2)}(X_0)\|^2 \,\mathrm{d}t. \qquad (69)
\end{aligned}$$

Applying Lemma B.5 with $M \asymp \sigma_{\min}^{-2}$ and $M_f \asymp \sigma_{\min}^{-2}\sqrt{D}$ together with Lemma B.4, we deduce that

$$\|\Psi_{0,t}^{(j)}(y)\| \lesssim \max_{0\leqslant k\leqslant N} \|Z_k^{(j)}\| \lesssim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(1 + \|y\|), \quad \text{for all } y \in \mathbb{R}^D,\ j \in \{1, 2\}. \qquad (70)$$

Furthermore, by Lemma B.5, we have that, for all X_0 satisfying

$$\|X_0\| \lesssim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\left(\log(1/\varepsilon) + \log(D\sigma_{\min}^{-2})\right), \qquad (71)$$

it follows that $\|\Psi_{0,t}^{(j)}(X_0)\| \leqslant R$ for every $0 \leqslant t \leqslant 1 - t_0$ and $j \in \{1, 2\}$. Then, using (62), we obtain that

$$\begin{aligned}
&\int_0^{1-t_0} \mathbb{E}\|v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) - v_{t,t}^{(2)}(\Psi_{0,t}^{(1)}(X_0))\|^2 \,\mathrm{d}t \lesssim \sigma_{\min}^{-8}(\sigma^{(1)} - \sigma^{(2)})^2 \int_0^{1-t_0} \mathbb{E}\|\Psi_{0,t}^{(1)}(X_0)\|^2 \,\mathrm{d}t \\
&\quad + D\sigma_{\min}^{-8}(\sigma^{(1)} - \sigma^{(2)})^2 + \sigma_{\min}^{-4}\|f^{(1)} - f^{(2)}\|_{L^\infty(\mathcal{B}(0,R)\times[0,1-t_0])}^2 \\
&\quad + \sigma_{\min}^{-4} D \cdot \mathbb{P}(\|X_0\| \gtrsim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\log(1/\varepsilon)).
\end{aligned}$$

Using the fact that

$$\int_0^{1-t_0} \mathbb{E}\|\Psi_{0,t}^{(1)}(X_0) - \Psi_{0,t}^{(2)}(X_0)\|^2 \,\mathrm{d}t \lesssim \int_0^{1-t_0} \mathbb{E} \max_{0\leqslant k\leqslant N} \|Z_k^{(1)} - Z_k^{(2)}\|^2 \,\mathrm{d}t,$$

we conclude from (68), (69), and (70) that

$$\int_0^{1-t_0} \mathbb{E}\|\partial_t[\Psi_{0,t}^{(1)}(X_0) - \Psi_{0,t}^{(2)}(X_0)] - v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) + v_{t,t}^{(2)}(\Psi_{0,t}^{(2)}(X_0))\|^2 \,\mathrm{d}t$$

$$\lesssim (h^{-2} + \sigma_{\min}^{-4}(1+L_f^2)) \int_0^{1-t_0} \mathbb{E} \max_{0\leqslant k\leqslant N} \|Z_k^{(1)} - Z_k^{(2)}\|^2 \,\mathrm{d}t \tag{72}$$

$$+ \sigma_{\min}^{-8}(\sigma^{(1)} - \sigma^{(2)})^2 D^2 \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} + \sigma_{\min}^{-4}\|f^{(1)} - f^{(2)}\|_{L^\infty(\mathcal{B}(0,R)\times[0,1-t_0])}^2 + \varepsilon^{4\beta}.$$

Similarly, from (62) we have that

$$\int_0^{1-t_0} \mathbb{E}\|\partial_t[\Psi_{0,t}^{(1)}(X_0) + \Psi_{0,t}^{(2)}(X_0)] - v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) - v_{t,t}^{(2)}(\Psi_{0,t}^{(2)}(X_0))\|^2 \,\mathrm{d}t$$

$$\lesssim \sum_{j=1}^{2} \int_0^{1-t_0} \mathbb{E}\left[\|\partial_t \Psi_{0,t}^{(j)}(X_0)\|^2 + \sigma_{\min}^{-4}\|\Psi_{0,t}^{(j)}(X_0)\|^2 + \sigma_{\min}^{-4} D\right] \mathrm{d}t.$$

Proceeding similarly to (70), we deduce that, for all $j \in \{1,2\}$,

$$\|\partial_t \Psi_{0,t}^{(j)}(y)\| \lesssim h^{-1} \max_{0\leqslant k\leqslant N} \|Z_k^{(j)}\| \lesssim h^{-1}\sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}(1+\|y\|), \quad \text{for all } y \in \mathbb{R}^D.$$

This bound together with (70) implies that

$$\int_0^{1-t_0} \mathbb{E}\|\partial_t[\Psi_{0,t}^{(1)}(X_0) + \Psi_{0,t}^{(2)}(X_0)] - v_{t,t}^{(1)}(\Psi_{0,t}^{(1)}(X_0)) - v_{t,t}^{(2)}(\Psi_{0,t}^{(2)}(X_0))\|^2 \,\mathrm{d}t$$

$$\lesssim h^{-2} D^2 \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}.$$

Using the above bound together with (67) and (72), we conclude that

$$|R(v^{(1)}) - R(v^{(2)})| \lesssim h^{-1} D \exp\{\mathcal{O}(\sigma_{\min}^{-2})\} \Bigg\{ (h^{-1} + L_f) \left[\int_0^{1-t_0} \mathbb{E} \max_{0\leqslant k\leqslant N} \|Z_k^{(1)} - Z_k^{(2)}\|^2 \,\mathrm{d}t \right]^{1/2}$$

$$+ D\sigma_{\min}^{-4} \exp\{\mathcal{O}(\sigma_{\min}^{-2})\}|\sigma^{(1)} - \sigma^{(2)}| + \sigma_{\min}^{-2}\|f^{(1)} - f^{(2)}\|_{L^\infty(\mathcal{B}(0,R)\times[0,1-t_0])} + \varepsilon^{2\beta} \Bigg\}. \tag{73}$$

The distance between $Z_k^{(1)}$ and $Z_k^{(2)}$. For $j \in \{1,2\}$, write

$$b_k^{(j)}(z) = c_1(\tau_k, \sigma^{(j)})z + c_2(\tau_k, \sigma^{(j)}) f_k^{(j)}(z), \qquad e_k = \|Z_k^{(1)} - Z_k^{(2)}\|.$$

Let $\Delta_\sigma = |\sigma^{(1)} - \sigma^{(2)}|$ and $\Delta_f = \max_k \|f_k^{(1)} - f_k^{(2)}\|_{L^\infty(\mathcal{B}(0,R))}$. On $\mathcal{B}(0,R)$ the velocity fields $b_k^{(j)}$ have Lipschitz constant $\ell \asymp \sigma_{\min}^{-4}$, and (62) yields

$$\|b_k^{(1)}(z) - b_k^{(2)}(z)\| \lesssim \sigma_{\min}^{-4}\left((\|z\| + \sqrt{D})\Delta_\sigma + \Delta_f\right).$$

For convenience, let $\eta(z)$ denote the right-hand side of this estimate. For initial points satisfying (71), choose the radius constants so that all sequences lie in $\mathcal{B}(0,R)$. We have that

$$\|Y_{k+1}^{(1)} - Y_{k+1}^{(2)}\| \leqslant (1+h\ell)e_k + h\eta(Z_k^{(1)}).$$

The corrected Heun update then implies

$$e_{k+1} \leqslant (1 + h\ell + h^2\ell^2/2)e_k + \frac{h}{2}(1+h\ell)\eta(Z_k^{(1)}) + \frac{h}{2}\eta(Y_{k+1}^{(1)})$$

$$\leqslant (1 + 3h\ell/2)e_k + h\eta(Z_k^{(1)}) + \frac{h}{2}\eta(Y_{k+1}^{(1)}),$$

where $h\ell \leqslant 1$. The growth part of Lemma B.5 bounds both terms by $C\sqrt{D}\exp\{C\sigma_{\min}^{-2}\}(1+\|X_0\|)$. Consequently, discrete Grönwall lemma, applied with coefficient $1 + 3h\ell/2$ and $e_0 = 0$, yields

$$\max_{0\leqslant k\leqslant N} e_k \lesssim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\Delta_\sigma(1 + \|X_0\|) + \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\Delta_f.$$

Using Lemma B.5, we obtain that

$$\begin{aligned}\int_0^{1-t_0} \mathbb{E}\max_{0\leqslant k\leqslant N}\|Z_k^{(1)} - Z_k^{(2)}\|^2\,\mathrm{d}t &\lesssim D\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}(\sigma^{(1)} - \sigma^{(2)})^2R^2\\ &+ \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\max_{0\leqslant l\leqslant N}\|f_l^{(1)} - f_l^{(2)}\|^2_{L^\infty(\mathcal{B}(0,R))}\\ &+ D\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\int_0^{1-t_0}\mathbb{E}[(1+\|X_0\|^2)\mathbb{1}(\|X_0\| \gtrsim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\log(1/\varepsilon))]\,\mathrm{d}t.\end{aligned}$$

Gaussian concentration bound in Lemma E.1 in conjunction with the Cauchy-Schwarz inequality implies that

$$\begin{aligned}\int_0^{1-t_0} \mathbb{E}\max_{0\leqslant k\leqslant N}\|Z_k^{(1)} - Z_k^{(2)}\|^2\,\mathrm{d}t &\lesssim D\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}(\sigma^{(1)} - \sigma^{(2)})^2R^2\\ &+ \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\max_{0\leqslant l\leqslant N}\|f_l^{(1)} - f_l^{(2)}\|^2_{L^\infty(\mathcal{B}(0,R))} + \varepsilon^{4\beta},\end{aligned}$$

Substituting the bound for R from the premise of the lemma, we obtain that

$$\begin{aligned}\left(\int_0^{1-t_0} \mathbb{E}\max_{0\leqslant k\leqslant N}\|Z_k^{(1)} - Z_k^{(2)}\|^2\,\mathrm{d}t\right)^{1/2} &\lesssim D^{3/2}\log(1/\varepsilon)\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}|\sigma^{(1)} - \sigma^{(2)}| + \varepsilon^{2\beta}\\ &+ \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\max_{0\leqslant l\leqslant N}\|f_l^{(1)} - f_l^{(2)}\|_{L^\infty(\mathcal{B}(0,R))}.\end{aligned}$$

Therefore, using (73) and the expression for L_f given by the initial conditions of the lemma, we conclude that

$$\begin{aligned}&|R(v^{(1)}) - R(v^{(2)})| \lesssim h^{-2}\exp\mathcal{O}(\sigma_{\min}^{-4})(D\log(1/\varepsilon))^{\mathcal{O}(1)} \qquad (74)\\ &\cdot\left(|\sigma^{(1)} - \sigma^{(2)}| + \|f^{(1)} - f^{(2)}\|_{L^\infty(\mathcal{B}(0,R)\times[0,1-t_0])} + \max_{0\leqslant l\leqslant N}\|f_l^{(1)} - f_l^{(2)}\|_{L^\infty(\mathcal{B}(0,R))} + \varepsilon^{2\beta}\right).\end{aligned}$$

Covering number evaluation. The following result evaluates the covering number of the class of feedforward neural networks with GELU activation function.

Lemma C.4 (Yakovlev et al. [2025], Lemma B.1, adapted)**.** *For every $R > 0$ and every*

$$0 < \tau \leqslant \sup\{\|f - g\|_{L^\infty([-R,R]^D)} \,:\, f, g \in \mathsf{NN}(L, W, S, B)\},$$

the covering number satisfies the bound

$$\log\mathcal{N}(\tau, \mathsf{NN}(L, W, S, B), \|\cdot\|_{L^\infty([-R,R]^D)}) \lesssim SL\log(\tau^{-1}L(B\vee 1)(\|W\|_\infty + 1)(R\vee 1)).$$

Let $\mathcal{F}$ be the subclass of $\mathsf{NN}(L, W, S, B)$ satisfying the two constraints and the global Lipschitz bound in Definition 3.5. Use the compact domain $\mathcal{B}(0, R + \widetilde{R} + R_{\mathrm{data}})$. Apply Lemma C.4 at half the desired radius. For each covering ball that meets $\mathcal{F}$, choose a representative from its intersection with $\mathcal{F}$. This yields a net $\mathcal{F}_{\varepsilon,k} \subseteq \mathcal{F}$ of radius $\sqrt{D}\varepsilon^{2\beta}$, up to a fixed constant, and

$$\log|\mathcal{F}_{\varepsilon,k}| \lesssim SL\log\big(\varepsilon^{-1}L(B\vee 1)(\|W\|_\infty + 1)D\big) + SL\sigma_{\min}^{-2} + SL\log(e\vee\log(n/\delta)). \qquad (75)$$

Here, the parameter R is taken from the initial conditions of the lemma. Now let $\mathcal{G}_\varepsilon$ be an $\varepsilon^{2\beta}$-net of the interval $[\sigma_{\min}, 1]$. Evidently, $|\mathcal{G}_\varepsilon| \lesssim \varepsilon^{-2\beta}$. Choose the scalar net inside $[\sigma_{\min}, 1]$ and form $\mathcal{V}_\varepsilon$

from its product with the constrained subnetwork nets. Every resulting candidate lies in $\mathcal{V}$. From (75) it follows that

$$\begin{aligned}\log|\mathcal{V}_\varepsilon| &\leqslant \log|\mathcal{G}_\varepsilon| + \sum_{k=0}^{N}\log|\mathcal{F}_{k,\varepsilon}| \\ &\lesssim \sigma_{\min}^{-2}(N+1)SL\log(\varepsilon^{-1}L(B\vee 1)(\|W\|_\infty+1)D) + (N+1)SL\log(e\vee\log(n/\delta)).\end{aligned}$$

By Lemma B.4, we have that

$$\|f^{(1)}-f^{(2)}\|_{L^\infty(\mathcal{B}(0,R+\widetilde{R}+R_{\text{data}})\times[0,1-t_0])} \lesssim \max_{0\leqslant k\leqslant N}\|f_k^{(1)}-f_k^{(2)}\|_{L^\infty(\mathcal{B}(0,R+\widetilde{R}+R_{\text{data}}))}. \tag{76}$$

Combining this with (65), we obtain that, with probability at least $(1-\delta)$,

$$\sup_{v\in\mathcal{V}}\inf_{v'\in\mathcal{V}_\varepsilon}\|\ell_{FM}(v,\cdot)-\ell_{FM}(v',\cdot)\|_{L^\infty(\mathcal{B}(0,R_{\text{data}}))} \lesssim (D\sigma_{\min}^{-2})^{\mathcal{O}(1)}\left(\log(n/\delta)+\log(1/\varepsilon)\right)\lambda(t_0)\varepsilon^{2\beta}.$$

Similarly, (66) yields that

$$\sup_{v\in\mathcal{V}}\inf_{v'\in\mathcal{V}_\varepsilon}|\mathbb{E}\ell_{FM}(v,X_1)-\mathbb{E}\ell_{FM}(v',X_1)| \lesssim (D\sigma_{\min}^{-2})^{\mathcal{O}(1)}\left(\log(n/\delta)+\log(1/\varepsilon)\right)\lambda(t_0)\varepsilon^{2\beta}.$$

Finally, from (74) and (76) we deduce that

$$\sup_{v\in\mathcal{V}}\inf_{v'\in\mathcal{V}_\varepsilon}|R(v)-R(v')| \lesssim h^{-2}\exp\{\mathcal{O}(\sigma_{\min}^{-4})\}(D\log(1/\varepsilon))^{\mathcal{O}(1)}\varepsilon^{2\beta}.$$

Since $N+1\lesssim h^{-1}$, the entropy bound follows, and (59) proves the lemma.

□

D Proof of Theorem 3.8

Let $\Psi_{0,t}(y)=y+t\cdot v_{0,t}(y)$, where $v\in\mathcal{V}(L,W,S,B,N,R,L_f)$ is arbitrary. It holds that

$$W_2(\text{Law}(\Psi_{0,1-t_0}(X_0)),\mathsf{p}^*) \leqslant W_2(\text{Law}(\Psi_{0,1-t_0}(X_0)),\mathsf{p}^*_{1-t_0}) + W_2(\mathsf{p}^*_{1-t_0},\mathsf{p}^*). \tag{77}$$

We now bound each term individually. For every $t\in[0,1-t_0]$ define $E_t=\mathbb{E}\|\Psi^*_{0,t}(X_0)-\Psi_{0,t}(X_0)\|^2$. Then, it follows that

$$\begin{aligned}\frac{\mathrm{d}E_t}{\mathrm{d}t} &= 2\mathbb{E}\langle\Psi^*_{0,t}(X_0)-\Psi_{0,t}(X_0), b^*(\Psi^*_{0,t}(X_0),t)-\partial_t\Psi_{0,t}(X_0)\rangle \\ &= 2\mathbb{E}\langle\Psi^*_{0,t}(X_0)-\Psi_{0,t}(X_0), b^*(\Psi^*_{0,t}(X_0),t)-v_{t,t}(\Psi^*_{0,t}(X_0))\rangle \\ &\quad + 2\mathbb{E}\langle\Psi^*_{0,t}(X_0)-\Psi_{0,t}(X_0), v_{t,t}(\Psi^*_{0,t}(X_0))-v_{t,t}(\Psi_{0,t}(X_0))\rangle \\ &\quad + 2\mathbb{E}\langle\Psi^*_{0,t}(X_0)-\Psi_{0,t}(X_0), v_{t,t}(\Psi_{0,t}(X_0))-\partial_t\Psi_{0,t}(X_0)\rangle.\end{aligned}$$

Applying the Cauchy-Schwarz inequality and the Young inequality, we obtain

$$\begin{aligned}\frac{\mathrm{d}E_t}{\mathrm{d}t} &\leqslant 2E_t + \mathbb{E}\|b^*(\Psi^*_{0,t}(X_0),t)-v_{t,t}(\Psi^*_{0,t}(X_0))\|^2 + \mathbb{E}\|v_{t,t}(\Psi_{0,t}(X_0))-\partial_t\Psi_{0,t}(X_0)\|^2 \\ &\quad + 2\mathbb{E}\left[\|\Psi^*_{0,t}(X_0)-\Psi_{0,t}(X_0)\|\cdot\|v_{t,t}(\Psi^*_{0,t}(X_0))-v_{t,t}(\Psi_{0,t}(X_0))\|\right].\end{aligned} \tag{78}$$

Using Lemma B.5 with $M\asymp\sigma_{\min}^{-2}$ and $M_f\asymp\sigma_{\min}^{-2}\sqrt{D}$ and Lemma B.4, we have that, for all X_0 satisfying

$$\|X_0\|\lesssim\sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\log(nD\sigma_{\min}^{-2}),$$

it follows that $\|\Psi^*_{0,t}(X_0)\|\vee\|\Psi_{0,t}(X_0)\|\leqslant R$ for every $0\leqslant t\leqslant 1-t_0$. Furthermore, we also have that

$$\|\Psi_{0,t}(y)\|\vee\|\Psi^*_{0,t}(y)\|\lesssim\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\sqrt{D}(1+\|y\|),\quad\text{for all } y\in\mathbb{R}^D. \tag{79}$$

For $z,z'\in\mathcal{B}(0,R)$, the linear term and the spline coefficients yield

$$\begin{aligned}\|v_{t,t}(z)-v_{t,t}(z')\| &\lesssim \sigma_{\min}^{-2}\|z-z'\|+\sigma_{\min}^{-2}\max_k\|f_k(z)-f_k(z')\| \\ &\lesssim \sigma_{\min}^{-4}\|z-z'\|.\end{aligned}$$

Here the fixed constants account for the extrapolated boundary coefficients, as in Lemma B.4. Apply this bound with $z = \Psi_{0,t}(X_0)$ and $z' = \Psi^*_{0,t}(X_0)$ on the event above. Hence, using the fact that $v_{t,t}$ is globally Lipschitz, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi^*_{0,t}(X_0) - \Psi_{0,t}(X_0)\| \cdot \|v_{t,t}(\Psi^*_{0,t}(X_0)) - v_{t,t}(\Psi_{0,t}(X_0))\|\right] \\
&\lesssim \sigma_{\min}^{-4}\mathbb{E}\|\Psi_{0,t}(X_0) - \Psi^*_{0,t}(X_0)\|^2 \\
&\quad + \sigma_{\min}^{-2}(1+L_f)\mathbb{E}\left[\|\Psi^*_{0,t}(X_0) - \Psi_{0,t}(X_0)\|^2 \cdot \mathbb{1}[\|X_0\| \gtrsim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\log(nD\sigma_{\min}^{-2})]\right].
\end{aligned}$$

By the Cauchy-Schwarz inequality, the bound from (79), and the fact that $X_0 \sim \mathcal{N}(0, I_D)$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi^*_{0,t}(X_0) - \Psi_{0,t}(X_0)\| \cdot \|v_{t,t}(\Psi^*_{0,t}(X_0)) - v_{t,t}(\Psi_{0,t}(X_0))\|\right] \\
&\lesssim \sigma_{\min}^{-4}\mathbb{E}\|\Psi_{0,t}(X_0) - \Psi^*_{0,t}(X_0)\|^2 \\
&\quad + \sigma_{\min}^{-2}(1+L_f)\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}D^{\mathcal{O}(1)}\mathbb{P}^{1/2}(\|X_0\| \gtrsim \sqrt{D}\exp\{\mathcal{O}(\sigma_{\min}^{-2})\}\log(nD\sigma_{\min}^{-2})).
\end{aligned}$$

Substituting the expression for L_f and using the Gaussian concentration bound from Lemma E.1, we conclude that

$$\mathbb{E}\left[\|\Psi^*_{0,t}(X_0) - \Psi_{0,t}(X_0)\| \cdot \|v_{t,t}(\Psi^*_{0,t}(X_0)) - v_{t,t}(\Psi_{0,t}(X_0))\|\right] \lesssim \sigma_{\min}^{-4}E_t + n^{-1}.$$

Substituting this bound into (78) and taking into account that $\mathrm{Law}(\Psi^*_{0,t}(X_0)) = \mathsf{p}^*_t$, we obtain

$$\frac{\mathrm{d}E_t}{\mathrm{d}t} \lesssim \sigma_{\min}^{-4}E_t + n^{-1} + \mathbb{E}\|b^*(X_t,t) - v_{t,t}(X_t)\|^2 + \mathbb{E}\|v_{t,t}(\Psi_{0,t}(X_0)) - \partial_t\Psi_{0,t}(X_0)\|^2.$$

By the Grönwall lemma, we have that

$$\begin{aligned}
E_{1-t_0} &\lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\left(\int_0^{1-t_0}\mathbb{E}\|v_{t,t}(X_t) - b^*(X_t,t)\|^2\,\mathrm{d}t + R(v) + n^{-1}\right) \\
&\lesssim \exp\{\mathcal{O}(\sigma_{\min}^{-4})\}\left(\mathbb{E}[\ell_{SD}(v,X_1) - \ell_{SD}(v^*,X_1)] + n^{-1}\right),
\end{aligned}$$

where the last inequality uses Proposition 2.1. Therefore, observing that $W_2^2(\mathrm{Law}(\Psi_{0,1-t_0}(X_0)), \mathsf{p}^*_{1-t_0}) \leqslant E_{1-t_0}$ and using the generalization bound established in Theorem 3.7, we conclude that

$$\begin{aligned}
&W_2^2(\mathrm{Law}(\widehat{\Psi}_{0,1-t_0}(X_0)), \mathsf{p}^*_{1-t_0}) \\
&\lesssim (D\log n)^{\mathcal{O}(1)}e^{\mathcal{O}(\sigma_{\min}^{-4})}\lambda(t_0)\left(\frac{\lambda(t_0)}{n}\right)^{\frac{2\beta}{5\beta+2d}}\log(\lambda(t_0))\log(4/\delta),
\end{aligned} \tag{80}$$

on an event of probability at least $(1-\delta)$. As for the second term in (77), by Assumption 3.1, we have that

$$W_2(\mathsf{p}^*_{1-t_0}, \mathsf{p}^*) \leqslant \mathbb{E}^{1/2}\|\alpha_{1-t_0}X_0 + \beta_{1-t_0}X_1 - X_1\|^2 \lesssim \sqrt{D}(\alpha_{1-t_0} \vee (1-\beta_{1-t_0})).$$

Combining this bound with (77) and (80) yields the desired result.

□

E Tools from probability theory

Lemma E.1 (Gaussian concentration of the norm Vershynin [2018], Theorem 3.1.1, adapted)**.** *Let $Z \sim \mathcal{N}(0, I_D)$ be a standard Gaussian random vector in $\mathbb{R}^D$. There exist absolute constants $C, c > 0$ such that,*

$$\mathbb{P}\left(\left|\|Z\| - \sqrt{D}\right| \geqslant t\right) \leqslant C\exp(-ct^2), \quad \textit{for all } t \geqslant 0.$$

Contents

1 Introduction **1**

1.1 Other Related Works 2

2 Preliminaries and notations **2**

3 Results **5**

3.1 Assumptions 5

3.2 Flow map approximation 6

3.3 Flow map estimation 8

3.4 Distribution estimation 9

4 Conclusion **10**

A Proofs of supporting propositions **14**

A.1 Proof of Proposition 2.1 14

A.2 Proof of Proposition 3.3 14

B Proof of Theorem 3.6 **15**

B.1 Proof of Lemma B.1 22

B.2 Proof of Lemma B.2 24

B.3 Proof of Lemma B.5 25

B.4 Proof of Lemma B.4 28

B.5 Proof of Lemma B.3 28

C Proof of Theorem 3.7 **30**

C.1 Proof of Lemma C.1 34

C.2 Proof of Lemma C.2 35

C.3 Proof of Lemma C.3 36

D Proof of Theorem 3.8 **41**

E Tools from probability theory **42**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The theoretical results presented in Section 3 were announced in the abstract and detailed in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As discussed in Section 4, the convergence rates can be improved through the use of higher-order B-splines and higher-order Runge–Kutta methods.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our theoretical results are presented in Section 3, with all proofs deferred to the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: This work presents a purely theoretical analysis without empirical validation or experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This paper presents a purely theoretical analysis.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper contains no experimental setting or implementation details, as it focuses exclusively on theoretical analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper contains no experimental results, hence no statistical significance measures or error bars are applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper contains no computational experiments, hence no resource requirements are applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper fully complies with the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents a theoretical analysis of the statistical efficiency of generative models, which is a purely mathematical investigation without direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper contains no external assets (code, data, or models) as it focuses exclusively on theoretical analysis. Consequently, there are no licensing requirements or assets to credit.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No assets released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and does not include research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and does not include research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large Language Models (LLMs) were used solely for grammar checking.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.