GENERALIZATION BOUNDS FOR FLOW-MATCHING GENERATIVE MODELS FOR INTRINSICALLY LOW-DIMENSIONAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite its remarkable empirical performance, the theoretical properties of flow-matching in terms of generalization accuracy remain largely unexplored. The current state-of-the-art generalization analyses not only impose stringent assumptions on the estimated velocity field, but also scale poorly with the ambient dimension—precisely the high-dimensional setting where flow matching is most commonly applied. In particular, the data distributions targeted by flow matching models, such as natural images, molecular geometries, etc., are commonly hypothesized to concentrate near a low-dimensional structures embedded in the high-dimensional ambient data space. However, existing analyses fail to reflect this geometry in their derived convergence rates. In this paper, we bridge the gap between the theory and practice of flow matching by establishing generalization bounds on the Wasserstein-p distance between the learned distribution and the target measure, for any $p \geq 1$. Specifically, given n independent and identically distributed samples from a target distribution P_{data} and for every $d > d_p^*(P_{\mathrm{data}})$, we show that for appropriate network and hyperparameter choices, $\mathbb{W}_p\left(\widehat{P}^{\mathrm{FM}}, P_{\mathrm{data}}\right) \lesssim n^{-1/d} + n^{-\frac{1}{2p}}(\log(1/\xi))^{\frac{1}{2p}}$ with probability at least $1-\xi$, where $d_p^*(P_{\mathrm{data}})$ denotes the Wasserstein-p dimension of the target measure. Our results demonstrate that flow matching naturally adapts to the intrinsic geometry of the data, effectively mitigating the curse of dimensionality: the convergence exponent depends only on the intrinsic dimension, $d_p^*(P_{\mathrm{data}})$, rather than the ambient dimension of the feature space, yielding convergence rates that remain meaningful in high-dimensional regimes and offering a principled theoretical explanation for the strong empirical performance of flow matching on structured data distributions with significantly relaxed assumptions, compared to the existing literature.

1 INTRODUCTION

Learning high-dimensional probability distributions from a handful of samples is a foundational problem in machine learning, with applications spanning image synthesis (Rombach et al., 2022; Saharia et al., 2022), scientific modeling (Watson et al., 2023; Geffner et al., 2025; Corso et al., 2023), and generative artificial intelligence. The dominant paradigm for this task has shifted decisively from learning a single generator/decoder (Kingma & Welling, 2014; Goodfellow et al., 2020) toward continuous-time generative models such as diffusion models and score-based methods that learn to reverse a fixed noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019; Song et al., 2021), while optimal transport-based approaches construct more direct paths between distributions (Korotin et al., 2023; Bunne et al., 2022). Flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023; Albergo et al., 2025; Chen et al., 2018) has emerged as a particularly compelling framework within this landscape, offering simulation-free training, straight probability paths, and strong empirical performance. The deterministic formulation of flow matching endows the resulting models with several attractive computational advantages, including stable optimization, flexible control over discretization during sampling, and seamless integration with modern Ordinary Differential Equation (ODE) solvers. In practice, flow-matching methods have

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

demonstrated state-of-the-art performance across a range of challenging high-dimensional applications, including text-to-image generation, video synthesis, and molecular design (Ma et al., 2024; Bose et al., 2024; Esser et al., 2024).

However, a rigorous understanding of when and why flow-matching generative models succeed remains incomplete, as prior theory has focused on sampling guarantees. Early work on score-based models established such guarantees under stochastic sampling (Block et al., 2020; Bortoli, 2022; Lee et al., 2022), with later results showing that an ϵ-approximate oracle score yields KL error $\mathcal{O}(\mathrm{poly}(D, \log(1/\delta))\epsilon^{-2})$ under mild regularity (Chen et al., 2023a; Lee et al., 2023; Albergo & Vanden-Eijnden, 2023; Albergo et al., 2025; Lu et al., 2022; Chen et al., 2023b), where D is the ambient dimension and δ the early stopping time. Stochastic interpolants bridge diffusion and flow matching (Albergo et al., 2025; Albergo & Vanden-Eijnden, 2023): Albergo & Vanden-Eijnden (2023) bound the 2-Wasserstein distance between observed and estimated flow ODE endpoints via the $\mathbb{L}_2$-error of the estimated velocity field, assuming Lipschitzness in time, while Albergo et al. (2025) analyze the KL error of flow matching under oracle approximations. For the deterministic sampler, Chen et al. (2023d) derive polynomial discretization bounds that do not extend to learned scores, and Chen et al. (2023c) obtain polynomial guarantees with learned scores via interleaved predictor-corrector updates. Most recently, Chen et al. (2026) established the first $\mathrm{polylog}(1/\epsilon)$-sampling guarantee, with $\widetilde{\mathcal{O}}(\overline{\dim}_M(P_{\text{data}})\,\mathrm{polylog}(1/\epsilon))$ steps, where $\overline{\dim}_M(P_{\text{data}})$ is the Minkowski dimension (Section 2.3) of the target P_{data}. All of these works, however, assume oracle access to an ϵ-accurate velocity or score estimate, without establishing its learnability.

Despite this progress, existing guarantees for flow matching suffer from the curse of dimensionality: their rates depend exponentially on the ambient dimension D, and thus, become uninformative precisely in the high-dimensional regimes where flow matching is most used. For example, Gao et al. (2024) derived $\mathbb{W}_2$ convergence guarantees for flow matching under the assumption that the target distribution is log-concave and admits a bounded density, obtaining $\mathbb{E}\mathbb{W}_2(\hat{P}^{\text{FM}}, P_{\text{data}}) = \widetilde{\mathcal{O}}\left(n^{-1/(D+5)}\right)$, when the estimated vector velocity field is uniformly Lipschitz. Zhou & Liu (2025) removed this Lipschitzness assumption on the learned velocity field at the cost of a slower rate, establishing $\mathbb{E}\mathbb{W}_2(\hat{P}^{\text{FM}}, P_{\text{data}}) = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{3(D+5)}}\right)$. Under additional Besov-smoothness assumptions on the target density, Fukumizu et al. (2025) showed that for any $\delta > 0$, $\mathbb{E}\mathbb{W}_2(\hat{P}^{\text{FM}}, P_{\text{data}}) = \mathcal{O}\left(n^{-\frac{s+(2\kappa)^{-1}-\delta}{2s+D}}\right)$, where s is the smoothness parameter of the density and $\kappa \geq 1/2$ is a tunable model parameter. This sits in sharp contrast with the widely held belief that high-dimensional data concentrate near low-dimensional structures (Pope et al., 2020), suggesting the D-dependence is an artifact of analysis rather than intrinsic to the problem—as in classical nonparametric statistics, where kernel estimators and Gaussian process regression achieve minimax rates depending only on intrinsic dimensionality (Bickel & Li, 2007; Kim et al., 2019). Dimension-adaptive, minimax-optimal guarantees have since been established for GANs (Dahal et al., 2022; Huang et al., 2022; Chakraborty & Bartlett, 2025), Wasserstein Autoencoders (Chakraborty & Bartlett, 2024), and diffusion models (Chen et al., 2023d; Oko et al., 2023; Tang & Yang, 2024; Chakraborty et al., 2026), but whether analogous results hold for flow matching remains open.

While recent studies have begun exploring flow matching on manifolds (Benton et al., 2024; Pi et al., 2026) and diffusion models for multi-modal low-dimensional data (Wu & Cai, 2026), existing guarantees either rely on restrictive manifold reach and density lower bounds or analyze in-expectation $\mathbb{W}_1$/$\mathbb{W}_2$ metrics. Kumar et al. (2026) partially address this, establishing $\mathbb{W}_2$ rates under oracle access to the ODE sampler for data on a smooth, lower-bounded-reach manifold with Hölder-smooth density; however, their analysis relies on strong, unverifiable geometric conditions and excludes even simple distributions (Gaussian, uniform, degenerate). In contrast, our framework provides the sharpest known non-asymptotic, high-probability concentration guarantees for arbitrary $p \geq 1$ without requiring manifold reach or global Lipschitz regularity on the learned vector field.

Summary of Contributions The main contributions of this paper are as follows.

- **High-probability Wasserstein guarantees for every $p \geq 1$.** We develop a general theoretical framework for analyzing the statistical performance of flow-matching generative models. For arbitrary $p \geq 1$, we derive non-asymptotic, high-probability bounds on the Wasserstein-p error $\mathbb{W}_p\left(\widehat{P}^{\text{FM}}, P_{\text{data}}\right)$. Specifically, when one has access to i.i.d. samples $x_1, \ldots, x_n \sim P_{\text{data}}$,

we show that, for any $d > d_p^*(P_{\text{data}})$, with probability at least $1-\xi$, $\mathbb{W}_p\Big(\widehat{P}^{\text{FM}}, P_{\text{data}}\Big) \lesssim n^{-1/d_p^*(P_{\text{data}})} + n^{-1/(2p)}\big(\log(1/\xi)\big)^{1/(2p)}$, under appropriate choices of the velocity-field network class, discretization scheme, and early-stopping rule. In contrast to existing analyses (Benton et al., 2024; Gao et al., 2024; Zhou & Liu, 2025; Kumar et al., 2026), which are largely expectation-based and often restricted to $p \in \{1,2\}$, our results hold for every $p \geq 1$ and provide guarantees at arbitrarily high confidence. We also give sufficient conditions for selecting the relevant hyperparameters as functions of the sample size and the population distribution.

- **Minimal distributional and model assumptions.** Our guarantees require no density, smoothness, or Lipschitz assumptions on either the data-generating distribution or the velocity-field class. Despite this generality, the resulting rates adapt to the intrinsic dimension of the target distribution, providing a rigorous statistical explanation for the effectiveness of flow matching on structured high-dimensional data, including images, videos, and molecular geometries that may concentrate near low-dimensional subsets of the ambient space.
- **Minimax rates.** For targets supported on $\tilde{d}$-dimensional regular sets, such as compact $\tilde{d}$-dimensional manifolds, compact convex sets of dimension $\tilde{d}$, and self-similar sets of similarity dimension $\tilde{d}$, our bound gives $\mathcal{O}(n^{-1/d})$ for every $d > \tilde{d}$ and every $1 \leq p < \tilde{d}/2$, matching the minimax rate for distribution estimation on such sets up to an arbitrarily small slack in the exponent.

2 Background and Problem Setting

2.1 Notations and Definitions

$B_\varrho(x,r)$ denotes the open ball of radius r around x w.r.t. the metric ϱ. For a probability measure γ, $\text{supp}(\gamma) = \{x : \gamma(B_\varrho(x,r)) > 0, \text{ for all } r > 0\}$. $\mathbb{1}(\cdot)$ denotes the indicator function. We write $A_n \lesssim B_n$ (equivalently $A_n = \mathcal{O}(B_n)$) if there exists $C > 0$, independent of n, such that $A_n \leq CB_n$; $A_n \precsim B_n$ (equivalently $A_n = \widetilde{\mathcal{O}}(B_n)$) if $A_n \leq B_n \cdot \log^C(en)$ for some $C > 0$, i.e. up to poly-log factors in n; and $A_n \asymp B_n$ if both $A_n \lesssim B_n$ and $B_n \lesssim A_n$ hold. For $k \in \mathbb{N}$, $[k] = \{1, \ldots, k\}$. For a function of t, ξ_t, we write $\dot{\xi}_t = \mathrm{d}\xi_t/\mathrm{d}t$. For a distribution μ, $\mathcal{T}_M(\mu) = \text{Law}(Z\mathbb{1}\{\|Z\| \leq M\})$ where $Z \sim \mu$. $\text{div}_x(\cdot)$ denotes the divergence operator w.r.t. x. We also recall the following standard definitions.

Definition 1 (Covering Number)**.** For a metric space $(\mathcal{S}, \varrho)$, the ϵ-covering number with respect to (w.r.t.) ϱ is $\mathcal{N}(\epsilon; \mathcal{S}, \varrho) = \inf\{n \in \mathbb{N} : \exists\, x_1, \ldots, x_n \text{ such that } \cup_{i=1}^n B_\varrho(x_i, \epsilon) \supseteq \mathcal{S}\}$.

Definition 2 (Neural networks)**.** Let $L \in \mathbb{N}$ and $\{N_i\}_{i\in[L]} \subset \mathbb{N}$. An L-layer neural network $f : \mathbb{R}^D \to \mathbb{R}^{N_L}$ is

$$f(x) = A_L \circ \sigma_{L-1} \circ A_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(x), \tag{1}$$

where $A_i(y) = W_i y + b_i$, with $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$, $b_i \in \mathbb{R}^{N_i}$, $N_0 = D$, and each σ_j applied component-wise. Here $\{W_i\}$ are weights, $\{b_i\}$ biases, and $\{\sigma_i\}_{1\leq i\leq L-1}$ activation functions. Define: depth $\mathcal{L}(f) := L$; number of weights $\mathcal{W}(f) = \sum_{i=1}^L N_i N_{i-1}$; maximum weight $\mathcal{B}(f) = \max_{1\leq j\leq L}(\|b_j\|_\infty \vee \|W_j\|_\infty)$; and maximum output $\mathcal{K}(f) = \sup_{x\in\mathbb{R}^D} \|f(x)\|_\infty$. Let

$$\mathcal{NN}_{\{\sigma_i\}_{1\leq i\leq L-1}}(L, W, B, K) = \big\{f \text{ of the form (1)} : \mathcal{L}(f) \leq L, \mathcal{W}(f) \leq W, \mathcal{B}(f) \leq B, \mathcal{K}(f) \leq K\big\}.$$

When $\sigma_j(x) = x \vee 0$ for all $j \in [L-1]$, we write $\mathcal{RNN}(L, W, B, K)$ for $\mathcal{NN}_{\{\sigma_i\}_{1\leq i\leq L-1}}(L, W, B, K)$.

Definition 3 (Wasserstein p-distance)**.** For probability measures μ, ν on $\mathbb{R}^D$, the Wasserstein-p distance (under ℓ_2) is $\mathbb{W}_p(\mu,\nu) = \big(\inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\|X-Y\|_2^p\big)^{1/p}$, where $\Gamma(\mu,\nu)$ is the set of couplings of μ and ν.

2.2 Flow-matching

The fundamental problem in generative modeling is to learn an unknown distribution P_{data} on a data space $\mathcal{X} \subseteq \mathbb{R}^D$ from n i.i.d. observations $x_1, \ldots, x_n \sim P_{\text{data}}$, where D can be extremely large—e.g., $D = 49{,}152$ for 128×128 color ImageNet images. Flow matching (Lipman et al., 2023) constructs a continuous probability path $\{P_t\}_{t\in[0,1]}$ from an easy-to-sample reference P_0 (typically standard Gaussian) to the target $P_1 \equiv P_{\text{data}}$, governed by a time-dependent vector field $v(\cdot,\cdot) : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$ whose induced flow transports P_0 to P_{data}: starting from $X_0 \sim P_0$,

$$\dot{X}_t = v(X_t, t), \quad X_0 \sim P_0. \tag{2}$$

Since many velocity fields transport P_0 to P_1, a canonical choice arises from the linear interpolant $I(x_0, x_1, t) = \alpha_t x_1 + \beta_t x_0$, with $\alpha_t, \beta_t \in [0, 1]$ satisfying $\alpha_0 = \beta_1 = 0$, $\alpha_1 = \beta_0 = 1$; this gives $I(x_0, x_1, 0) = x_0$ and $I(x_0, x_1, 1) = x_1$. The process $\{X_t\}_{0 \le t \le 1}$ constructed as $X_t = I(X_0, X_1, t)$ for independent $X_0 \sim P_0$, $X_1 \sim P_1 \equiv P_{\text{data}}$, with marginal law $P_t = \text{Law}(X_t)$. Since P_{data} is unknown, it is replaced in practice by $P_n^{\text{emp}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, so $X_1 \sim P_n^{\text{emp}}$. Among velocity fields consistent with $\{P_t\}$, flow matching targets the conditional expectation $v_\star(x, t) = \mathbb{E}\big(\dot{X}_t \mid X_t = x\big)$, satisfying the continuity equation $\partial_t \varrho_t + \text{div}_x(v_\star \varrho_t) = 0$, $t \in [0, 1)$, where ϱ_t is the density of P_t. Since $v_\star(\cdot, t)$ may explode as $t \to 1$, the flow is instead run until $t = 1 - \delta$. Theorem 4 shows the resulting ODE

$$\dot{X}_t = v_\star(X_t, t), \quad X_0 \sim P_0, \quad t \in [0, 1 - \delta] \tag{3}$$

is well-defined: for any initialization the solution exists and is unique, and X_t has density ϱ_t.

Theorem 4. *For all $t \in [0, 1-\delta]$, P_t admits a density, ϱ_t, w.r.t. the Lebesgue measure and satisfies the continuity equation (in the weak sense), $\partial_t \varrho_t + \text{div}_x(\varrho_t v_\star(\cdot, t)) = 0$, where $v_\star(x, t) = \beta_t^{-1}\dot{\beta}_t x + \beta_t^{-1}(\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t)\left(\sum_{i=1}^n \exp\left(-\frac{\|x-\alpha_t x_i\|^2}{2\beta_t^2}\right)\right)^{-1}\sum_{i=1}^n x_i \exp\left(-\frac{\|x-\alpha_t x_i\|^2}{2\beta_t^2}\right)$. Let*

$$y_x'(t) = v_\star(y_x(t), t), \text{ with } y_x(0) = x, \tag{4}$$

be the Initial Value Problem (IVP) with initial point x. Then (4) has a unique solution for every $x \in \mathbb{R}^D$, and $P_t = (Y_t)_\sharp P_0$, where $Y_t : x \mapsto y_x(t)$.

Flow matching trains a neural network $v : \mathbb{R}^D \times [0, 1] \to \mathbb{R}^D$ to approximate $v_\star$ by minimizing $\mathcal{L}(v) = \int_0^1 \mathbb{E}\|v(X_t, t) - v_\star(X_t, t)\|^2 \text{d}t$. A key observation of Lipman et al. (2023) is that minimizing $\mathcal{L}(v)$ is equivalent (Lemma 11) to minimizing the conditional flow matching objective

$$\mathcal{L}_{\text{CFM}}(v) = \int_0^1 \mathbb{E}\|v(X_t, t) - \dot{X}_t\|^2 \text{d}t = \int_0^1 \mathbb{E}\left\|v(X_t, t) - (\dot{\alpha}_t X_1 + \dot{\beta}_t X_0)\right\|^2 \text{d}t, \tag{5}$$

which can be estimated from data without knowledge of $v_\star$. Substituting P_n^{emp} for P_{data} gives the empirical objective, $\mathcal{L}_{\text{CNF}}(v) = \frac{1}{n}\sum_{i=1}^n \int_0^1 \mathbb{E}\left[\|v(X_{i,t}, t) - (\dot{\alpha}_t x_i + \dot{\beta}_t X_0)\|^2\right] \text{d}t$, where $X_{i,t} = \alpha_t x_i + \beta_t X_0$, $X_0 \sim P_0$. Since this integral is not computed exactly in practice, we study its numerical approximation

$$\hat{\mathcal{L}}_n(v) = \frac{1}{n}\sum_{i=1}^n \sum_{k=0}^{N-1} (t_{k+1} - t_k)\mathbb{E}\|v(X_{i,t_k}, t_k) - (\dot{\alpha}_{t_k} x_i + \dot{\beta}_{t_k} X_0)\|^2, \tag{6}$$

where $0 = t_0 < t_1 < \cdots < t_N = 1 - \delta$ partitions $[0, 1 - \delta]$; we use a uniform step size, $t_{k+1} = t_k + h$ for $k \in [N - 1]$. The expectation in (6) is typically further approximated via Monte Carlo samples $X_0^{(j)} \sim \mathcal{N}(0, I_D)$, $X_1^{(j)} \sim \text{Unif}\left(\{X_1, \ldots X_n\}\right)$, $X_t^{(j)} = \alpha_t X_1^{(j)} + \beta_t X_0^{(j)}$, giving the Monte Carlo estimate

$$\hat{\mathcal{L}}_{n,m}^{\text{mc}}(v) = \frac{1}{m}\sum_{j=1}^m \sum_{k=0}^{N-1} (t_{k+1} - t_k)\left\|v(X_{t_k}^{(j)}, t_k) - (\dot{\alpha}_{t_k} X_1^{(j)} + \dot{\beta}_{t_k} X_0^{(j)})\right\|^2. \tag{7}$$

The resulting velocity field estimates are $\hat{v}_n = \text{argmin}_{v \in \mathcal{V}} \hat{\mathcal{L}}_n(v)$ and $\hat{v}_{n,m}^{\text{mc}} = \text{argmin}_{v \in \mathcal{V}} \hat{\mathcal{L}}_{n,m}^{\text{mc}}(v)$.

Stochastic Sampling After the velocity field is learned, one can generate new samples by following the ODE, $\text{d}X_t = \hat{v}(X_t, t)\text{d}t$, $X_0 \sim \mathcal{N}(0, I_D)$. However, analyzing the above ODE is rather difficult, as stability of the flow relies critically on the regularity of the learned velocity field. In particular, standard perturbation arguments require $\hat{v}(\cdot, t)$ to be uniformly globally Lipschitz, and the error between trajectories generated by the learned and optimal velocity fields typically grows according to Grönwall's inequality, leading to an exponential dependence on the Lipschitz constant. Since this Lipschitz constant may potentially blow up as $t \to 1$, these bounds become increasingly loose near the terminal time, making it difficult to obtain meaningful control of the Wasserstein distance between the generated and target distributions. The existing literature artificially imposes a boundedness on the Lipschitz constant of the learned velocity field (Kumar et al., 2026; Gao et al., 2024). Alternatively, one can perform the generation through an SDE. This is because the Fokker-Plank equation corresponding to (3),

$$\partial_t \varrho_t = -\text{div}_x(\varrho_t v_\star) = -\text{div}_x\left(\varrho_t\left(v_\star + \frac{1}{2}\sigma^2 \nabla \log \varrho_t\right)\right) + \frac{1}{2}\sigma^2 \Delta \varrho_t,$$

which is satisfied by the SDE, $\mathrm{d}X_t = \big(v_\star(X_t,t) + \frac{\sigma}{2}\nabla \log \varrho_t(X_t)\big)\,\mathrm{d}t + \sigma \mathrm{d}W_t$. A formal statement that ϱ_t satisfies the above SDE appears in Appendix A.3. Here, $\sigma > 0$. Since, $v_\star(x,t) = \frac{\dot\alpha_t}{\alpha_t}x + \beta_t^2\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)\nabla \log \varrho_t(x)$, One can generate samples by following the SDE,

$$\mathrm{d}X_t = (a_t v_\star(X_t,t) - b_t X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad 0 \le t \le 1-\delta, \tag{8}$$

where, let $a_t = 1 + \Big(2\beta_t(\dot\alpha_t\beta_t - \dot\beta_t\alpha_t)\Big)^{-1}\sigma^2\alpha_t$ and $b_t = \Big(2\beta_t(\dot\alpha_t\beta_t - \dot\beta_t\alpha_t)\Big)^{-1}\sigma^2\dot\alpha_t$. After the velocity field is learned, one generates new samples through the Euler–Maruyama discretization of the SDE as,

$$\mathrm{d}\widehat{X}_t = \Big(a_{t_k}\hat v(\widehat{X}_{t_k},t_k) - b_t\widehat{X}_t\Big)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad t_k \le t < t_{k+1}, \tag{9}$$

where $\hat v$ is either $\hat v_n$ or $\hat v^{\mathrm{mc}}_{n,m}$. The distribution of $\widehat{X}_{1-\delta}$ is denoted as $\hat P_{1-\delta}(\hat v) := \mathrm{Law}\left(\widehat{X}_{1-\delta}\right)$. Letting $B(s,t) = \int_s^t b_r \mathrm{d}r$, the solution to (9) satisfies

$$\widehat{X}_{t_{k+1}} = e^{-B(t_k,t_{k+1})}\widehat{X}_{t_k} + a_{t_k}\hat v(\widehat{X}_{t_k},t_k)\int_{t_k}^{t_{k+1}} e^{-B(s,t_{k+1})}\mathrm{d}s + \sigma\,\xi_k\sqrt{\int_{t_k}^{t_{k+1}} e^{-2B(s,t_{k+1})}\mathrm{d}s},$$

where $\xi_k \sim \mathcal{N}(0, I_D)$. Such a data generation procedure has been used by practitioners such as Singh & Fischer (2024); Ma et al. (2024), who report empirical superiority compared to deterministic sampling in flow-matching.

2.3 Intrinsic Data Dimension

Several notions of the *effective dimension* of a probability measure formalize the intuition of intrinsic dimension, most characterizing the growth rate of covering numbers on the logarithmic scale. Let $(\mathcal{S}, \varrho)$ be a Polish space endowed with the ℓ_∞ metric and let μ be a probability measure on $\mathcal{S}$. Recall the (ϵ,τ)-covering number of μ (Posner et al., 1967), $\mathcal{N}_\epsilon(\mu,\tau) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : S \subseteq \mathcal{S},\, \mu(S) \ge 1-\tau\}$, the minimal number of ϵ-balls needed to cover a set carrying at least $1-\tau$ of the mass of μ.

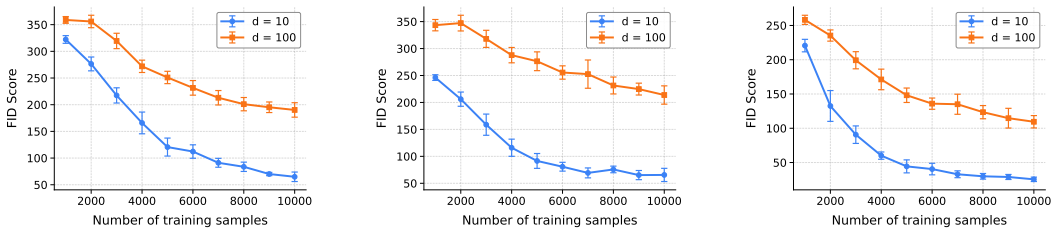
A natural measure of dimensionality is the *upper Minkowski dimension* $\overline{\dim}_M(\mu) = \limsup_{\epsilon\downarrow 0}\left(\log(1/\epsilon)\right)^{-1}\log \mathcal{N}(\epsilon;\, \mathrm{supp}(\mu),\, \ell_\infty)$, which depends only on the metric entropy of the support and so requires no smoothness or manifold structure, accommodating irregular geometries such as fractals. Its statistical implications are well established: covering numbers of smooth function classes are governed by the Minkowski dimension of their domain (Kolmogorov & Tikhomirov, 1961), and deep networks, WGANs and Wasserstein Autoencoders all attain rates reflecting it rather than the ambient dimension (Nakada & Imaizumi, 2020; Huang et al., 2022; Chakraborty & Bartlett, 2024). Being a property of the support alone, however, it ignores how mass is distributed and can be as large as D even when μ concentrates near a low-dimensional region (Chakraborty & Bartlett, 2025). Weed & Bach (2019) therefore refined Dudley's (Dudley, 1969) entropic dimension into a notion sensitive to the mass distribution of μ, and showed that it governs the convergence rate of the empirical Wasserstein distance.

Definition 5 (Wasserstein Dimension (Weed & Bach, 2019))**.** For $p > 0$, the p-upper Wasserstein dimension of μ is $d_p^*(\mu) = \inf\left\{s > 2p : \limsup_{\epsilon\downarrow 0}\left(\log(1/\epsilon)\right)^{-1}\log \mathcal{N}_\epsilon\Big(\mu,\, \epsilon^{\frac{sp}{s-2p}}\Big) \le s\right\}$, and the lower Wasserstein dimension is $d_*(\mu) = \lim_{\tau\downarrow 0}\,\liminf_{\epsilon\downarrow 0}\frac{\log \mathcal{N}_\epsilon(\mu,\tau)}{\log(1/\epsilon)}$.

Weed & Bach (2019) established that these jointly control the convergence rate of the empirical Wasserstein distance, $n^{-1/d_*(\mu)} \lesssim \mathbb{E}\,\mathbb{W}_p(\hat\mu_n,\mu) \lesssim n^{-1/d_p^*(\mu)}$. The upper Wasserstein dimension has recently been used to bound error rates for GANs (Chakraborty & Bartlett, 2025) and score-matching diffusion models (Chakraborty et al., 2026).

3 A Proof of Concept Experiment

Before presenting our theoretical analysis, we give an experiment illustrating that Flow Matching (FM) error rates depend primarily on the *intrinsic* dimension of the data, not the ambient dimension. We establish finite-sample generalization bounds using the Fokker-Planck equivalent SDE, bypassing artificial Lipschitz assumptions on the velocity field. Empirically, probability-flow ODE simulations confirm that our predicted intrinsic-dimensional error decay holds under standard deterministic sampling. Since intrinsic dimensionality is difficult to control in real image data, we



(a) Results on the basenji class (b) Results on the bubble class (c) Results on the goldfish class

Figure 1: Average generalization error (in terms of FID scores) for different values of n for flow-matching. The error bars denote the standard deviation out of 10 replications.

follow the synthetic construction of Pope et al. (2020) and Chakraborty & Bartlett (2024): we use a pre-trained BigGAN (Brock et al., 2019) (128-dim latent space, $128 \times 128 \times 3$ output, trained on ImageNet (Deng et al., 2009)) to generate 11,000 images each for the `basenji`, `bubble`, and `goldfish` classes, fixing all but $d \in \{10, 100\}$ latent coordinates to zero so that generated data lie on a d-dimensional manifold. Images are downsampled to 28×28 via bilinear interpolation. For each d, we train a standard UNet FM model with the usual objective $\mathbb{E}\|v_t - v_\theta(x_t, t \cdot 1000)\|_2^2$, varying the training set size over $\{1000, 2000, \ldots, 10000\}$ (Adam, lr 10^{-4}, batch size 64, 50 epochs; NVIDIA RTX 6000 Blackwell GPUs). At evaluation, we simulate the learned probability-flow ODE (100 steps) to draw 1,000 samples and compute FID (Heusel et al., 2017) against a held-out set of equal size (indices 10,000–11,000), averaged over 10 repetitions. The empirical results, presented in Fig. 1, clearly validate our hypothesis. As the number of training samples increases from 1000 to 10000, the FM model trained on data with intrinsic dimension $d = 10$ consistently achieves substantially lower FID scores than the model trained on data with intrinsic dimension $d = 100$. The gap persists across large sample sizes, and in fact widens slightly as more data become available, indicating that lower intrinsic-dimensional structure yields faster error decay under flow-based generative learning. This behavior aligns precisely with our theoretical predictions: the intrinsic dimension, rather than the ambient pixel dimension, governs the sample complexity and achievable error rates of flow-matching generative models.

4 Theoretical Analyses

4.1 Main Result and Inference

To begin our theoretical analyses, we make the following standard assumption (Huang et al., 2022; Oko et al., 2023; Chakraborty & Bartlett, 2025) that the data is independent and identically distributed and supported on the unit hypercube, $[0, 1]^D$.

Assumption 1. *We assume that $x_1, \ldots, x_n$ are independent and identically distributed according to the distribution P_{data}, such that $P_{data}\left([0, 1]^D\right) = 1$.*

Our approach relaxes several restrictive assumptions found in recent literature. Unlike Kumar et al. (2026), we do not require the existence of densities, smoothness, or a support characterized as a smooth manifold with bounded reach. Additionally, we avoid the uniform score-Lipschitzness and log-concavity of the target measure imposed by Gao et al. (2024) and Zhou & Liu (2025). Further, unlike Fukumizu et al. (2025), we do not require that P_{data} to have a lower-bounded smooth density.

Design Choices To ensure that the interpolating path is regular enough, we impose the following design choices on α_t, and β_t. Design choices D 1(a)–(c) are satisfied by the commonly used interpolation schedules, $(\alpha_t = t, \beta_t = 1 - t)$, $(\alpha_t = t, \beta_t = \sqrt{1 - t^2})$.

D 1. *We assume that for any $t, s \in [0, 1 - \delta]$, the following hold:*

(a) $\left|\frac{\dot\beta_t}{\beta_t}\right|$, $\left|\frac{\dot\alpha_t\beta_t - \dot\beta_t\alpha_t}{\beta_t}\right|$, a_t^2, $|b_t|$, β_t^{-1}, $\left|\frac{\ddot\beta_t}{\beta_t} - \left(\frac{\dot\beta_t}{\beta_t}\right)^2\right|$ $\left|\beta_t^{-4}\left(3\alpha_t\beta_t\dot\alpha_t\dot\beta_t - \beta_t^2\dot\alpha_t^2 - 2\alpha_t^2\dot\beta_t^2\right)\right|$, $\left|\ddot\alpha_t - \alpha_t\ddot\beta_t\,\beta_t^{-1} - \dot\alpha_t\dot\beta_t\,\beta_t^{-1} + \alpha_t\dot\beta_t^2\,\beta_t^{-2}\right|$, $\left|\alpha_t\left(\dot\alpha_t\beta_t - \dot\beta_t\alpha_t\right)^2\beta_t^{-4}\right| \lesssim \delta^{-\kappa_1}$.

(b) $|\alpha_t - \alpha_s|$, $|\beta_t - \beta_s|$, $|a_t - a_s| \lesssim \delta^{-\kappa_2}|t - s|$.

(c) $1-\alpha_{1-\delta},\ \beta_{1-\delta} \lesssim \delta^{\kappa_3}$.

Note that for the linear interpolation, ($\alpha_t = t$, $\beta_t = 1-t$), it suffices the above hold with $\kappa_1 = 4$, $\kappa_2 = 2$ and $\kappa_3 = 1$, where as for the Variance Preserving (VP) choice ($\alpha_t = t$, $\beta_t = \sqrt{1-t^2}$), the above hold with $\kappa_1 = 3$, $\kappa_2 = 0.5$ and $\kappa_3 = 0.5$.

Under Assumption 1, we establish high-probability upper bounds on the Wasserstein-p distance between the flow-matching estimators $\hat{P}_{1-\delta}(\hat{v}_n)$ and $\hat{P}_{1-\delta}(\hat{v}^{\mathrm{mc}}_{n,m})$ and the target distribution P_{data}. These bounds depend on the training sample size through a rate exponent governed solely by the upper p-Wasserstein dimension of P_{data}, rather than the ambient dimension D. Consequently, with data-driven choices of the discretization step-size h, early stopping time δ, and velocity-field function class $\mathcal{V}$, minimizing the flow-matching MSE loss yields an estimator that automatically adapts to the low-dimensional structure of P_{data}. Theorem 6 derives a high-probability bound in terms of the Wasserstein-p distance.

Theorem 6 (Main Theorem). *Suppose that $d > d^*_p(P_{data})$ and Assumption 1 holds. Then under Design Choices 1, for any $p \geq 1$, there exists a constant $n_0 \in \mathbb{N}$, such that if $n \geq n_0$, and $\mathcal{V} \equiv \mathcal{RN}(W, L, B, K)$ with $\delta \asymp n^{-\frac{1}{d\kappa_3}}$, $h \asymp n^{-\frac{2p(2p+1)}{d} - \frac{10\kappa_1+2\kappa_2}{d\kappa_3}}$, $M \asymp n^{\frac{2p}{d}}$, then, with probability at least $1-\xi$,*

$$\mathbb{W}_p\left(P_{data}, \mathcal{T}_M\left(\hat{P}_{1-\delta}(\hat{v}_n)\right)\right) \lesssim n^{-1/d} + n^{-\frac{1}{2p}} \log^{\frac{1}{2p}}(1/\xi). \tag{10}$$

Here, $W \asymp n^{\frac{p(2p+1)(D+2)}{d} + \frac{(20+7D)\kappa_1+4\kappa_2}{2d\kappa_3}} \log^{D/2+1} n$, $L \asymp \log n$, $B \asymp n^{\left(\frac{2p(2p+1)}{d} + \frac{20\kappa_1+4\kappa_2}{2d\kappa_3}\right) \vee \frac{(3\kappa_1)\vee 1}{d\kappa_3}} \sqrt{\log n}$, and $K \asymp n^{\frac{(3\kappa_1)\vee 1}{d\kappa_3}} \sqrt{\log n}$. Further, if $m \asymp n^{\frac{p(2p+1)(D+6)}{d} + \frac{(24+7D)\kappa_1+4\kappa_2+4((6\kappa_1)\vee 2)}{2d\kappa_3}} \log^2 n \left(\log^{D/2+3} n + \log(1/\xi)\right)$, with probability at least $1-\xi$,

$$\mathbb{W}_p\left(P_{data}, \mathcal{T}_M\left(\hat{P}_{1-\delta}(\hat{v}^{mc}_{n,m})\right)\right) \lesssim n^{-1/d} + n^{-\frac{1}{2p}} \log^{\frac{1}{2p}}(1/\xi). \tag{11}$$

Inference for Data Supported on a Manifold Recall that a set $\mathcal{A}$ is $\tilde{d}$-regular w.r.t. the $\tilde{d}$-dimensional Hausdorff measure $\mathbb{H}^{\tilde{d}}$ if $\mathbb{H}^{\tilde{d}}(B_\varrho(x,r)) \asymp r^{\tilde{d}}$ for all $x \in \mathcal{A}$ (Weed & Bach, 2019, Definition 6). By Weed & Bach (2019, Proposition 8), if $\mathrm{supp}(P_{\text{data}})$ is $\tilde{d}$-regular and $P_{\text{data}} \ll \mathbb{H}^{\tilde{d}}$, then $d^*_p(P_{\text{data}}) = \tilde{d}$ for every $p < \tilde{d}/2$; in particular this holds for every compact $\tilde{d}$-dimensional differentiable manifold, nonempty compact convex set of dimension $\tilde{d}$, its relative boundary, and self-similar sets of similarity dimension $\tilde{d}$. For such supports and any $1 \leq p < \tilde{d}/2$, Theorem 6 applies with every $d > \tilde{d}$, giving $\mathcal{O}\big(n^{-1/d} + n^{-1/2p}\log^{1/2p}(1/\xi)\big)$; as $d \downarrow \tilde{d}$ this matches the minimax rate for distribution estimation on $\tilde{d}$-dimensional manifolds up to the slack $d - \tilde{d}$ (Chakraborty, 2025), and improves on Kumar et al. (2026), whose comparable guarantee is not minimax optimal and additionally requires bounded reach and Hölder-smooth densities.

Comparison with Existing Bounds Unlike Gao et al. (2024); Zhou & Liu (2025), whose $\mathcal{O}(n^{-1/(D+5)})$-type rates depend on the log-concavity or Lipschitzness of the score/velocity field and scale with the ambient dimension D, and unlike Fukumizu et al. (2025), who require Besov smoothness of the density, our bound in Theorem 6 holds for every $p \geq 1$, is non-asymptotic and high-probability rather than in-expectation, and imposes no density, smoothness, or Lipschitzness condition on P_{data} or the velocity class. The claim concerns the exponent of n; the prescribed architecture still depends on D, and more than polynomially (see below). This places flow matching on the same footing as the dimension-adaptive guarantees recently established for GANs (Chakraborty & Bartlett, 2025) and diffusion models (Chakraborty et al., 2026), while relaxing the compact-support and manifold-structure assumptions common to both lines of work.

Choice of Early Stopping and Discretization Theorem 6 sets the early-stopping offset $\delta \asymp n^{-1/(d\kappa_3)}$, which by D 1(c) keeps the truncation error $\mathbb{W}_p(P_{1-\delta}, P_{\text{data}}) = \mathcal{O}(\delta^{\kappa_3})$ on the same order as the sampling error, while preventing $v_\star$ from blowing up as $t \to 1^-$. Because the sampler is an SDE with a nondegenerate diffusion coefficient, its increments over a step are of mean square order h rather than h^2, and the cumulative Euler–Maruyama error is $\mathcal{O}(\delta^{-(10\kappa_1+2\kappa_2)}h)$ (Lemma 8); matching this to the sampling error gives the step size $h \asymp n^{-\frac{2p(2p+1)}{d} - \frac{10\kappa_1+2\kappa_2}{d\kappa_3}}$ above, and a number of partitions $N \asymp h^{-1}$ that still grows only polynomially in n. The proof also fixes internally

the approximation accuracy ϵ and the excursion probability η, both polynomial in n; the latter sets the radius $R \asymp \sqrt{D \log(1/\eta)} \asymp \sqrt{D \log n}$ of the box on which $v_\star$ is approximated.

Network and Monte Carlo Scaling The velocity network width $W \asymp n^{\frac{p(2p+1)(D+2)}{d} + \frac{(20+7D)\kappa_1 + 4\kappa_2}{2d\kappa_3}} \log^{D/2+1} n$ is polynomial in n for each fixed D, with an exponent that grows linearly in the ambient dimension D and shrinks as the intrinsic dimension d increases, while depth $L \asymp \log n$ suffices; the caps B and K carry an extra $D\sqrt{\log n}$, the magnitude $v_\star$ attains on the truncation box. The polylogarithmic factors reflect the cost of covering the box $[-R, R]^D$ on which $v_\star$ is approximated, and do not affect the exponent of n, which is what governs the rate. The Monte Carlo sample size $m \asymp n^{\frac{p(2p+1)(D+6)}{d} + \frac{(24+7D)\kappa_1 + 4\kappa_2 + 4((6\kappa_1)\vee 2)}{2d\kappa_3}} \log^2 n \big(\log^{D/2+3} n + \log(1/\xi) \big)$ inherits this (D, d)-dependence, inflated by the covering number of $\mathcal{V}$ and by K^4 from the sub-exponential concentration in Lemma 14; larger D/d ratios thus demand proportionally more generated samples, mirroring the scaling seen for GANs (Chakraborty & Bartlett, 2025; Huang et al., 2022) and diffusion models (Chakraborty et al., 2026).

4.2 Proof Sketch

Error Decomposition To derive an upper bound on the excess risk, we first decompose the KL estimation error into a statistical error due to velocity field estimation and a numerical error resulting from the forward Euler-Maruyama discretization of the SDE. This decomposotion is shown in Lemma 7 with proof appearing in Appendix B.1. Through out the remainder of the paper, all expectations are taken conditioned on the data i.e., $\mathbb{E}(\cdot \mid x_1, \ldots, x_n)$.

Lemma 7. *Under Assumption 1 and Design Choices D 1, for any estimator $\hat{v} \in \mathcal{V}$ of $v_\star$,*

$$\mathrm{KL}\left(P_{1-\delta} \,\|\, \hat{P}_{1-\delta}(\hat{v})\right) \le \frac{2}{\sigma^2} \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} a_t^2 \, \mathbb{E}\big\|v_\star(X_t, t) - v_\star(X_{t_k}, t_k)\big\|^2 \mathrm{d}t$$

$$+ \frac{2}{\sigma^2} \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (a_t - a_{t_k})^2 \, \mathbb{E}\big\|v_\star(X_{t_k}, t_k)\big\|^2 \mathrm{d}t + \frac{1}{\sigma^2} \sum_{k=0}^{N-1} h_k \, a_{t_k}^2 \, \mathbb{E}\big\|v_\star(X_{t_k}, t_k) - \hat{v}(X_{t_k}, t_k)\big\|^2,$$

where every expectation is taken under the law of the process $\{X_t\}_{t \in [0, 1-\delta]}$ solving (8)*.*

Using Lemma 7, we aim to bound the terms individually in the subsequent paragraphs.

Discretization Error The first in Lemma 7, $\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\|a_t v_\star(X_t, t) - a_{t_k} v_\star(X_{t_k}, t_k)\|^2 \mathrm{d}t$ captures the *discretization error* incurred when the continuous-time SDE generated by the velocity field $v_\star$ is approximated using the forward Euler-Maruyama scheme. This term measures the cumulative discrepancy between the velocity evaluated at the left endpoint of each interval and the true velocity along the continuous trajectory. Consequently, it quantifies the bias introduced by replacing the exact SDE dynamics with piecewise-constant updates over the partition $\{t_j\}_{j=0}^N$. The following lemma shows that this error decreases as the uniform discretization step size h tends to zero.

Lemma 8. *Under Assumption 1 and Design Choices D 1, with uniform step size $h \le 1$, the following hold, all expectations being taken under the law of the process $\{X_t\}$ solving* (8)*:*

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} a_t^2 \, \mathbb{E}\big\|v_\star(X_t, t) - v_\star(X_{t_k}, t_k)\big\|^2 \mathrm{d}t \;\lesssim\; D^3 \big(\sigma^2 \vee 1\big) \, \delta^{-10\kappa_1} \, h, \tag{12}$$

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (a_t - a_{t_k})^2 \, \mathbb{E}\big\|v_\star(X_{t_k}, t_k)\big\|^2 \mathrm{d}t \;\lesssim\; D \, \delta^{-2(\kappa_1 + \kappa_2)} \, h^2. \tag{13}$$

In particular, writing $c_1 := 10\kappa_1 + 2\kappa_2$, the sum of the two is $\lesssim D^3(\sigma^2 \vee 1) \, \delta^{-c_1} h$.

Approximation Error A key component in controlling the first term in Lemma 7 is the approximation error of the function class $\mathcal{V}$ used to approximate $v_\star$. Neural network approximation theory has been studied extensively: Cybenko (1989) and Hornik (1991) established universal approximation for sigmoid-type single-hidden-layer networks on compact domains, and more recent work has characterized the expressive power and approximation rates of deep networks (Yarotsky, 2017; Petersen & Voigtlaender, 2018; Shen et al., 2022; Nakada & Imaizumi, 2020; Schmidt-Hieber, 2020).

Lemma 9. *For any $\epsilon \in (0,1)$ and any $R \geq 1$, there exists a network $v_0 \in \mathcal{R}(W, L, B, K)$ with $\sup_{x\in[-R,R]^D} \|v_0(x,t_i) - v_\star(x,t_i)\| \leq \epsilon$ for all $i = 0,\dots,N$, and $\mathcal{W}(v_0) \lesssim NR^D\delta^{-3\kappa_1 D}\epsilon^{-D}\big(\log(1/\delta)+\log(1/\epsilon)+\log R\big)$, $\mathcal{L}(v_0) \lesssim \log(1/\delta)+\log(1/\epsilon)+\log R$, $\mathcal{B}(v_0) \lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D}+N$, and $\mathcal{K}(v_0) \lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D}$. Further, $v_0(x,t) = 0$ if $x \notin [-(R+1), R+1]^D$, and $\mathrm{Lip}(v_0(\cdot,t)) \lesssim \delta^{-3\kappa_1}R$ for all $t \in [0, 1-\delta]$, where the implied constants depend only on D.*

The proof (detailed in Appendix C.1) then proceeds to control the first term in Lemma 7. For the estimator $\hat{v}_n$, this term can be upper bounded by a function of the approximation error, $\inf_{v\in\mathcal{V}}\sum_{j=0}^{N-1} h_j\mathbb{E}\left\|v(X_{t_j},t_j) - v_\star(X_{t_j},t_j)\right\|^2$ as shown in Lemma 9. For the estimator $\hat{v}^{\mathrm{mc}}_{n,m}$, controlling this term becomes more nuanced, as this involves controlling both an approximation error and a generalization error that accounts for estimating the expectation via Monte-Carlo in Equation (7) as shown in Lemma 14. Finally, we use the bound on $\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\right)$ to bound $\mathbb{W}_p\big(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\big)$ via truncation as shown in Lemma 21.

5 Conclusion and Discussions

In this work, we developed a theoretical understanding of flow matching for data distributions whose complexity is substantially lower than their ambient dimension. We established the first intrinsic-dimension-adaptive, high-probability generalization bound for flow matching in Wasserstein-p distance, for any $p \geq 1$: with appropriate choices of network architecture, early stopping, discretization scheme and truncation level, given n i.i.d. samples from P_{data}, and for every $d > d^*_p(P_{\text{data}})$, with probability at least $1-\xi$, $\mathbb{W}_p\big(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}_n))\big) \lesssim n^{-1/d} + n^{-\frac{1}{2p}}\log^{\frac{1}{2p}}(1/\xi)$, where $d^*_p(\cdot)$ is the upper Wasserstein-p dimension (Weed & Bach, 2019) and $\mathcal{T}_M$ truncates at level $M \asymp n^{2p/d}$. Since d may be taken arbitrarily close to $d^*_p(P_{\text{data}})$, the exponent is governed by the intrinsic dimension up to an arbitrarily small slack. Our analysis dispenses with the uniform Lipschitz assumption on the *learned* velocity field standard in state-of-the-art analyses—the regularity of $v_\star$ itself is established rather than assumed—and prescribes network sizes polynomial in n for each fixed D. The resulting rates avoid the ambient-dimensional curse of dimensionality and, for intrinsically low-dimensional distributions—including those on compact smooth manifolds—approach the corresponding minimax rates, giving a theoretical basis for the empirically observed effectiveness of flow matching in high dimensions with low-dimensional structure.

We also address several algorithmic considerations—*discretization*, *early stopping*, and *truncation of the generated law*—with theoretically motivated prescriptions. Our bounds imply an early stopping time $\delta = \Theta(n^{-\frac{1}{d\kappa_3}})$, ensuring the learned velocity field stays close to the ground truth while avoiding blow-up as $t \to 1^-$, and a truncation level $M = \Theta(n^{2p/d})$, balancing the tail of $P_{1-\delta}$ against the transport cost it induces. For numerical integration on $[0, 1-\delta]$, we recommend a uniform step size $h = \Theta\left(n^{-\frac{2p(2p+1)}{d} - \frac{10\kappa_1+2\kappa_2}{d\kappa_3}}\right)$. This is dictated by the Brownian component of the sampler: over a step of length h the increments of the generating SDE have mean square of order h rather than h^2, so the accumulated Euler–Maruyama error is $\mathcal{O}(\delta^{-(10\kappa_1+2\kappa_2)}h)$, and the step size above is what places it on the same order as the velocity estimation error. For the velocity network, ReLU feed-forward architectures suffice, with depth $L \asymp \log n$ and $W \asymp n^{\frac{p(2p+1)(D+2)}{d} + \frac{(20+7D)\kappa_1+4\kappa_2}{2d\kappa_3}}\log^{D/2+1} n$ weights. The exponent of n here is linear in D and decreasing in d.

While our results give detailed statistical guarantees, the *optimization error* from training non-convex, time-dependent velocity networks remains hard to control; our bounds are optimization-agnostic and can be combined with future optimization-theoretic advances. Two further directions remain open. First, our approximation analysis uses only the Lipschitz regularity of $v_\star$; exploiting its Gaussian-mixture structure, or additional smoothness, should yield smaller networks. Second, we assume a ReLU velocity network—standard in the statistical learning literature and compatible with existing approximation results for Hölder and Besov-type classes—whereas practical models use structured architectures such as U-Nets or transformers. These often exhibit strong empirical smoothness and hierarchical approximation properties, but rigorous guarantees for them remain unavailable.

DISCLOSURE OF AI USE

We used generative AI tools to assist with writing and editing experimental code, checking mathematical proofs developed by the authors, and polishing the manuscript for clarity and readability. The research ideas and original mathematical proofs were developed by the authors without generative AI assistance. The authors reviewed the AI-assisted writing, checked the experimental code, and independently verified the final mathematical proofs. We take full responsibility for the final content of this paper, including its text, mathematical claims and proofs, experimental results, and accompanying code.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric 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.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

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

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pp. 177–186, 2007.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising autoencoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Vladimir I. Bogachev, Nikolai V. Krylov, Michael Röckner, and Stanislav V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2015.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=MhK5aXo3gB. Expert Certification.

Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian FATRAS, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael M. Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kJFIH23hXb.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Charlotte Bunne, Andreas Krause, and marco cuturi. Supervised training of conditional monge maps. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=sPNtVVUq7wi.

Saptarshi Chakraborty. Minimax lower bounds for estimating distributions on low-dimensional spaces. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=wIgRV336hC.

Saptarshi Chakraborty and Peter Bartlett. A statistical analysis of wasserstein autoencoders for intrinsically low-dimensional data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WjRPZsfeBO.

Saptarshi Chakraborty and Peter 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. URL http://jmlr.org/papers/v26/24-0054.html.

Saptarshi Chakraborty, Quentin Berthet, and Peter L. Bartlett. Generalization properties of score-matching diffusion models for intrinsically low-dimensional data, 2026. URL https://arxiv.org/abs/2603.03700.

Fan Chen, Sinho Chewi, Constantinos Costis Daskalakis, and Alexander Rakhlin. High-accuracy sampling for diffusion models and log-concave distributions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GW3umRqsZZ.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/chen23q.html.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/chen23o.html.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c. URL https://openreview.net/forum?id=KD6MFeWSAd.

Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4462–4484. PMLR, 23–29 Jul 2023d. URL https://proceedings.mlr.press/v202/chen23e.html.

Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1955. ISBN 978-0070115422.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Biraj Dahal, Alexander Havrilla, Minshuo Chen, Tuo Zhao, and Wenjing Liao. On deep generative models for approximation and estimation of distributions on manifolds. *Advances in Neural Information Processing Systems*, 35:10615–10628, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin 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.

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

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

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TVQLu34bdw.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. *Journal of Machine Learning Research*, 23(116):1–43, 2022. URL http://jmlr.org/papers/v23/21-0732.html.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, Berlin, 2nd edition, 1991.

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning*, pp. 3398–3407. PMLR, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Andrey N Kolmogorov and Vladimir Mikhaılovich Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364, 1961.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow matching is adaptive to manifold structures, 2026. URL https://arxiv.org/abs/2602.22486.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=dUSI4vFyMK.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/lee23a.html.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

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

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion ODEs by high order denoising score matching. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14429–14460. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/lu22f.html.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020. URL http://jmlr.org/papers/v21/20-002.html.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pp. 38–50. Springer, 2003.

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

Sophia Pi, Mingcheng Lu, Maojiang Su, Weimin Wu, Jerry Yao-Chieh Hu, and Han Liu. Learning manifold data with flow matching. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=9mA5d705Gd.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

Edward C Posner, Eugene R Rodemich, and Howard Rumsey Jr. Epsilon entropy of stochastic processes. *The Annals of Mathematical Statistics*, pp. 1000–1020, 1967.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, 2015. ISBN 978-3-319-20827-5. doi: 10.1007/978-3-319-20828-2. URL https://doi.org/10.1007/978-3-319-20828-2.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Saurabh Singh and Ian Fischer. Stochastic sampling from deterministic flow models. *arXiv preprint arXiv:2410.02217*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1648–1656. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tang24a.html.

Terence Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011. ISBN 978-0-8218-6919-2. URL https://bookstore.ams.org/gsm-126.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2 edition, 2026.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648, 2019. doi: 10.3150/18-BEJ1065. URL https://doi.org/10.3150/18-BEJ1065.

Jingda Wu and Changxiao Cai. Diffusion models are statistically optimal for learning low-dimensional multi-modal distributions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=L5JTAPUdbQ.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=vES22INUKm.

APPENDICES

A Omitted Proofs from Section 2 **15**

A.1 Existence and Uniqueness of the True Vector Velocity Field (Proof of Theorem 4) . 15

A.2 Equivalence to the Conditional Flow-matching Objective 17

A.3 Equivalence of the probability-flow ODE and the SDE family 18

B Omitted Proofs from Section 4.2 **19**

B.1 Error Decomposition 19

B.2 Discretization Error 20

B.3 Approximation Result 22

C Proof of the Main Result **23**

C.1 Proof of Theorem 6 23

C.2 Supporting Lemmata 26

D Supporting Results **28**

D.1 Results for the Velocity Field 28

D.2 Results for the Wasserstein Distance 33

E Supporting Results from the Literature **34**

A OMITTED PROOFS FROM SECTION 2

A.1 EXISTENCE AND UNIQUENESS OF THE TRUE VECTOR VELOCITY FIELD (PROOF OF THEOREM 4)

Theorem 4. *For all $t \in [0, 1-\delta]$, P_t admits a density, ϱ_t, w.r.t. the Lebesgue measure and satisfies the continuity equation (in the weak sense), $\partial_t \varrho_t + \mathrm{div}_x(\varrho_t v_\star(\cdot, t)) = 0$, where $v_\star(x,t) = \beta_t^{-1}\dot{\beta}_t x + \beta_t^{-1}(\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t)\left(\sum_{i=1}^n \exp\left(-\frac{\|x-\alpha_t x_i\|^2}{2\beta_t^2}\right)\right)^{-1}\sum_{i=1}^n x_i \exp\left(-\frac{\|x-\alpha_t x_i\|^2}{2\beta_t^2}\right)$. Let*

$$y_x'(t) = v_\star(y_x(t), t), \text{ with } y_x(0) = x, \tag{4}$$

be the Initial Value Problem (IVP) with initial point x. Then (4) has a unique solution for every $x \in \mathbb{R}^D$, and $P_t = (Y_t)_\sharp P_0$, where $Y_t : x \mapsto y_x(t)$.

Proof. The characteristic function of X_t is given by,

$$\varphi_t(\kappa) = \mathbb{E}\exp(\iota\kappa^\top X_t) = \int \exp(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0))\mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1).$$

For any $t \in [0, 1-\delta]$, $\int |\varphi_t(\kappa)|\mathrm{d}\kappa \lesssim \delta^{-2\kappa_1}(2\pi)^{D/2} < \infty$, by Lemma 10. Hence, applying the Fourier inversion theorem, X_t has a density and is given by,

$$\varrho_t(x) = (2\pi)^{-D}\int \exp(-\iota\kappa^\top x)\varphi_t(\kappa)\mathrm{d}\kappa.$$

By Lemma 10, since,

$$\int \left\| \int (x_1 - x_0) \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1) \right\| \mathrm{d}\kappa$$
$$\lesssim \frac{\delta^{\kappa_1}\sqrt{2}\Gamma\left(\frac{D+1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)} \times \frac{\delta^{\kappa_1}(2\pi)^{D/2}}{\delta^{\kappa_1(D+1)}} + \frac{\sqrt{D}(2\pi)^{D/2}}{\delta^{\kappa_1 D}} < \infty, \tag{14}$$

Applying the Fubini-Tonelli Theorem (Tao, 2011, Corollary 1.7.23),

$$\begin{aligned}
&\partial_t \varrho_t(x) \\
=&(2\pi)^{-D} \int \exp(-\iota\kappa^\top x)\varphi_t'(\kappa)\mathrm{d}\kappa \\
=&(2\pi)^{-D} \int \iota\kappa^\top(\dot\alpha_t x_1 + \dot\beta_t x_0) \exp(-\iota\kappa^\top x) \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1)\mathrm{d}\kappa \\
=&-\operatorname{div}_x \left((2\pi)^{-D} \int (\dot\alpha_t x_1 + \dot\beta_t x_0) \exp(-\iota\kappa^\top x) \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1)\mathrm{d}\kappa \right) \\
=&-\operatorname{div}_x \Bigg(\int (\dot\alpha_t x_1 + \dot\beta_t x_0) \left((2\pi)^{-D} \int \exp(-\iota\kappa^\top x) \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}\kappa \right) \\
&\qquad \mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1) \Bigg).
\end{aligned} \tag{15}$$

We note that,

$$\begin{aligned}
&\int \exp(\iota\kappa^\top x) v_\star(x,t)\varrho_t(x)\mathrm{d}x \\
=&\mathbb{E}\left(\exp(\iota\kappa^\top X_t) v_\star(X_t,t)\right) \\
=&\mathbb{E}\left(\exp(\iota\kappa^\top X_t)\mathbb{E}\left(\dot X_t \mid X_t\right)\right) \\
=&\mathbb{E}\mathbb{E}\left(\exp(\iota\kappa^\top X_t)\dot X_t \mid X_t\right) \\
=&\mathbb{E}\left(\exp(\iota\kappa^\top X_t)\dot X_t\right) \\
=&\mathbb{E}\left(\exp(\iota\kappa^\top(\alpha_t X_1 + \beta_t X_0))(\dot\alpha_t X_1 + \dot\beta_t X_0)\right) \\
=&\int \exp(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0))(\dot\alpha_t x_1 + \dot\beta_t x_0)\mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1)
\end{aligned} \tag{16}$$

Taking an inverse Fourier transform on both sides, we get,

$$\begin{aligned}
&v_\star(x,t)\varrho_t(x) \\
=&(2\pi)^{-D} \int \exp(-\iota\kappa^\top x) \exp(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0))(\dot\alpha_t x_1 + \dot\beta_t x_0)\mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1)\mathrm{d}\kappa.
\end{aligned} \tag{17}$$

Note that we can apply the inverse Fourier transform since (14) is satisfied. Hence, from (15) and (17), we get,

$$\partial_t \varrho_t(x) + \operatorname{div}_x(v_\star(x,t)\varrho_t(x)) = 0. \tag{18}$$

Note that since $v_\star(\cdot,t)$ is uniformly (in $t \in [0, 1-\delta]$) $D\delta^{-3\kappa_1}$-Lipschitz, by Corollary 16, The Picard–Lindelöf theorem (Coddington & Levinson, 1955, Theorem 1.3.1) implies that (4) has a unique solution for all initialization $x \in \mathbb{R}^D$. From Santambrogio (2015, Theorem 4.4), $(Y_t)_\sharp P_0$ satisfies the continuity equation (18), which admits a unique solution. □

Lemma 10. *For any $0 \le t < 1$, the following holds,*

$$\int \left| \int \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{emp}(x_1) \right| \mathrm{d}\kappa \le \frac{(2\pi)^{D/2}}{\beta_t^D}, \tag{19}$$

$$\begin{aligned}
&\int \left\| \int (x_1 - z) \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{emp}(x_1) \right\| \mathrm{d}\kappa \\
\le&\frac{\dot\beta_t\sqrt{2}\Gamma\left(\frac{D+1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)} \times \frac{(2\pi)^{D/2}}{\beta_t^{D+1}} + \frac{\dot\alpha_t\sqrt{D}(2\pi)^{D/2}}{\beta_t^D}.
\end{aligned} \tag{20}$$

Proof.

$$
\begin{aligned}
&\int \left| \int \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1) \right| \mathrm{d}\kappa \\
\leq &\frac{1}{n}\sum_{i=1}^n \int \left| \int \exp\left(\iota\kappa^\top(\alpha_t x_i + \beta_t x_0)\right) \mathrm{d}P_0(x_0) \right| \mathrm{d}\kappa \\
= &\frac{1}{n}\sum_{i=1}^n \int \left| \exp(\iota\alpha_t\kappa^\top x_i) \exp\left(-\frac{1}{2}\beta_t^2\|\kappa\|^2\right) \right| \mathrm{d}\kappa \\
\leq &\int \exp\left(-\frac{1}{2}\beta_t^2\|\kappa\|^2\right) \mathrm{d}\kappa \\
= &\left(\frac{2\pi}{\beta_t^2}\right)^{D/2}.
\end{aligned}
$$

$$
\begin{aligned}
&\int \left\| \int (\dot\alpha_t x_1 + \dot\beta_t x_0) \exp\left(\iota\kappa^\top(\alpha_t x_1 + \beta_t x_0)\right) \mathrm{d}P_0(x_0)\mathrm{d}P_n^{\text{emp}}(x_1) \right\| \mathrm{d}\kappa \\
\leq &\frac{1}{n}\sum_{i=1}^n \int \left\| \int (\dot\alpha_t x_i + \dot\beta_t x_0) \exp\left(\iota\kappa^\top(\alpha_t x_i + \beta_t x_0)\right) \mathrm{d}P_0(x_0) \right\| \mathrm{d}\kappa \\
\leq &\frac{\dot\beta_t}{n}\sum_{i=1}^n \int \left\| \int x_0 \exp\left(\iota\kappa^\top(\alpha_t x_i + \beta_t x_0)\right) \mathrm{d}P_0(x_0) \right\| \mathrm{d}\kappa \\
&+ \frac{\dot\alpha_t}{n}\sum_{i=1}^n \int \left\| \int x_i \exp\left(\iota\kappa^\top(\alpha_t x_i + \beta_t x_0)\right) \mathrm{d}P_0(x_0) \right\| \mathrm{d}\kappa \\
\leq &\frac{\dot\beta_t}{n}\sum_{i=1}^n \int \left\| \int x_0 \exp\left(\iota\kappa^\top(\alpha_t x_i + \beta_t x_0)\right) \mathrm{d}P_0(x_0) \right\| \mathrm{d}\kappa \\
&+ \frac{\dot\alpha_t}{n}\sum_{i=1}^n \int \left\| x_i \exp\left(\iota\alpha_t\kappa^\top x_i\right) \int \exp\left(\iota\kappa^\top\beta_t x_0\right) \mathrm{d}P_0(x_0) \right\| \mathrm{d}\kappa \\
= &\frac{\dot\beta_t}{n}\sum_{i=1}^n \int \left\| \exp(\iota\alpha_t\kappa^\top x_i)\cdot \iota\beta_t\kappa \exp\left(-\frac{\beta_t^2\|\kappa\|^2}{2}\right) \right\| \mathrm{d}\kappa \\
&+ \frac{\dot\alpha_t}{n}\sum_{i=1}^n \int \left\| x_i \exp(\iota\alpha_t\kappa^\top x_i) \exp\left(-\frac{\beta_t^2\|\kappa\|^2}{2}\right) \right\| \mathrm{d}\kappa \\
\leq &\dot\beta_t \int \|\kappa\| \exp\left(-\frac{\beta_t^2\|\kappa\|^2}{2}\right) \mathrm{d}\kappa + \frac{\dot\alpha_t}{n}\sum_{i=1}^n \|x_i\| \int \exp\left(-\frac{\beta_t^2\|\kappa\|^2}{2}\right) \mathrm{d}\kappa \\
= &\frac{\sqrt{2}\Gamma\left(\frac{D+1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)} \times \frac{(2\pi)^{D/2}}{\beta_t^{D+1}} + \frac{\dot\alpha_t}{n}\sum_{i=1}^n \|x_i\| \times \frac{(2\pi)^{D/2}}{\beta_t^D} \\
\leq &\frac{\dot\beta_t\sqrt{2}\Gamma\left(\frac{D+1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)} \times \frac{(2\pi)^{D/2}}{\beta_t^{D+1}} + \frac{\dot\alpha_t\sqrt{D}(2\pi)^{D/2}}{\beta_t^D}.
\end{aligned}
$$

□

A.2 Equivalence to the Conditional Flow-matching Objective

Lemma 11. $\mathcal{L}_{\text{CFM}}(v) = \mathcal{L}(v) + \int_0^1 \mathbb{E}\|v_\star(X_t, t) - \dot X_t\|^2 \mathrm{d}t$.

Proof. Note that,

$\mathcal{L}_{\text{CFM}}(v)$

$$
\begin{aligned}
&= \int_0^1 \mathbb{E}\|v(X_t,t) - \dot{X}_t\|^2 \mathrm{d}t \\
&= \int_0^1 \mathbb{E}\left(\|v(X_t,t) - v_\star(X_t,t)\|^2 + \|v_\star(X_t,t) - \dot{X}_t\|^2 + 2\langle v(X_t,t) - v_\star(X_t,t), v_\star(X_t,t) - \dot{X}_t\rangle\right) \mathrm{d}t \\
&= \int_0^1 \mathbb{E}\|v(X_t,t) - v_\star(X_t,t)\|^2 \mathrm{d}t + \int_0^1 \mathbb{E}\|v_\star(X_t,t) - \dot{X}_t\|^2 \mathrm{d}t \\
&\quad + 2\int_0^1 \mathbb{E}\mathbb{E}\left(\langle v(X_t,t) - v_\star(X_t,t), v_\star(X_t,t) - \dot{X}_t\rangle \mid X_t\right) \mathrm{d}t \\
&= \mathcal{L}(v) + \int_0^1 \mathbb{E}\|v_\star(X_t,t) - \dot{X}_t\|^2 \mathrm{d}t.
\end{aligned}
$$

□

A.3 Equivalence of the probability-flow ODE and the SDE family

Let $\{\varrho_t\}_{t\in[0,1]}$ be the marginal law of the stochastic interpolant $X_t = \alpha_t X_1 + \beta_t X_0$, absolutely continuous with density ϱ_t for $t \in (0,1)$, and let $v_\star(x,t)$ denote the interpolant velocity field. For a constant $\sigma > 0$ define

$$
\mathfrak{b}_t(x) \;:=\; v_\star(x,t) + \frac{\sigma^2}{2}\nabla \log \varrho_t(x). \tag{21}
$$

By Theorem 4, $v_\star$ solves, in the sense of distributions on $(0,1)\times\mathbb{R}^D$,

$$
\partial_t \varrho_t + \nabla\cdot(v_\star \varrho_t) = 0, \qquad \varrho_t\big|_{t\to 0^+} = \varrho_0 = \mathcal{N}(0, I_D) \text{ weakly}. \tag{22}
$$

We first record the regularity of the score, which is what makes (23) below well posed.

Lemma 12 (Lipschitzness of the score)**.** *Under Assumption 1 and Design Choices D 1, for every $t \in [0, 1-\delta]$ the map $x \mapsto \nabla \log \varrho_t(x)$ is differentiable with*

$$
\nabla^2 \log \varrho_t(x) \;=\; \frac{\alpha_t^2}{\beta_t^4}\,\mathrm{Var}\big(X_1 \mid X_t = x\big) \;-\; \frac{1}{\beta_t^2}\, I_D, \qquad \big\|\nabla^2 \log \varrho_t(x)\big\| \;\lesssim\; D\,\delta^{-4\kappa_1},
$$

uniformly in $x \in \mathbb{R}^D$ and $t \in [0, 1-\delta]$. In particular $\nabla \log \varrho_t(\cdot)$ is globally Lipschitz with constant $\mathcal{O}(D\delta^{-4\kappa_1})$, uniformly in $t \in [0, 1-\delta]$.

Proof. By Lemma 15, $\nabla \log \varrho_t(x) = -\beta_t^{-2}\big(x - \alpha_t m(x,t)\big)$ with $m(x,t) = \mathbb{E}[X_1 \mid X_t = x]$, and $\nabla_x m(x,t) = \alpha_t \beta_t^{-2}\,\mathrm{Var}(X_1 \mid X_t = x)$. Differentiating the first identity in x and substituting the second gives the display. For the bound, $\mathrm{Var}(X_1 \mid X_t = x) \succeq 0$ and, since $X_1 \sim P_n^{\text{emp}}$ is supported in $[0,1]^D$ by Assumption 1, each conditional coordinate variance is at most $1/4$, so

$$
\big\|\mathrm{Var}(X_1 \mid X_t = x)\big\| \;\le\; \mathrm{trace}\,(\mathrm{Var}(X_1 \mid X_t = x)) \;\le\; \tfrac{D}{4}.
$$

With $\alpha_t \le 1$ and $\beta_t^{-1} \lesssim \delta^{-\kappa_1}$ by D 1(a), $\|\nabla^2 \log \varrho_t(x)\| \le \frac{D}{4}\alpha_t^2\beta_t^{-4} + \beta_t^{-2} \lesssim D\delta^{-4\kappa_1}$. □

Theorem 13 (Marginal equivalence of the ODE and the SDE family)**.** *Fix $\delta \in (0, 1/2)$. Under Assumption 1 and Design Choices D 1, there is a pathwise unique strong solution $\{X_t\}_{t\in[0,1-\delta]}$ of*

$$
dX_t = \mathfrak{b}_t(X_t)\,dt + \sigma\, dW_t, \qquad X_0 \sim \mathcal{N}(0, I_D), \tag{23}
$$

with $\mathrm{Law}(X_t) = \varrho_t$ for every $t \in [0, 1-\delta]$.

Proof. Let $\mathcal{L}_t^\sigma f := \mathfrak{b}_t \cdot \nabla f + \frac{\sigma^2}{2}\Delta f$ be the generator of (23), with formal adjoint $(\mathcal{L}_t^\sigma)^*$. Since ϱ_t is a Gaussian smoothing of a compactly supported measure it is smooth and strictly positive on $\mathbb{R}^D$ for every $t \in (0,1)$, so the following manipulations are classical. Using the score identity $\varrho_t \nabla \log \varrho_t = \nabla \varrho_t$,

$$
\begin{aligned}
(\mathcal{L}_t^\sigma)^* \varrho_t &= -\nabla\cdot(\mathfrak{b}_t \varrho_t) + \frac{\sigma^2}{2}\Delta \varrho_t \\
&= -\nabla\cdot\Big(v_\star \varrho_t + \frac{\sigma^2}{2}\nabla \varrho_t\Big) + \frac{\sigma^2}{2}\Delta \varrho_t \\
&= -\nabla\cdot(v_\star \varrho_t) \\
&= \partial_t \varrho_t.
\end{aligned} \tag{24}
$$

In the above calculations, (24) follows from (22). This holds *for every* $\sigma \geq 0$ simultaneously: the diffusive term $\frac{\sigma^2}{2}\Delta\varrho_t$ contributed by the noise is exactly cancelled by the score correction in (21), not merely asymptotically, and (24) does not depend on σ at all. Hence $t \mapsto \varrho_t\, dx$ is a probability solution of

$$\partial_t \mu_t = (\mathcal{L}_t^\sigma)^* \mu_t, \qquad t \in [0, 1-\delta], \qquad \mu_0 = \varrho_0 = \mathcal{N}(0, I_D), \tag{25}$$

for every admissible σ, the initial condition being the weak limit in (22).

By Corollary 16, $v_\star(\cdot, t)$ is $\mathcal{O}(D\delta^{-3\kappa_1})$-Lipschitz uniformly over $t \in [0, 1-\delta]$, and by Lemma 12, $\nabla \log \varrho_t(\cdot)$ is $\mathcal{O}(D\delta^{-4\kappa_1})$-Lipschitz uniformly over the same range. Hence

$$\mathfrak{b}_t(x) = v_\star(x,t) + \tfrac{\sigma^2}{2}\nabla \log \varrho_t(x)$$

is globally Lipschitz in x, uniformly over $t \in [0, 1-\delta]$, with Lipschitz constant $\Lambda_\delta := \mathcal{O}\big(D(1 \vee \sigma^2)\delta^{-4\kappa_1}\big)$, and is of at most linear growth, since $\|\mathfrak{b}_t(0)\| \lesssim \delta^{-\kappa_1}\sqrt{D}$ by Corollary 16 and $\|\nabla \log \varrho_t(0)\| \leq \beta_t^{-2}\alpha_t \mathbb{E}\|X_1\| \lesssim \delta^{-2\kappa_1}\sqrt{D}$. Since σ is a fixed positive constant, the classical Itô existence–uniqueness theorem for SDEs with globally Lipschitz coefficients of linear growth (Karatzas & Shreve, 1991, Theorem 2.9, Chapter 5) gives a pathwise unique strong solution $\{X_t\}_{t\in[0,1-\delta]}$ of (23) for any initial law with finite second moment, in particular for $\mathrm{Law}(X_0) = \mathcal{N}(0, I_D)$.

The drift $\mathfrak{b}$ is globally Lipschitz in x uniformly in $t \in [0, 1-\delta]$ and of linear growth, and the diffusion coefficient σI_D is constant and nondegenerate, so the Cauchy problem (25) admits at most one probability solution on $[0, 1-\delta]$ with initial datum ϱ_0 (Bogachev et al., 2015, Ch. 9). Equivalently, and without appeal to that theory: $t \mapsto \mathrm{Law}(X_t)$ for the pathwise-unique solution of (23) is itself a probability solution of (25) with the same initial datum, and two probability solutions of a well-posed SDE with the same initial law have equal marginals. From the above arguments, the map, $t \mapsto \varrho_t$ is also a probability solution of (25) with $\varrho_t|_{t=0} = \varrho_0 = \mathcal{N}(0, I_D)$. Uniqueness therefore forces $\mathrm{Law}(X_t) = \varrho_t$ for every $t \in [0, 1-\delta]$. □

B Omitted Proofs from Section 4.2

B.1 Error Decomposition

Lemma 7. *Under Assumption 1 and Design Choices D 1, for any estimator $\hat{v} \in \mathcal{V}$ of $v_\star$,*

$$\mathrm{KL}\left(P_{1-\delta} \,\|\, \hat{P}_{1-\delta}(\hat{v})\right) \leq \frac{2}{\sigma^2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} a_t^2\, \mathbb{E}\big\|v_\star(X_t,t) - v_\star(X_{t_k},t_k)\big\|^2 \mathrm{d}t$$

$$+ \frac{2}{\sigma^2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} (a_t - a_{t_k})^2\, \mathbb{E}\big\|v_\star(X_{t_k},t_k)\big\|^2 \mathrm{d}t + \frac{1}{\sigma^2}\sum_{k=0}^{N-1} h_k\, a_{t_k}^2\, \mathbb{E}\big\|v_\star(X_{t_k},t_k) - \hat{v}(X_{t_k},t_k)\big\|^2,$$

where every expectation is taken under the law of the process $\{X_t\}_{t\in[0,1-\delta]}$ solving (8)*.*

Proof. Write $\Omega = C([0, 1-\delta]; \mathbb{R}^D)$ for path space and let $\boldsymbol{P}$ denote the law on Ω of the solution $\{X_t\}$ of

$$\mathrm{d}X_t = \big(a_t v_\star(X_t,t) - b_t X_t\big)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad X_0 \sim \mathcal{N}(0, I_D),$$

which exists and is pathwise unique by Theorem 13. Let $\boldsymbol{Q}$ denote the law on Ω of the Euler–Maruyama process (9), namely the solution of

$$\mathrm{d}\widehat{X}_t = \big(a_{t_k}\hat{v}(\widehat{X}_{t_k},t_k) - b_t\widehat{X}_t\big)\mathrm{d}t + \sigma \mathrm{d}W_t, \quad t \in [t_k, t_{k+1}), \qquad \widehat{X}_0 \sim \mathcal{N}(0, I_D).$$

Both are driven by the same diffusion coefficient $\sigma > 0$ and started from the same law, so $\boldsymbol{P}$ and $\boldsymbol{Q}$ differ only through their drifts. Writing these as functionals of the path,

$$\mu_t^{\boldsymbol{P}}(\omega) = a_t v_\star(\omega_t, t) - b_t\omega_t, \qquad \mu_t^{\boldsymbol{Q}}(\omega) = a_{t_k}\hat{v}(\omega_{t_k}, t_k) - b_t\omega_t \quad \text{for } t \in [t_k, t_{k+1}),$$

both are adapted, and the linear terms $-b_t\omega_t$ coincide, so that

$$\mu_t^{\boldsymbol{P}}(\omega) - \mu_t^{\boldsymbol{Q}}(\omega) = a_t v_\star(\omega_t,t) - a_{t_k}\hat{v}(\omega_{t_k},t_k), \qquad t \in [t_k, t_{k+1}). \tag{26}$$

By Corollary 16, $\|v_\star(x,t)\| \lesssim \delta^{-\kappa_1}(\|x\|+\sqrt{D})$ uniformly on $[0,1-\delta]$; by D 1(a), $\sup_{t\le 1-\delta}|a_t| \lesssim \delta^{-\kappa_1/2}$; and $\|\hat v(\cdot,\cdot)\| \le \sqrt{D}\,\mathcal{K}(\hat v) \le \sqrt{D}K$. Hence the drift difference (26) satisfies

$$\big\|\mu_t^{\boldsymbol{P}}(\omega)-\mu_t^{\boldsymbol{Q}}(\omega)\big\| \;\le\; C_\delta\Big(1+\sup_{s\le t}\|\omega_s\|\Big), \qquad t\in[0,1-\delta], \qquad C_\delta \;\lesssim\; \delta^{-\kappa_1/2}\sqrt{D}\,\big(\delta^{-\kappa_1}+K\big),$$

which is finite for each fixed $\delta>0$. Since Beneš's condition requires only linear growth of the drift difference in the sup-norm of the path, with no restriction on the constant, it holds here (Karatzas & Shreve, 1991, Corollary 3.5.16), and the stochastic exponential below is a true martingale. Girsanov's theorem (Øksendal, 2003, Theorem 8.6.5) therefore gives $\boldsymbol{Q}\sim\boldsymbol{P}$ with

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}\boldsymbol{Q}}(\omega)=\exp\left(\frac{1}{\sigma}\int_0^{1-\delta}\big\langle \mu_t^{\boldsymbol{P}}-\mu_t^{\boldsymbol{Q}},\mathrm{d}W_t\big\rangle-\frac{1}{2\sigma^2}\int_0^{1-\delta}\big\|\mu_t^{\boldsymbol{P}}-\mu_t^{\boldsymbol{Q}}\big\|^2\mathrm{d}t\right),$$

and taking $\mathbb{E}_{\boldsymbol{P}}\log(\mathrm{d}\boldsymbol{P}/\mathrm{d}\boldsymbol{Q})$, the stochastic integral being a true martingale under $\boldsymbol{P}$,

$$\mathrm{KL}(\boldsymbol{P}\,\|\,\boldsymbol{Q})=\frac{1}{2\sigma^2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\big\|a_t v_\star(X_t,t)-a_{t_k}\hat v(X_{t_k},t_k)\big\|^2\mathrm{d}t. \tag{27}$$

The marginal laws at the terminal time are $P_{1-\delta}=\boldsymbol{P}\circ\pi_{1-\delta}^{-1}$ and $\hat P_{1-\delta}(\hat v)=\boldsymbol{Q}\circ\pi_{1-\delta}^{-1}$, where $\pi_{1-\delta}(\omega)=\omega_{1-\delta}$, the first identification being Theorem 13. Since the Kullback–Leibler divergence contracts under the pushforward by $\pi_{1-\delta}$,

$$\mathrm{KL}\Big(P_{1-\delta}\,\|\,\hat P_{1-\delta}(\hat v)\Big)\le\mathrm{KL}(\boldsymbol{P}\,\|\,\boldsymbol{Q}). \tag{28}$$

It remains to split (27). Adding and subtracting $a_{t_k}v_\star(X_{t_k},t_k)$ and using $\|u+v\|^2\le 2\|u\|^2+2\|v\|^2$,

$$\begin{aligned}\mathrm{KL}(\boldsymbol{P}\,\|\,\boldsymbol{Q})\le{}&\frac{1}{\sigma^2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\big\|a_t v_\star(X_t,t)-a_{t_k}v_\star(X_{t_k},t_k)\big\|^2\mathrm{d}t\\&+\frac{1}{\sigma^2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}a_{t_k}^2\,\mathbb{E}\big\|v_\star(X_{t_k},t_k)-\hat v(X_{t_k},t_k)\big\|^2\mathrm{d}t.\end{aligned}$$

The second sum has an integrand constant in t on each $[t_k,t_{k+1})$ and so equals $\sigma^{-2}\sum_k h_k a_{t_k}^2\,\mathbb{E}\|v_\star(X_{t_k},t_k)-\hat v(X_{t_k},t_k)\|^2$. For the first, adding and subtracting $a_t v_\star(X_{t_k},t_k)$ and splitting once more,

$$\big\|a_t v_\star(X_t,t)-a_{t_k}v_\star(X_{t_k},t_k)\big\|^2\le 2a_t^2\big\|v_\star(X_t,t)-v_\star(X_{t_k},t_k)\big\|^2+2(a_t-a_{t_k})^2\big\|v_\star(X_{t_k},t_k)\big\|^2.$$

Combining the three displays with (28) gives the claim. □

B.2 Discretization Error

Lemma 8. *Under Assumption 1 and Design Choices D 1, with uniform step size $h\le 1$, the following hold, all expectations being taken under the law of the process $\{X_t\}$ solving* (8)*:*

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}a_t^2\,\mathbb{E}\big\|v_\star(X_t,t)-v_\star(X_{t_k},t_k)\big\|^2\mathrm{d}t\;\lesssim\;D^3\big(\sigma^2\vee 1\big)\,\delta^{-10\kappa_1}\,h, \tag{12}$$

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}(a_t-a_{t_k})^2\,\mathbb{E}\big\|v_\star(X_{t_k},t_k)\big\|^2\mathrm{d}t\;\lesssim\;D\,\delta^{-2(\kappa_1+\kappa_2)}\,h^2. \tag{13}$$

Proof. Throughout, $t\in[t_k,t_{k+1})$ for the relevant k, and all expectations are under the law of $\{X_t\}_{t\in[0,1-\delta]}$ solving (8). By Theorem 13, $\mathrm{Law}(X_s)=\varrho_s=\mathrm{Law}(\alpha_s X_1+\beta_s X_0)$ for each fixed s. Since $X_1\sim P_n^{\mathrm{emp}}$ is supported in $[0,1]^D$ by Assumption 1 and $X_0\sim\mathcal{N}(0,I_D)$,

$$\mathbb{E}\|X_s\|^2\le 2\alpha_s^2\,\mathbb{E}\|X_1\|^2+2\beta_s^2\,\mathbb{E}\|X_0\|^2\lesssim D,\qquad s\in[0,1-\delta]. \tag{29}$$

Combining (29) with Corollary 16,

$$\mathbb{E}\big\|v_\star(X_s,s)\big\|^2 \lesssim \delta^{-2\kappa_1}\Big(\mathbb{E}\|X_s\|^2 + D\Big) \lesssim D\,\delta^{-2\kappa_1}. \tag{30}$$

We write $\mu_s(x) := a_s v_\star(x,s) - b_s x$ for the drift of (8). By (30), D 1(a) (which gives $a_s^2 \lesssim \delta^{-\kappa_1}$ and $|b_s| \lesssim \delta^{-\kappa_1}$) and (29),

$$\mathbb{E}\|\mu_s(X_s)\|^2 \le 2a_s^2\,\mathbb{E}\|v_\star(X_s,s)\|^2 + 2b_s^2\,\mathbb{E}\|X_s\|^2 \lesssim D\,\delta^{-3\kappa_1}. \tag{31}$$

Integrating (8) from t_k to t,

$$X_t - X_{t_k} = \int_{t_k}^{t} \mu_s(X_s)\,\mathrm{d}s + \sigma\big(W_t - W_{t_k}\big).$$

By $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$, the Cauchy–Schwarz inequality applied to the drift integral, and $\mathbb{E}\|W_t - W_{t_k}\|^2 = D(t-t_k)$,

$$\begin{aligned}\mathbb{E}\big\|X_t - X_{t_k}\big\|^2 &\le 2(t-t_k)\int_{t_k}^{t}\mathbb{E}\|\mu_s(X_s)\|^2\,\mathrm{d}s + 2\sigma^2 D\,(t-t_k)\\ &\lesssim D\,\delta^{-3\kappa_1}(t-t_k)^2 + \sigma^2 D\,(t-t_k)\\ &\lesssim\ D\big(\sigma^2 + \delta^{-3\kappa_1}h\big)(t-t_k).\end{aligned} \tag{32}$$

Note that the σ^2 appearing here is exactly cancelled by the factor σ^{-2} carried by Lemma 7, so the resulting contribution to the KL bound is independent of σ; see Remark 1 below.

Splitting in space and in time,

$$\begin{aligned}&\mathbb{E}\big\|v_\star(X_t,t) - v_\star(X_{t_k},t_k)\big\|^2\\ &\le 2\,\mathbb{E}\big\|v_\star(X_t,t) - v_\star(X_{t_k},t)\big\|^2 + 2\,\mathbb{E}\big\|v_\star(X_{t_k},t) - v_\star(X_{t_k},t_k)\big\|^2.\end{aligned} \tag{33}$$

For the first summand, Corollary 16 gives $\mathrm{Lip}(v_\star(\cdot,t)) \lesssim D\delta^{-3\kappa_1}$ uniformly in $t \in [0, 1-\delta]$. Thus, from (32),

$$\mathbb{E}\big\|v_\star(X_t,t) - v_\star(X_{t_k},t)\big\|^2 \lesssim D^2\delta^{-6\kappa_1}\,\mathbb{E}\|X_t - X_{t_k}\|^2 \lesssim D^3\delta^{-6\kappa_1}\big(\sigma^2 + \delta^{-3\kappa_1}h\big)(t-t_k). \tag{34}$$

For the second summand, the two evaluations share the spatial argument X_{t_k}, so only the one-time marginal of X_{t_k} is involved. By Lemma 18, $\|\partial_t v_\star(x,t)\| \lesssim \delta^{-\kappa_1}(1+\|x\|)$ uniformly on $[0, 1-\delta]$. Hence, by the mean value theorem and (29),

$$\mathbb{E}\big\|v_\star(X_{t_k},t) - v_\star(X_{t_k},t_k)\big\|^2 \lesssim \delta^{-2\kappa_1}\,\mathbb{E}\big(1+\|X_{t_k}\|^2\big)(t-t_k)^2 \lesssim D\,\delta^{-2\kappa_1}(t-t_k)^2. \tag{35}$$

Using $a_t^2 \lesssim \delta^{-\kappa_1}$, then (33), (34) and (35), and finally $\int_{t_k}^{t_{k+1}}(t-t_k)\mathrm{d}t = h^2/2$ and $\int_{t_k}^{t_{k+1}}(t-t_k)^2\mathrm{d}t = h^3/3$,

$$\begin{aligned}&\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} a_t^2\,\mathbb{E}\big\|v_\star(X_t,t) - v_\star(X_{t_k},t_k)\big\|^2\mathrm{d}t\\ &\lesssim \delta^{-\kappa_1}\left[D^3\delta^{-6\kappa_1}\big(\sigma^2 + \delta^{-3\kappa_1}h\big)\sum_{k=0}^{N-1}\frac{h^2}{2} + D\,\delta^{-2\kappa_1}\sum_{k=0}^{N-1}\frac{h^3}{3}\right]\\ &\lesssim D^3\delta^{-7\kappa_1}\big(\sigma^2 + \delta^{-3\kappa_1}h\big)\,h + D\,\delta^{-3\kappa_1}h^2\\ &\lesssim D^3\big(\sigma^2 \vee 1\big)\,\delta^{-10\kappa_1}\,h,\end{aligned}$$

where the second line uses $Nh = 1-\delta \le 1$, so that $\sum_k h^2 \le h$ and $\sum_k h^3 \le h^2$, and the last uses $h \le 1$ to absorb h^2 into h. This is (12).

To bound (13), we note that this term involves $v_\star$ at the single time t_k only, so (30) applies directly. By D 1(b), $|a_t - a_{t_k}| \lesssim \delta^{-\kappa_2}(t-t_k)$, whence

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}(a_t - a_{t_k})^2\,\mathbb{E}\big\|v_\star(X_{t_k},t_k)\big\|^2\mathrm{d}t \lesssim D\,\delta^{-2\kappa_1-2\kappa_2}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}(t-t_k)^2\mathrm{d}t \lesssim D\,\delta^{-2(\kappa_1+\kappa_2)}h^2,$$

using $\sum_k h^3 \le h^2$. This is (13). Since $h \le 1$ and $2(\kappa_1+\kappa_2) \le 10\kappa_1 + 2\kappa_2$, the two bounds combine as stated. □

Remark 1. The diffusion coefficient cannot be tuned to remove the $O(h)$ term. Lemma 7 carries a factor σ^{-2}, while the leading term of (12) carries σ^2 through the Brownian increment in (32); the two cancel, and the discretization contribution to $\mathrm{KL}(P_{1-\delta}\|\hat{P}_{1-\delta}(\hat{v}))$ is of order $D^3\delta^{-c_1}h$, uniformly in σ.

B.3 Approximation Result

Lemma 9. *For any $\epsilon \in (0,1)$ and any $R \geq 1$, there exists a network $v_0 \in \mathcal{R}(W,L,B,K)$ with $\sup_{x\in[-R,R]^D}\|v_0(x,t_i) - v_\star(x,t_i)\| \leq \epsilon$ for all $i = 0,\dots,N$, and $\mathcal{W}(v_0) \lesssim NR^D\delta^{-3\kappa_1 D}\epsilon^{-D}\big(\log(1/\delta)+\log(1/\epsilon)+\log R\big)$, $\mathcal{L}(v_0)\lesssim \log(1/\delta)+\log(1/\epsilon)+\log R$, $\mathcal{B}(v_0)\lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D}+N$, and $\mathcal{K}(v_0)\lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D}$. Further, $v_0(x,t)=0$ if $x\notin[-(R+1),R+1]^D$, and $\mathrm{Lip}(v_0(\cdot,t))\lesssim \delta^{-3\kappa_1}R$ for all $t\in[0,1-\delta]$, where the implied constants depend only on D.*

Proof. Fix $\epsilon\in(0,1)$ and $R\geq 1$. By Corollary 16, for every $t\in[0,1-\delta]$,

$$\mathrm{Lip}\big(v_\star(\cdot,t)\big) =: L_t \;\lesssim\; L_\delta \asymp \delta^{-3\kappa_1}, \quad \sup_{x\in[-R,R]^D}\|v_\star(x,t)\| \;\lesssim\; \delta^{-\kappa_1}\big(R\sqrt{D}+\sqrt{D}\big) \;\lesssim\; \delta^{-\kappa_1}R\sqrt{D}. \tag{36}$$

Let,

$$K_0 \;:=\; c\,\delta^{-(3\kappa_1)\vee 1}R\sqrt{D} \tag{37}$$

with c a large enough absolute constant that $K_0 \geq 1+\sup_{t\in[0,1-\delta]}\sup_{x\in[-R,R]^D}\|v_\star(x,t)\|_\infty$, which is possible by (36). Fix $i\in\{0,\dots,N\}$ and $\ell\in[D]$. Define $g_i^{(\ell)}:[0,1]^D\to\mathbb{R}$ by

$$g_i^{(\ell)}(y) \;=\; \frac{1}{2RL_{t_i}}\Big(v_\star^{(\ell)}\big(2Ry - R\mathbf{1},\,t_i\big) - v_\star^{(\ell)}\big(-R\mathbf{1},\,t_i\big)\Big).$$

Since $x\mapsto v_\star^{(\ell)}(x,t_i)$ is L_{t_i}-Lipschitz, $g_i^{(\ell)}$ is 1-Lipschitz on $[0,1]^D$ and $g_i^{(\ell)}(0)=0$. Let $h_i^{(\ell)}$ be the ε-approximating ReLU network for $g_i^{(\ell)}$ furnished by Lemma 22, and set

$$\tilde{v}_{t_i}^{(\ell)}(x) \;=\; 2RL_{t_i}\,h_i^{(\ell)}\left(\frac{x+R\mathbf{1}}{2R}\right) + v_\star^{(\ell)}\big(-R\mathbf{1},t_i\big).$$

By construction,

$$\begin{aligned}\sup_{x\in[-R,R]^D}\big|\tilde{v}_{t_i}^{(\ell)}(x) - v_\star^{(\ell)}(x,t_i)\big| &= 2RL_{t_i}\sup_{y\in[0,1]^D}\Big|h_i^{(\ell)}(y)-g_i^{(\ell)}(y)\Big| \\ &\leq 2RL_{t_i}\varepsilon \\ &\lesssim R\,L_\delta\,\varepsilon.\end{aligned} \tag{38}$$

Let $\tilde{v}_{t_i}$ be obtained by stacking $\{\tilde{v}_{t_i}^{(\ell)}\}_{\ell=1}^D$. By Lemma 22, $\|h_i^{(\ell)}\|_\infty \leq \sqrt{D}$, so on all of $\mathbb{R}^D$ $\mathcal{W}(\tilde{v}_{t_i})\lesssim D\big(\varepsilon^{-D}\log(1/\varepsilon)+D\big)$, $\mathcal{L}(\tilde{v}_{t_i})\lesssim\log(1/\varepsilon)+D$, $\mathcal{B}(\tilde{v}_{t_i})\lesssim RL_\delta+\delta^{-\kappa_1}R\sqrt{D}\lesssim K_0$, and, since $h_i^{(\ell)}$ is $10D$-Lipschitz in ℓ_1 and the inner affine map has gain $1/(2R)$,

$$\mathrm{Lip}(\tilde{v}_{t_i}) \;\leq\; 2RL_{t_i}\cdot 10D\cdot\tfrac{1}{2R}\cdot\sqrt{D} \;\lesssim\; D^{3/2}L_\delta. \tag{39}$$

We define the clipping function, $\mathrm{cl}_{K_0}(u) = \mathrm{ReLU}(u+K_0)-\mathrm{ReLU}(u-K_0)-K_0$, so that $\mathrm{cl}_{K_0}(u)=u$ for $|u|\leq K_0$, $|\,\mathrm{cl}_{K_0}|\leq K_0$ everywhere, and cl_{K_0} is 1-Lipschitz; write $\mathrm{cl}_{K_0}(\tilde{v}_{t_i})$ for its coordinatewise application. By (36), (37) and (38), clipping does not alter $\tilde{v}_{t_i}$ on $[-R,R]^D$ once $\varepsilon\leq 1$. Next set

$$\chi(x) \;=\; \mathrm{ReLU}\left(1-\sum_{\ell=1}^{D}\mathrm{ReLU}\big(\mathrm{ReLU}(x_\ell)+\mathrm{ReLU}(-x_\ell)-R\big)\right),$$

a ReLU network of width $O(D)$ and depth 3 with $\chi\equiv 1$ on $[-R,R]^D$, $\chi\equiv 0$ off $[-(R+1),R+1]^D$, $0\leq\chi\leq 1$, and $\mathrm{Lip}(\chi)\lesssim D$. We define $t_{-1}=-1$ and $t_{N+1}=1$, let $\delta_i=\frac{1}{2}\big((t_i-t_{i-1})\wedge(t_{i+1}-t_i)\big)$ and

$$\xi_{a,b}(x) = \mathrm{ReLU}\left(\frac{x+a}{a-b}\right)-\mathrm{ReLU}\left(\frac{x+b}{a-b}\right)-\mathrm{ReLU}\left(\frac{x-b}{a-b}\right)+\mathrm{ReLU}\left(\frac{x-a}{a-b}\right),$$

so that $\xi_i := \xi_{\delta_i/2,\,\delta_i/4}(\cdot - t_i)$ satisfies $0 \le \xi_i \le 1$, $\xi_i \equiv 1$ on $|t - t_i| \le \delta_i/4$, and $\operatorname{supp}\xi_i \subseteq \{|t - t_i| \le \delta_i/2\}$; its weights are of order $1/\delta_i \asymp N$. Since $\delta_i \le \frac{1}{2}|t_i - t_j|$ for $j \ne i$, the supports of $\{\xi_i\}$ are pairwise disjoint and $\xi_j(t_i) = 0$ for $j \ne i$. Let

$$\psi_i(x,t) \;=\; \min\big(\chi(x), \xi_i(t)\big) \;=\; \chi(x) - \mathrm{ReLU}\big(\chi(x) - \xi_i(t)\big),$$

which is again a ReLU network, takes values in $[0,1]$, equals 1 on $[-R,R]^D \times \{|t - t_i| \le \delta_i/4\}$, and vanishes whenever $\chi(x) = 0$ or $\xi_i(t) = 0$.

Finally, let Π denote the ReLU approximate-multiplication network on $[-K_0, K_0] \times [0,1]$ of accuracy $\epsilon/(4\sqrt{D})$, of width and depth $\mathcal{O}\big(\log(K_0\sqrt{D}/\epsilon)\big)$, and satisfying $\Pi(u,0) = 0$ identically. We define

$$v_0^{(\ell)}(x,t) \;=\; \sum_{i=0}^{N} \Pi\Big(\mathrm{cl}_{K_0}\big(\tilde{v}_{t_i}^{(\ell)}(x)\big),\; \psi_i(x,t)\Big), \qquad \ell \in [D].$$

Fix $i \in \{0,\dots,N\}$ and $x \in [-R,R]^D$. Then $\psi_j(x,t_i) = 0$ for $j \ne i$, so those summands vanish exactly, while $\psi_i(x,t_i) = 1$. Hence, by (38),

$$\big|v_0^{(\ell)}(x,t_i) - v_\star^{(\ell)}(x,t_i)\big| \;\le\; \frac{\epsilon}{4\sqrt{D}} + 2RL_{t_i}\varepsilon.$$

Choosing $\varepsilon = \epsilon\big/\big(8\sqrt{D}\,RL_\delta\big)$ makes the second term at most $\epsilon/(4\sqrt{D})$, so that the coordinatewise error is at most $\epsilon/(2\sqrt{D})$ and

$$\sup_{x\in[-R,R]^D} \big\|v_0(x,t_i) - v_\star(x,t_i)\big\| \;\le\; \sqrt{D}\cdot\frac{\epsilon}{2\sqrt{D}} \;\le\; \epsilon \qquad \text{for every } i = 0,\dots,N.$$

With this ε we have $\varepsilon^{-D} \lesssim R^D\delta^{-3\kappa_1 D}\epsilon^{-D}$ and $\log(1/\varepsilon) \asymp \log(K_0\sqrt{D}/\epsilon) \asymp \log(1/\delta) + \log(1/\epsilon) + \log R$, whence, summing the $N+1$ parallel branches and recalling that D is fixed,

$$\mathcal{W}(v_0) \lesssim N\Big(\varepsilon^{-D}\log(1/\varepsilon) + \log(K_0\sqrt{D}/\epsilon) + D\Big) \lesssim NR^D\delta^{-3\kappa_1 D}\epsilon^{-D}\big(\log(1/\delta) + \log(1/\epsilon) + \log R\big),$$

$$\mathcal{L}(v_0) \lesssim \log(1/\varepsilon) + \log(K_0\sqrt{D}/\epsilon) + D \lesssim \log(1/\delta) + \log(1/\epsilon) + \log R,$$

$$\mathcal{B}(v_0) \lesssim \max\big\{K_0,\, 1/\delta_i,\, 1\big\} \lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D} + N,$$

$$\mathcal{K}(v_0) \lesssim K_0 \lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D}.$$

Moreover $v_0(x,t) = 0$ for $x \notin [-(R+1), R+1]^D$, since then $\chi(x) = 0$, hence $\psi_i(x,t) = 0$ for every i and $\Pi(\cdot,0) = 0$. Finally, for fixed t the map $x \mapsto \Pi(u(x),\psi(x))$ has $|\partial_u\Pi| \le 1$ and $|\partial_\psi\Pi| \lesssim K_0$, so by (39) and $\mathrm{Lip}(\chi) \lesssim D$,

$$\mathrm{Lip}\big(v_0(\cdot,t)\big) \;\lesssim\; D^{3/2}L_\delta + DK_0 \;\lesssim\; D^{5/2}\,\delta^{-3\kappa_1}R, \qquad t \in [0, 1-\delta]. \tag{40}$$

□

C Proof of the Main Result

C.1 Proof of Theorem 6

Theorem 6 (Main Theorem). *Suppose that $d > d_p^*(P_{data})$ and Assumption 1 holds. Then under Design Choices 1, for any $p \ge 1$, there exists a constant $n_0 \in \mathbb{N}$, such that if $n \ge n_0$, and $\mathcal{V} \equiv \mathcal{RN}(W,L,B,K)$ with $\delta \asymp n^{-\frac{1}{d\kappa_3}}$, $h \asymp n^{-\frac{2p(2p+1)}{d} - \frac{10\kappa_1+2\kappa_2}{d\kappa_3}}$, $M \asymp n^{\frac{2p}{d}}$, then, with probability at least $1-\xi$,*

$$\mathbb{W}_p\left(P_{data}, \mathcal{T}_M\left(\hat{P}_{1-\delta}(\hat{v}_n)\right)\right) \lesssim n^{-1/d} + n^{-\frac{1}{2p}}\log^{\frac{1}{2p}}(1/\xi). \tag{10}$$

Here, $W \asymp n^{\frac{p(2p+1)(D+2)}{d} + \frac{(20+7D)\kappa_1+4\kappa_2}{2d\kappa_3}}\log^{D/2+1} n$, $L \asymp \log n$, $B \asymp n^{\left(\frac{2p(2p+1)}{d} + \frac{20\kappa_1+4\kappa_2}{2d\kappa_3}\right)\vee\frac{(3\kappa_1)\vee 1}{d\kappa_3}}\sqrt{\log n}$, and $K \asymp n^{\frac{(3\kappa_1)\vee 1}{d\kappa_3}}\sqrt{\log n}$. Further, if $m \asymp n^{\frac{p(2p+1)(D+6)}{d} + \frac{(24+7D)\kappa_1+4\kappa_2+4((6\kappa_1)\vee 2)}{2d\kappa_3}}\log^2 n\left(\log^{D/2+3} n + \log(1/\xi)\right)$, with probability at least $1-\xi$,

$$\mathbb{W}_p\left(P_{data}, \mathcal{T}_M\left(\hat{P}_{1-\delta}(\hat{v}_{n,m}^{mc})\right)\right) \lesssim n^{-1/d} + n^{-\frac{1}{2p}}\log^{\frac{1}{2p}}(1/\xi). \tag{11}$$

Proof. Throughout, $R \asymp \sqrt{D\log(1/\eta)}$ denotes the truncation radius of Lemma 9, and K denotes the output bound of the approximating network v_0 of that lemma, namely

$$K \lesssim \delta^{-(3\kappa_1)\vee 1} R\sqrt{D} \asymp \delta^{-(3\kappa_1)\vee 1} D\sqrt{\log(1/\eta)}.$$

We abbreviate

$$c_1 := 10\kappa_1 + 2\kappa_2,$$

the exponent appearing in Lemma 8. The ambient dimension D and the diffusion coefficient σ are held fixed, so factors of D and σ are absorbed by $\lesssim$ in the final rate; we nonetheless track those of D explicitly below. By Remark 1, the factor σ^2 carried by Lemma 8 cancels against the σ^{-2} of Lemma 7, so the discretization contribution below is in fact uniform in σ.

Proof for estimator $\hat{v}_n$:
Since conditioning on X_{t_k} eliminates the cross term in $\mathbb{E}\|v(X_{t_k},t_k) - \dot{X}_{t_k}\|^2$ exactly as in the proof of the conditional flow-matching identity (Lemma 11), so $\widehat{\mathcal{L}}_n(v) = \sum_{k=0}^{N-1} h_k \mathbb{E}\|v(X_{t_k},t_k) - v_\star(X_{t_k},t_k)\|^2 + \widehat{\mathcal{L}}_n(v_\star)$, the second term being independent of v. Hence $\hat{v}_n$, the minimizer of $\widehat{\mathcal{L}}_n$ over $\mathcal{V}$, also minimizes the left-hand side below, and

$$\begin{aligned}
&\sum_{k=0}^{N-1} h_k \mathbb{E}\left\|\hat{v}_n(X_{t_k},t_k) - v_\star(X_{t_k},t_k)\right\|^2 \\
&= \inf_{v\in\mathcal{V}} \sum_{k=0}^{N-1} h_k \mathbb{E}\left\|v(X_{t_k},t_k) - v_\star(X_{t_k},t_k)\right\|^2 \\
&\le \sum_{k=0}^{N-1} h_k \mathbb{E}\left\|v_0(X_{t_k},t_k) - v_\star(X_{t_k},t_k)\right\|^2 \\
&= \sum_{k=0}^{N-1} h_k \mathbb{E}\left\|v_0(X_{t_k},t_k) - v_\star(X_{t_k},t_k)\right\|^2 \mathbb{1}\left(\|X_{t_k}\| \le R\right) \\
&\quad + \sum_{k=0}^{N-1} h_k \mathbb{E}\left\|v_0(X_{t_k},t_k) - v_\star(X_{t_k},t_k)\right\|^2 \mathbb{1}\left(\|X_{t_k}\| > R\right) \\
&\lesssim \sum_{k=0}^{N-1} h_k \epsilon^2 + \sum_{k=0}^{N-1} h_k \left(DK^2\eta + \delta^{-2\kappa_1} D^2 \eta^{1/2}\right) \qquad (41) \\
&\lesssim \epsilon^2 + D^2\delta^{-2((3\kappa_1)\vee 1)}\eta^{1/2}. \qquad (42)
\end{aligned}$$

Here (41) follows from Lemmata 9 and 17, bounding the second summand by $2\mathbb{E}\|v_0\|^2\mathbb{1}(\cdot) + 2\mathbb{E}\|v_\star\|^2\mathbb{1}(\cdot)$; the indicator confines X_{t_k} to the Euclidean ball of radius R, which is contained in $[-R,R]^D$, so Lemma 9 applies on the first event. Inequality (42) uses $\sum_k h_k \le 1$ together with $DK^2\eta \lesssim D^2\delta^{-2((3\kappa_1)\vee 1)}\log(1/\eta)\,\eta$ and $\log(1/\eta)\,\eta \le \eta^{1/2}$ for $\eta \le \eta_0$, and $2\kappa_1 \le 2\big((3\kappa_1)\vee 1\big)$.

From Lemmata 7 and 8, we observe that,

$$\begin{aligned}
&\mathrm{KL}\left(P_{1-\delta}\|\hat{P}_{1-\delta}(\hat{v}_n)\right) \\
&\lesssim \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} a_t^2\, \mathbb{E}\|v_\star(X_t,t) - v_\star(X_{t_k},t_k)\|^2 \mathrm{d}t + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} (a_t - a_{t_k})^2 \mathbb{E}\|v_\star(X_{t_k},t_k)\|^2 \mathrm{d}t \\
&\quad + \sum_{k=0}^{N-1} h_k a_{t_k}^2 \mathbb{E}\|v_\star(X_{t_k},t_k) - \hat{v}_n(X_{t_k},t_k)\|^2 \\
&\lesssim D^3\delta^{-c_1}h + D\,\delta^{-2(\kappa_1+\kappa_2)}h^2 + \delta^{-\kappa_1}\left(\epsilon^2 + D^2\delta^{-2((3\kappa_1)\vee 1)}\eta^{1/2}\right) \\
&\lesssim D^3\delta^{-c_1}h + \delta^{-\kappa_1}\left(\epsilon^2 + D^2\delta^{-2((3\kappa_1)\vee 1)}\eta^{1/2}\right),
\end{aligned}$$

where the last step using $h \le 1$ and $2(\kappa_1+\kappa_2) \le c_1$, so that the second summand is absorbed by the first.

Hence, from Lemma 21,

$$\begin{aligned}
&\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}_n))) \\
&\lesssim \mathbb{W}_p(P_{\text{data}}, P_n^{\text{emp}}) + \delta^{\kappa_3} + M^{-\frac{1}{2p}} + M\left(\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v}_n)\right)\right)^{\frac{1}{2p}} \\
&\lesssim \mathbb{W}_p(P_{\text{data}}, P_n^{\text{emp}}) + \delta^{\kappa_3} + M^{-\frac{1}{2p}} \\
&\quad + M\left(D^3\delta^{-c_1}h + \delta^{-\kappa_1}\left(\epsilon^2 + D^2\delta^{-2((3\kappa_1)\vee 1)}\eta^{1/2}\right)\right)^{\frac{1}{2p}} \\
&\le \mathbb{W}_p(P_{\text{data}}, P_n^{\text{emp}}) + \delta^{\kappa_3} + M^{-\frac{1}{2p}} \\
&\quad + M\left(D^{\frac{3}{2p}}\delta^{-\frac{c_1}{2p}}h^{\frac{1}{2p}} + \delta^{-\frac{\kappa_1}{2p}}\epsilon^{1/p} + D^{\frac{1}{p}}\delta^{-\frac{\kappa_1}{2p}-\frac{(3\kappa_1)\vee 1}{p}}\eta^{\frac{1}{4p}}\right),
\end{aligned} \tag{43}$$

by the sub-additivity of $u \mapsto u^{1/2p}$.

We choose $M = n^{\frac{2p}{d}}$, $\delta = n^{-\frac{1}{d\kappa_3}}$, $h \asymp n^{-2p\left(\frac{2p+1}{d}+\frac{c_1}{2dp\kappa_3}\right)} = n^{-\frac{2p(2p+1)}{d}-\frac{10\kappa_1+2\kappa_2}{d\kappa_3}}$, $\epsilon = n^{-p\left(\frac{2p+1}{d}+\frac{\kappa_1}{2dp\kappa_3}\right)}$, and $\eta = n^{-4p\left(\frac{2p+1}{d}+\frac{\kappa_1+2((3\kappa_1)\vee 1)}{2dp\kappa_3}\right)}$. Each of the remaining terms of (43) is then $O(n^{-1/d})$. Hence

$$\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}_n))) \lesssim \mathbb{W}_p(P_{\text{data}}, P_n^{\text{emp}}) + n^{-1/d} \lesssim n^{-1/d} + n^{-\frac{1}{2p}}\log^{\frac{1}{2p}}(1/\xi),$$

with probability at least $1-\xi$, by Lemma 20.

Proof for Estimator $\hat{v}_{n,m}^{\text{mc}}$:
Write ϵ_{c} for the radius of the ℓ_∞-cover of $\mathcal{V}$ appearing in Lemma 14, to distinguish it from the approximation accuracy ϵ of Lemma 9. From Lemma 14 we recall that, with probability at least $1-\beta$,

$$\begin{aligned}
&\sum_{k=0}^{N-1} h_k \mathbb{E}\|\hat{v}_{n,m}^{\text{mc}}(X_{t_k}, t_k) - v_\star(X_{t_k}, t_k)\|^2 \\
&\lesssim \sum_{k=0}^{N-1} h_k \mathbb{E}\|v_0(X_{t_k}, t_k) - v_\star(X_{t_k}, t_k)\|^2 + DK^2\sqrt{\frac{\Lambda_\beta}{m}} + D\,\epsilon_{\text{c}}K,
\end{aligned}$$

where, $\Lambda_\beta = \log\frac{N(1+\mathcal{N}(\epsilon_{\text{c}};\mathcal{V},\ell_\infty))}{\beta}$. We take $m \asymp \epsilon_{\text{c}}^{-2}K^2\,\Lambda_{\xi'}$, which makes the middle term $O(D\,\epsilon_{\text{c}}K)$. This ensures that with probability at least $1-\xi'$,

$$\begin{aligned}
&\sum_{j=0}^{N-1} h_j \mathbb{E}\left\|\hat{v}_{n,m}^{\text{mc}}(X_{t_j}, t_j) - v_\star(X_{t_j}, t_j)\right\|^2 \\
&\lesssim \sum_{j=0}^{N-1} h_j \mathbb{E}\left\|v_0(X_{t_j}, t_j) - v_\star(X_{t_j}, t_j)\right\|^2 + D\,\epsilon_{\text{c}}K \\
&= \sum_{j=0}^{N-1} h_j \mathbb{E}\left\|v_0(X_{t_j}, t_j) - v_\star(X_{t_j}, t_j)\right\|^2 \mathbb{1}\left(\|X_{t_j}\| \le R\right) \\
&\quad + \sum_{j=0}^{N-1} h_j \mathbb{E}\left\|v_0(X_{t_j}, t_j) - v_\star(X_{t_j}, t_j)\right\|^2 \mathbb{1}\left(\|X_{t_j}\| > R\right) + D\,\epsilon_{\text{c}}K \\
&\lesssim \sum_{k=0}^{N-1} h_k\epsilon^2 + \sum_{k=0}^{N-1} h_k\left(DK^2\eta + \delta^{-2\kappa_1}D^2\eta^{1/2}\right) + D\,\epsilon_{\text{c}}K \qquad (44) \\
&\lesssim \epsilon^2 + D\,\epsilon_{\text{c}}K + D^2\delta^{-2((3\kappa_1)\vee 1)}\eta^{1/2}. \qquad (45)
\end{aligned}$$

From Lemmata 7 and 8, exactly as above,

$$\mathrm{KL}\left(P_{1-\delta}\|\hat{P}_{1-\delta}(\hat{v}_{n,m}^{\text{mc}})\right) \lesssim D^3\delta^{-c_1}h + \delta^{-\kappa_1}\left(\epsilon^2 + D\,\epsilon_{\text{c}}K + D^2\delta^{-2((3\kappa_1)\vee 1)}\eta^{1/2}\right).$$

Hence, from Lemma 21,

$$\begin{aligned}
&\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}^{\text{mc}}_{n,m})))\\
&\lesssim \mathbb{W}_p(P_{\text{data}}, P^{\text{emp}}_n) + \delta^{\kappa_3} + M^{-\frac{1}{2p}} + M\left(\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v}^{\text{mc}}_{n,m})\right)\right)^{\frac{1}{2p}}\\
&\le \mathbb{W}_p(P_{\text{data}}, P^{\text{emp}}_n) + \delta^{\kappa_3} + M^{-\frac{1}{2p}}\\
&\quad + M\left(D^{\frac{3}{2p}}\delta^{-\frac{c_1}{2p}}h^{\frac{1}{2p}} + \delta^{-\frac{\kappa_1}{2p}}\epsilon^{\frac{1}{p}} + D^{\frac{1}{2p}}\delta^{-\frac{\kappa_1}{2p}}\epsilon_{\mathrm{c}}^{\frac{1}{2p}}K^{\frac{1}{2p}} + D^{\frac{1}{p}}\delta^{-\frac{\kappa_1}{2p}-\frac{(3\kappa_1)\vee 1}{p}}\eta^{\frac{1}{4p}}\right). \qquad (46)
\end{aligned}$$

Keeping M, δ, h, ϵ and η as above and setting $\epsilon_{\mathrm{c}} = \epsilon^2/K$, the third term of (46) coincides with the second up to the fixed factor $D^{1/2p}$, and every term is $O(n^{-1/d})$ by the computation above. Hence

$$\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}^{\text{mc}}_{n,m}))) \lesssim \mathbb{W}_p(P_{\text{data}}, P^{\text{emp}}_n) + n^{-1/d} \lesssim n^{-1/d} + n^{-\frac{1}{2p}}\log^{\frac{1}{2p}}(1/\xi),$$

with probability at least $1 - \xi - \xi'$, by Lemma 20. Replacing ξ and ξ' with $\xi/2$ gives the desired result.

From Lemma 9, $W \lesssim NR^D\delta^{-3\kappa_1 D}\epsilon^{-D}\big(\log(1/\delta) + \log(1/\epsilon) + \log R\big)$, with $N = h^{-1}$. Substituting the choices above gives

$$N = n^{\frac{2p(2p+1)}{d} + \frac{20\kappa_1 + 4\kappa_2}{2d\kappa_3}}, \qquad \delta^{-3\kappa_1 D} = n^{\frac{3\kappa_1 D}{d\kappa_3}} = n^{\frac{6\kappa_1 D}{2d\kappa_3}}, \qquad \epsilon^{-D} = n^{\frac{Dp(2p+1)}{d} + \frac{D\kappa_1}{2d\kappa_3}}.$$

Since η is polynomial in n, $R \asymp \sqrt{D\log(1/\eta)} \asymp \sqrt{D\log n}$, so $R^D \asymp (D\log n)^{D/2}$ and $\log(1/\delta) + \log(1/\epsilon) + \log R \asymp \log n$. Summing exponents, the n-exponents combine as $\frac{2p(2p+1)}{d} + \frac{Dp(2p+1)}{d} = \frac{p(2p+1)(D+2)}{d}$ and the κ-exponents as $\frac{(20\kappa_1+4\kappa_2)+6\kappa_1 D + D\kappa_1}{2d\kappa_3} = \frac{(20+7D)\kappa_1 + 4\kappa_2}{2d\kappa_3}$, so

$$W \lesssim n^{\frac{p(2p+1)(D+2)}{d} + \frac{(20+7D)\kappa_1 + 4\kappa_2}{2d\kappa_3}}(D\log n)^{D/2}\log n.$$

Further, $L \lesssim \log(1/\delta) + \log(1/\epsilon) + \log R \lesssim \log n$.

From Lemma 9, $B \lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D} + N$, where $\delta^{-(3\kappa_1)\vee 1} = n^{\frac{(3\kappa_1)\vee 1}{d\kappa_3}}$ and N is as above, so $B \lesssim n^{\left(\frac{2p(2p+1)}{d} + \frac{10\kappa_1 + 2\kappa_2}{d\kappa_3}\right)\vee\frac{(3\kappa_1)\vee 1}{d\kappa_3}} D\sqrt{\log n}$, and likewise $K \lesssim \delta^{-(3\kappa_1)\vee 1}R\sqrt{D} \lesssim n^{\frac{(3\kappa_1)\vee 1}{d\kappa_3}} D\sqrt{\log n}$, the latter unchanged.

Finally, by Lemma 23, $\log\mathcal{N}(\epsilon_{\mathrm{c}}; \mathcal{V}, \ell_\infty) \le W\log\big(2LB^L(W+1)^L/\epsilon_{\mathrm{c}}\big) \lesssim W\log^2 n$, since $\log L, \log B, \log W, \log(1/\epsilon_{\mathrm{c}}) \lesssim \log n$ and $L \lesssim \log n$. With $\epsilon_{\mathrm{c}} = \epsilon^2/K$ we have $\epsilon_{\mathrm{c}}^{-2}K^2 = \epsilon^{-4}K^4$, whence

$$\begin{aligned}
m &\asymp \epsilon^{-4}K^4\Big(W\log^2 n + \log(1/\xi)\Big)\\
&\asymp n^{\frac{p(2p+1)(D+6)}{d} + \frac{(24+7D)\kappa_1 + 4\kappa_2 + 4((6\kappa_1)\vee 2)}{2d\kappa_3}}\ \log^2 n\ \Big((D\log n)^{D/2}\log^3 n + \log(1/\xi)\Big),
\end{aligned}$$

using $\epsilon^{-4} = n^{\frac{4p(2p+1)}{d} + \frac{4\kappa_1}{2d\kappa_3}}$ and $K^4 \asymp n^{\frac{8((3\kappa_1)\vee 1)}{2d\kappa_3}} D^4\log^2 n$, together with $8\big((3\kappa_1)\vee 1\big) = 4\big((6\kappa_1)\vee 2\big)$ and $\frac{4p(2p+1)}{d} + \frac{p(2p+1)(D+2)}{d} = \frac{p(2p+1)(D+6)}{d}$. □

C.2 Supporting Lemmata

Lemma 14. *Let $\epsilon_{\mathrm{c}} \in (0,1)$, let $v_0 \in \mathcal{V}$ be the network of Lemma 9, and write K for the output bound of $\mathcal{V} \equiv \mathcal{RN}(W, L, B, K)$ and $\Lambda_\beta := \log\Big(N\big(1 + \mathcal{N}(\epsilon_{\mathrm{c}}; \mathcal{V}, \ell_\infty)\big)/\beta\Big)$. If $m \ge \Lambda_\beta$, then with probability at least $1 - \beta$ over the Monte Carlo sampling,*

$$\sum_{k=0}^{N-1} h_k\mathbb{E}\|\hat{v}^{mc}_{n,m} - v_\star\|^2 \lesssim \sum_{k=0}^{N-1} h_k\mathbb{E}\|v_0 - v_\star\|^2 + DK^2\sqrt{\frac{\Lambda_\beta}{m}} + D\,\epsilon_{\mathrm{c}}K. \qquad (47)$$

Proof. All expectations are conditional on $x_1, \dots, x_n$. Fix $k \in \{0, 1, \dots, N-1\}$ and $v \in \mathcal{V}$. Since the Monte Carlo sample is i.i.d.,

$$\mathbb{E}\left(\frac{1}{m}\sum_{j=1}^{m}\left\|v(X_{t_k}^{(j)}, t_k) - (\dot\alpha_{t_k}X_1^{(j)} + \dot\beta_{t_k}X_0^{(j)})\right\|^2\right) = \mathbb{E}\left\|v(X_{t_k}, t_k) - (\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0)\right\|^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X_0}\|v(X_{i,t_k}, t_k) - (\dot\alpha_{t_k}x_i + \dot\beta_{t_k}X_0)\|^2,$$

where $X_{i,t} = \alpha_t x_i + \beta_t X_0$ and $X_0 \sim P_0$. By Bernstein's inequality for sub-exponential variables (Vershynin, 2026, Theorem 2.9.1), for any fixed k and v, with probability at least $1-\beta$,

$$\left|\frac{1}{m}\sum_{j=1}^{m}\left\|v(X_{t_k}^{(j)}, t_k) - (\dot\alpha_{t_k}X_1^{(j)} + \dot\beta_{t_k}X_0^{(j)})\right\|^2 - \mathbb{E}\|v(X_{t_k}, t_k) - (\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0)\|^2\right|$$
$$\lesssim \left(\left\|\|v(X_{t_k}, t_k)\|^2\right\|_{\psi_1} + \left\|\|\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0\|^2\right\|_{\psi_1}\right)\left(\sqrt{\frac{\log(1/\beta)}{m}} \vee \frac{\log(1/\beta)}{m}\right). \tag{48}$$

Since $\mathcal{K}(v) \le K$ bounds $\|v(\cdot,\cdot)\|_\infty$ coordinatewise, $\|v(X_{t_k}, t_k)\|^2 \le DK^2$ almost surely, so

$$\left\|\|v(X_{t_k}, t_k)\|^2\right\|_{\psi_1} \lesssim DK^2.$$

Moreover, by D1(a), $|\dot\alpha_t| \vee |\dot\beta_t| \lesssim \delta^{-\kappa_1}$ for $t \in [0, 1-\delta]$, so $\left\|\|\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0\|\right\|_{\psi_2} \lesssim \delta^{-\kappa_1}\sqrt{D}$ and hence

$$\left\|\|\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0\|^2\right\|_{\psi_1} = \left\|\|\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0\|\right\|_{\psi_2}^2 \lesssim D\delta^{-2\kappa_1} \lesssim DK^2,$$

the last step because $K \gtrsim \delta^{-(3\kappa_1)\vee 1} \ge \delta^{-\kappa_1}$. Hence from (48), for fixed k and v, with probability at least $1-\beta$,

$$\left|\frac{1}{m}\sum_{j=1}^{m}\left\|v(X_{t_k}^{(j)}, t_k) - (\dot\alpha_{t_k}X_1^{(j)} + \dot\beta_{t_k}X_0^{(j)})\right\|^2 - \mathbb{E}\|v(X_{t_k}, t_k) - (\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0)\|^2\right|$$
$$\lesssim DK^2\sqrt{\frac{\log(1/\beta)}{m}}. \tag{49}$$

Let $\mathcal{C}(\epsilon_c; \mathcal{V}, \ell_\infty)$ be an ϵ_c-cover of $\mathcal{V}$ in the ℓ_∞ norm, of cardinality $\mathcal{N}(\epsilon_c; \mathcal{V}, \ell_\infty)$. Applying (49) with β replaced by $\beta/\big(N(1+\mathcal{N}(\epsilon_c; \mathcal{V}, \ell_\infty))\big)$ and taking a union bound over the N time indices and the $\mathcal{N}(\epsilon_c; \mathcal{V}, \ell_\infty)+1$ networks in $\mathcal{C}(\epsilon_c; \mathcal{V}, \ell_\infty) \cup \{v_0\}$, we obtain: with probability at least $1-\beta$, simultaneously for all $v \in \mathcal{C}(\epsilon_c; \mathcal{V}, \ell_\infty) \cup \{v_0\}$,

$$\sum_{k=0}^{N-1} h_k \left|\frac{1}{m}\sum_{j=1}^{m}\left\|v(X_{t_k}^{(j)}, t_k) - (\dot\alpha_{t_k}X_1^{(j)} + \dot\beta_{t_k}X_0^{(j)})\right\|^2 - \mathbb{E}\|v(X_{t_k}, t_k) - (\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0)\|^2\right|$$
$$\lesssim DK^2\sqrt{\frac{\Lambda_\beta}{m}}, \tag{50}$$

using $\sum_k h_k \le 1$. We work on this event for the remainder of the proof.

Let $\tilde v \in \mathcal{C}(\epsilon_c; \mathcal{V}, \ell_\infty)$ be such that $\|\hat v_{n,m}^{\mathrm{mc}} - \tilde v\|_\infty \le \epsilon_c$. Using $\|a\|^2 - \|b\|^2 = \langle a-b, a+b\rangle$, for any k and any x,

$$\left|\|\hat v_{n,m}^{\mathrm{mc}}(x, t_k) - (\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0)\|^2 - \|\tilde v(x, t_k) - (\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0)\|^2\right|$$
$$\le \sqrt{D}\,\|\hat v_{n,m}^{\mathrm{mc}} - \tilde v\|_\infty\left(\|\hat v_{n,m}^{\mathrm{mc}}\|_\infty + \|\tilde v\|_\infty + 2\|\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0\|\right)$$
$$\le 2\sqrt{D}\,\epsilon_c\left(\sqrt{D}K + \|\dot\alpha_{t_k}X_1 + \dot\beta_{t_k}X_0\|\right). \tag{51}$$

Taking expectations in (51) and using $\mathbb{E}\|\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0\|\lesssim\delta^{-\kappa_1}\sqrt{D}\lesssim\sqrt{D}K$ gives, for every k,

$$\left|\mathbb{E}\|\hat v^{\mathrm{mc}}_{n,m}(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2-\mathbb{E}\|\tilde v(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2\right|\ \lesssim\ D\,\epsilon_{\mathrm{c}}K,$$

and the same bound holds with $\mathbb{E}$ replaced by the empirical average $\frac{1}{m}\sum_{j=1}^m$, since (51) is pointwise in the sample. Chaining these with (50) applied first to $\tilde v$ and then to v_0, and using that $\hat v^{\mathrm{mc}}_{n,m}$ minimises the Monte Carlo objective over $\mathcal{V}\ni v_0$,

$$\begin{aligned}
&\sum_{k=0}^{N-1}h_k\mathbb{E}\|\hat v^{\mathrm{mc}}_{n,m}(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2\\
\le\ &\sum_{k=0}^{N-1}h_k\mathbb{E}\|\tilde v(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2+\mathcal{O}\big(D\epsilon_{\mathrm{c}}K\big)\\
\le\ &\frac{1}{m}\sum_{j=1}^m\sum_{k=0}^{N-1}h_k\|\tilde v(X^{(j)}_{t_k},t_k)-(\dot\alpha_{t_k}X_1^{(j)}+\dot\beta_{t_k}X_0^{(j)})\|^2+\mathcal{O}\left(DK^2\sqrt{\frac{\Lambda_\beta}{m}}\right)+\mathcal{O}\big(D\epsilon_{\mathrm{c}}K\big)\\
\le\ &\frac{1}{m}\sum_{j=1}^m\sum_{k=0}^{N-1}h_k\|\hat v^{\mathrm{mc}}_{n,m}(X^{(j)}_{t_k},t_k)-(\dot\alpha_{t_k}X_1^{(j)}+\dot\beta_{t_k}X_0^{(j)})\|^2+\mathcal{O}\left(DK^2\sqrt{\frac{\Lambda_\beta}{m}}\right)+\mathcal{O}\big(D\epsilon_{\mathrm{c}}K\big)\\
\le\ &\frac{1}{m}\sum_{j=1}^m\sum_{k=0}^{N-1}h_k\|v_0(X^{(j)}_{t_k},t_k)-(\dot\alpha_{t_k}X_1^{(j)}+\dot\beta_{t_k}X_0^{(j)})\|^2+\mathcal{O}\left(DK^2\sqrt{\frac{\Lambda_\beta}{m}}\right)+\mathcal{O}\big(D\epsilon_{\mathrm{c}}K\big)\\
\le\ &\sum_{k=0}^{N-1}h_k\mathbb{E}\|v_0(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2+\mathcal{O}\left(DK^2\sqrt{\frac{\Lambda_\beta}{m}}\right)+\mathcal{O}\big(D\epsilon_{\mathrm{c}}K\big).
\end{aligned}$$

Finally, conditioning on X_{t_k} makes the cross term vanish, so that for every v,

$$\begin{aligned}
&\sum_{k=0}^{N-1}h_k\mathbb{E}\|v(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2\\
=\ &\sum_{k=0}^{N-1}h_k\mathbb{E}\|v(X_{t_k},t_k)-v_\star(X_{t_k},t_k)\|^2+\sum_{k=0}^{N-1}h_k\mathbb{E}\|v_\star(X_{t_k},t_k)-(\dot\alpha_{t_k}X_1+\dot\beta_{t_k}X_0)\|^2,
\end{aligned}$$

this being the discrete analogue of Lemma 11. Applying it to $v=\hat v^{\mathrm{mc}}_{n,m}$ and to $v=v_0$ and subtracting the common second term, which does not depend on v, yields

$$\sum_{k=0}^{N-1}h_k\mathbb{E}\|\hat v^{\mathrm{mc}}_{n,m}-v_\star\|^2\lesssim\sum_{k=0}^{N-1}h_k\mathbb{E}\|v_0-v_\star\|^2+DK^2\sqrt{\frac{\Lambda_\beta}{m}}+D\epsilon_{\mathrm{c}}K.\qquad\square$$

D Supporting Results

D.1 Results for the Velocity Field

Lemma 15 (Spatial Derivative of $v_\star$)**.** *The following holds for any $x\in\mathbb{R}^D$ and $t>0$,*

$$v_\star(x,t)=\frac{\dot\beta_t}{\beta_t}x+\frac{\dot\alpha_t\beta_t-\dot\beta_t\alpha_t}{\beta_t}\mathbb{E}(X_1|X_t=x)=\frac{\dot\alpha_t}{\alpha_t}x+\beta_t^2\left(\frac{\dot\alpha_t}{\alpha_t}-\frac{\dot\beta_t}{\beta_t}\right)\nabla\log\varrho_t(x),\tag{52}$$

$$\nabla v_\star(x,t)=\frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t}-\frac{\dot\beta_t}{\beta_t}\right)\mathrm{Var}(X_1|X_t=x)+\frac{\dot\beta_t}{\beta_t}I_D=\frac{1}{2}\partial_t\left(\frac{\alpha_t^2}{\beta_t^2}\right)\mathrm{Var}(X_1|X_t=x)+\frac{\dot\beta_t}{\beta_t}I_D.\tag{53}$$

Proof. Throughout, fix $t\in(0,1)$ with $\alpha_t\neq0$, $\beta_t>0$, and recall the interpolant is given by, $X_t=\alpha_tX_1+\beta_tX_0$, $X_1\sim P_n^{\mathrm{emp}}$, $X_0\sim N(0,I_D)$ independent of X_1. Hence, ϱ_t, the law of X_t, is the Gaussian smoothing

$$\varrho_t(x)=\int_{\mathbb{R}^D}\phi(x;\alpha_ty,\beta_t^2I_D)\,P_n^{\mathrm{emp}}(dy),$$

where, $\phi(x;\mu,\Sigma) := (2\pi)^{-D/2}\det(\Sigma)^{-1/2}\exp\Big(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\Big)$. By definition, $v_\star(x,t) = \mathbb{E}[\dot{X}_t \mid X_t = x]$. For notational simplicity, we write

$$m(x,t) := \mathbb{E}[X_1 \mid X_t = x] \in \mathbb{R}^D, \qquad \Sigma(x,t) := \mathrm{Var}(X_1 \mid X_t = x) \in \mathbb{R}^{D\times D}.$$

Since $\beta_t > 0$, we may write $X_0 = \beta_t^{-1}(X_t - \alpha_t X_1)$. Hence

$$v_\star(x,t) = \mathbb{E}[\dot{\alpha}_t X_1 + \dot{\beta}_t X_0 \mid X_t = x] = \dot{\alpha}_t\, m(x,t) + \frac{\dot{\beta}_t}{\beta_t}\big(x - \alpha_t m(x,t)\big),$$

using $\mathbb{E}[X_t \mid X_t = x] = x$. Collecting terms in $m(x,t)$ gives

$$v_\star(x,t) = \frac{\dot{\beta}_t}{\beta_t}\, x + \frac{\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t}{\beta_t}\, m(x,t). \tag{54}$$

To prove equation (53), we note that, $\nabla_x\phi(x;\alpha_t y,\beta_t^2 I_D) = -\beta_t^{-2}(x-\alpha_t y)\,\phi(x;\alpha_t y,\beta_t^2 I_D)$, differentiation under the integral sign (justified since $\beta_t > 0$ and P_n^{emp} has finite second moments) gives

$$\nabla\varrho_t(x) = -\frac{1}{\beta_t^2}\int (x-\alpha_t y)\,\phi(x;\alpha_t y,\beta_t^2 I_D)\, P_n^{\text{emp}}(dy).$$

Dividing by $\varrho_t(x) > 0$ and using that $\phi(x;\alpha_t y,\beta_t^2 I_D)P_n^{\text{emp}}(dy)/\varrho_t(x)$ is exactly the conditional law of X_1 given $X_t = x$,

$$\nabla\log\varrho_t(x) = -\frac{1}{\beta_t^2}\,\mathbb{E}[x - \alpha_t X_1 \mid X_t = x] = -\frac{x - \alpha_t m(x,t)}{\beta_t^2}. \tag{55}$$

Solving for $m(x,t)$ (using $\alpha_t \neq 0$):

$$m(x,t) = \frac{x + \beta_t^2\nabla\log\varrho_t(x)}{\alpha_t}. \tag{56}$$

Substituting (56) into (52),

$$v_\star(x,t) = \frac{\dot{\beta}_t}{\beta_t}x + \frac{\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t}{\alpha_t\beta_t}\, x + \frac{(\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t)\beta_t}{\alpha_t}\,\nabla\log\varrho_t(x).$$

The coefficient of x simplifies as $\frac{\dot{\beta}_t}{\beta_t} + \frac{\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t}{\alpha_t\beta_t} = \frac{\dot{\beta}_t\alpha_t + \dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t}{\alpha_t\beta_t} = \frac{\dot{\alpha}_t}{\alpha_t}$, and the coefficient of $\nabla\log\varrho_t(x)$ as

$$\frac{\beta_t(\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t)}{\alpha_t} = \beta_t^2\cdot\frac{\dot{\alpha}_t\beta_t - \dot{\beta}_t\alpha_t}{\alpha_t\beta_t} = \beta_t^2\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right).$$

Hence

$$v_\star(x,t) = \frac{\dot{\alpha}_t}{\alpha_t}x + \beta_t^2\left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\dot{\beta}_t}{\beta_t}\right)\nabla\log\varrho_t(x), \tag{57}$$

which together with (52) establishes the first display of the lemma.

To complete the proof, we make the following claim,

$$\nabla_x m(x,t) = \frac{\alpha_t}{\beta_t^2}\,\Sigma(x,t). \tag{58}$$

Write $N_i(x) := \int y_i\,\phi(x;\alpha_t y,\beta_t^2 I_D)\,P_n^{\text{emp}}(dy)$, so that $m_i(x,t) = N_i(x)/\varrho_t(x)$ for $i = 1,\dots,D$. Differentiating under the integral sign as before,

$$\begin{aligned}\partial_{x_j}N_i(x) &= -\frac{1}{\beta_t^2}\int y_i(x_j - \alpha_t y_j)\,\phi(x;\alpha_t y,\beta_t^2 I_D)\,P_n^{\text{emp}}(dy)\\ &= -\frac{x_j}{\beta_t^2}N_i(x) + \frac{\alpha_t}{\beta_t^2}\,\varrho_t(x)\,\mathbb{E}[X_{1,i}X_{1,j} \mid X_t = x].\end{aligned}$$

Dividing by $\varrho_t(x)$, $\frac{\partial_{x_j} N_i(x)}{\varrho_t(x)} = -\frac{x_j}{\beta_t^2}\, m_i(x,t) + \frac{\alpha_t}{\beta_t^2}\, \mathbb{E}[X_{1,i} X_{1,j} \mid X_t = x]$. By the quotient rule, $\partial_{x_j} m_i(x,t) = \dfrac{\partial_{x_j} N_i(x)}{\varrho_t(x)} - m_i(x,t)\, \partial_{x_j} \log \varrho_t(x)$. Substituting the previous display and (2) (i.e. $\partial_{x_j} \log \varrho_t(x) = -\beta_t^{-2}(x_j - \alpha_t m_j(x,t))$),

$$\partial_{x_j} m_i(x,t) = \frac{\alpha_t}{\beta_t^2}\Big(\mathbb{E}[X_{1,i} X_{1,j} \mid X_t = x] - m_i(x,t) m_j(x,t)\Big) = \frac{\alpha_t}{\beta_t^2}\, \Sigma_{ij}(x,t),$$

which is (58) in coordinates.

Differentiate (52) with respect to x and apply (58):

$$\nabla v_\star(x,t) = \frac{\dot\beta_t}{\beta_t} I_D + \frac{\dot\alpha_t \beta_t - \dot\beta_t \alpha_t}{\beta_t}\, \nabla_x m(x,t) = \frac{\dot\beta_t}{\beta_t} I_D + \frac{\alpha_t(\dot\alpha_t \beta_t - \dot\beta_t \alpha_t)}{\beta_t^3}\, \Sigma(x,t).$$

Since $\frac{\alpha_t(\dot\alpha_t\beta_t - \dot\beta_t\alpha_t)}{\beta_t^3} = \frac{\alpha_t^2}{\beta_t^2} \cdot \frac{\dot\alpha_t\beta_t - \dot\beta_t\alpha_t}{\alpha_t\beta_t} = \frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)$, we obtain

$$\nabla v_\star(x,t) = \frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right) \mathrm{Var}(X_1 \mid X_t = x) + \frac{\dot\beta_t}{\beta_t} I_D. \tag{59}$$

By the quotient rule, $\partial_t\left(\frac{\alpha_t^2}{\beta_t^2}\right) = \frac{2\alpha_t\dot\alpha_t\beta_t^2 - 2\alpha_t^2\beta_t\dot\beta_t}{\beta_t^4} = \frac{2\alpha_t\beta_t(\dot\alpha_t\beta_t - \alpha_t\dot\beta_t)}{\beta_t^4} = 2\,\frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)$. Hence $\frac{1}{2}\,\partial_t\left(\frac{\alpha_t^2}{\beta_t^2}\right) = \frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)$, and substituting into (59) gives

$$\nabla v_\star(x,t) = \frac{1}{2}\,\partial_t\left(\frac{\alpha_t^2}{\beta_t^2}\right) \mathrm{Var}(X_1 \mid X_t = x) + \frac{\dot\beta_t}{\beta_t} I_D.$$

□

Corollary 16. *For any $t \in [0, 1-\delta]$,*

$$\|v_\star(x,t)\| \lesssim \delta^{-\kappa_1}\|x\| + \delta^{-\kappa_1}\sqrt{D}, \tag{60}$$
$$\|\nabla v_\star(x,t)\| \lesssim D\delta^{-3\kappa_1}. \tag{61}$$

Proof. By construction, for any $t \in [0, 1-\delta]$,

$$\begin{aligned}\|v_\star(x,t)\| &= \left\| \frac{\dot\beta_t}{\beta_t} x + \frac{\dot\alpha_t\beta_t - \dot\beta_t\alpha_t}{\beta_t} \mathbb{E}(X_1 | X_t = x) \right\| \\ &\lesssim \delta^{-\kappa_1}\|x\| + \delta^{-\kappa_1}\sqrt{D}.\end{aligned}$$

Similarly,

$$\begin{aligned}\|\nabla v_\star(x,t)\| &= \left\| \frac{\alpha_t^2}{\beta_t^2}\left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right) \mathrm{Var}(X_1 | X_t = x) + \frac{\dot\beta_t}{\beta_t} I_D \right\| \\ &\lesssim \frac{D}{\beta_t^2}\left|\frac{\dot\alpha_t\beta_t - \dot\beta_t\alpha_t}{\beta_t}\right| + \left|\frac{\dot\beta_t}{\beta_t}\right| \\ &\lesssim D\delta^{-3\kappa_1}.\end{aligned}$$

□

Lemma 17. *There exists an absolute constant C such that, for every $\eta \le \eta_0$, writing $R = C\sqrt{D\log(1/\eta)}$, we have for every $t \in [0, 1-\delta]$*

$$\mathbb{P}\big(\|X_t\| > R\big) \;\le\; \eta \qquad \textit{and} \qquad \mathbb{E}\|v_\star(X_t,t)\|^2 \mathbb{1}\big(\|X_t\| > R\big) \;\lesssim\; \delta^{-2\kappa_1} D^2 \eta^{1/2}.$$

Proof. Fix $t \in [0, 1-\delta]$. By Theorem 13, $\mathrm{Law}(X_t) = \varrho_t = \mathrm{Law}(\alpha_t X_1 + \beta_t X_0)$, a statement about the one-time marginal, which is all that is used below. Since $X_1 \sim P_n^{\mathrm{emp}}$ is supported in $[0,1]^D$ by Assumption 1 we have $\|X_1\| \le \sqrt{D}$, and since $\alpha_t, \beta_t \in [0,1]$,

$$\|X_t\| \;\le\; \alpha_t\|X_1\| + \beta_t\|X_0\| \;\le\; \sqrt{D} + \|X_0\|, \qquad X_0 \sim \mathcal{N}(0, I_D). \tag{62}$$

Hence $\big\| \|X_t\| \big\|_{\psi_2} \le \sqrt{D} + \big\| \|X_0\| \big\|_{\psi_2} \lesssim \sqrt{D}$, and by Proposition 2.6.1 of Vershynin (2026), $\mathbb{E}\|X_t\|^4 \lesssim D^2$. Moreover, by (62) and Gaussian concentration of the norm, $\mathbb{P}\big(\|X_t\| \ge 2\sqrt{D}+s\big) \le \mathbb{P}\big(\|X_0\| \ge \sqrt{D}+s\big) \le e^{-s^2/2}$ for every $s \ge 0$; taking $s = \sqrt{2\log(1/\eta)}$ gives the first claim, for a suitable absolute constant C and $\eta \le \eta_0$.

For the second claim, by Corollary 16 and then the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\|v_\star(X_t,t)\|^2 \mathbb{1}\big(\|X_t\| > R\big) \\
&\lesssim \delta^{-2\kappa_1}\, \mathbb{E}\big(\|X_t\|^2 + D\big)\, \mathbb{1}\big(\|X_t\| > R\big) \qquad (63)\\
&\le \delta^{-2\kappa_1} \Big(\mathbb{E}\big(\|X_t\|^2 + D\big)^2\Big)^{1/2} \Big(\mathbb{P}\big(\|X_t\| > R\big)\Big)^{1/2} \\
&\lesssim \delta^{-2\kappa_1} \big(\mathbb{E}\|X_t\|^4 + D^2\big)^{1/2}\, \eta^{1/2} \\
&\lesssim \delta^{-2\kappa_1} D\, \eta^{1/2} \;\le\; \delta^{-2\kappa_1} D^2\, \eta^{1/2}.
\end{aligned}$$

In the above calculations, (63) follows from Corollary 16. □

Lemma 18 (Temporal velocity of $v_\star$)**.** *Fix $t \in (0,1)$ with $\alpha_t \neq 0$, $\beta_t > 0$, and suppose α_t, β_t are twice continuously differentiable in a neighborhood of t. Define*

$$u_t := \frac{\dot\alpha_t}{\alpha_t}, \qquad w_t := \frac{\dot\beta_t}{\beta_t}, \qquad \kappa_t := u_t - w_t = \frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}.$$

Then, with $E_{1,t}(x) = \mathbb{E}(X_1 \mid X_t = x)$, $E_{2,t}(x) = \mathbb{E}(\|X_1\|^2 \mid X_t = x)$, $E_{3,t}(x) = \mathbb{E}(\|X_1\|^2 X_1 \mid X_t = x)$, and $V_{1,t}(x) = \mathrm{Var}(X_1 \mid X_t = x)$ as in Lemma 15,

$$\begin{aligned}
\partial_t v_\star(x,t) = \dot w_t\, x \;-\; \frac{\alpha_t^2 \kappa_t (2w_t - u_t)}{\beta_t^2}\, V_{1,t}(x)\, x \;+\; \alpha_t\big(\dot u_t - \dot w_t + u_t\kappa_t\big)\, E_{1,t}(x) \\
- \frac{\alpha_t^3 \kappa_t^2}{\beta_t^2}\big(E_{3,t}(x) - E_{1,t}(x) E_{2,t}(x)\big).
\end{aligned} \tag{64}$$

Further, for any $t \in [0, 1-\delta]$ and $x \in \mathbb{R}^D$, $\|\partial_t v_\star(x,t)\| \lesssim \delta^{-\kappa_1}(1 + \|x\|)$.

Proof. Throughout, write E_1, E_2, E_3, V_1 for $E_{1,t}(x), E_{2,t}(x), E_{3,t}(x), V_{1,t}(x)$, and $g(x) := \nabla_x \log \varrho_t(x)$.

From Lemma 19, taking $f(y) = y$ recovers $\nabla_x E_1 = \frac{\alpha_t}{\beta_t^2} V_1$; taking $f(y) = \|y\|^2$ gives $\nabla_x E_2 = \frac{\alpha_t}{\beta_t^2}(E_3 - E_2 E_1)$. Since $\mathrm{tr}(V_1) = E_2 - \|E_1\|^2$,

$$\nabla_x \,\mathrm{tr}\,\big(V_{1,t}(x)\big) = \nabla_x E_2 - 2(\nabla_x E_1)^\top E_1 = \frac{\alpha_t}{\beta_t^2}\big(E_3 - E_2 E_1 - 2V_1 E_1\big). \tag{65}$$

The continuity equation $\partial_t \varrho_t + \nabla_x \cdot (\varrho_t v_\star) = 0$ gives, for fixed x,

$$\partial_t g(x) = -\Big[\nabla_x^2 \log \varrho_t(x)\, v_\star(x,t) + \nabla_x v_\star(x,t)\, g(x) + \nabla_x\big(\nabla_x \cdot v_\star(x,t)\big)\Big]. \tag{66}$$

Differentiating $g(x) = -\frac{x - \alpha_t E_1}{\beta_t^2}$ in x and using $\nabla_x E_1 = \frac{\alpha_t}{\beta_t^2} V_1$,

$$\nabla_x^2 \log \varrho_t(x) = \frac{\alpha_t^2}{\beta_t^4} V_1 - \frac{1}{\beta_t^2} I_D =: H. \tag{67}$$

Using (57), (53), and (67),

$$H\, v_\star(x,t) = \left(\frac{\alpha_t^2}{\beta_t^4} V_1 - \frac{1}{\beta_t^2} I_D\right)\big(u_t x + q_t g(x)\big) = \frac{\alpha_t^2 u_t}{\beta_t^4} V_1 x + \frac{\alpha_t^2 q_t}{\beta_t^4} V_1 g(x) - \frac{u_t}{\beta_t^2} x - \frac{q_t}{\beta_t^2} g(x),$$

$$\nabla_x v_\star(x,t)\, g(x) = \left(\frac{\alpha_t^2}{\beta_t^2}\kappa_t V_1 + w_t I_D\right) g(x) = \frac{\alpha_t^2\kappa_t}{\beta_t^2} V_1 g(x) + w_t g(x).$$

Since $q_t = \beta_t^2\kappa_t$, the coefficient of $V_1 g(x)$ in the sum is $\frac{\alpha_t^2 q_t}{\beta_t^4} + \frac{\alpha_t^2\kappa_t}{\beta_t^2} = 2\frac{\alpha_t^2\kappa_t}{\beta_t^2}$, and the coefficient of $g(x)$ is $w_t - \frac{q_t}{\beta_t^2} = w_t - \kappa_t = 2w_t - u_t$. Substituting $g(x) = -\frac{x-\alpha_t E_1}{\beta_t^2}$ throughout and collecting terms by type ($V_1 x$, x, $V_1 E_1$, E_1),

$$Hv_\star(x,t) + \nabla_x v_\star(x,t) g(x) = \frac{\alpha_t^2(2w_t - u_t)}{\beta_t^4} V_1 x - \frac{2w_t}{\beta_t^2} x + \frac{2\alpha_t^3\kappa_t}{\beta_t^4} V_1 E_1 + \frac{\alpha_t(2w_t - u_t)}{\beta_t^2} E_1. \tag{68}$$

From (53), $\nabla_x \cdot v_\star(x,t) = \frac{\alpha_t^2}{\beta_t^2}\kappa_t \operatorname{tr}(V_1) + D w_t$, so by (65),

$$\nabla_x\big(\nabla_x \cdot v_\star(x,t)\big) = \frac{\alpha_t^3\kappa_t}{\beta_t^4}\big(E_3 - E_2 E_1 - 2V_1 E_1\big). \tag{69}$$

Adding (69) to (68), the $V_1 E_1$ contributions cancel exactly ($\frac{2\alpha_t^3\kappa_t}{\beta_t^4} - \frac{2\alpha_t^3\kappa_t}{\beta_t^4} = 0$), leaving

$$\partial_t g(x) = -\frac{\alpha_t^2(2w_t - u_t)}{\beta_t^4} V_1 x + \frac{2w_t}{\beta_t^2} x - \frac{\alpha_t(2w_t - u_t)}{\beta_t^2} E_1 - \frac{\alpha_t^3\kappa_t}{\beta_t^4}(E_3 - E_2 E_1).$$

By the product rule,

$$\partial_t v_\star(x,t) = \dot{u}_t\, x + \dot{q}_t\, g(x) + q_t\, \partial_t g(x), \qquad \dot{q}_t = \beta_t^2\big(2w_t\kappa_t + \dot{\kappa}_t\big), \quad \dot{\kappa}_t = \dot{u}_t - \dot{w}_t.$$

Substituting $g(x) = -\frac{x-\alpha_t E_1}{\beta_t^2}$ and $q_t \partial_t g(x)$, then collecting terms of each type (x; $V_1 x$; E_1; $E_3 - E_2 E_1$), we get the final result. Finally,

$$\begin{aligned}\|\partial_t v_\star(x,t)\| \le& |\dot{w}_t|\|x\| + \left|\frac{\alpha_t^2\kappa_t(2w_t - u_t)}{\beta_t^2}\right| \|V_{1,t}(x)\|\,\|x\| + \big|\alpha_t\big(\dot{u}_t - \dot{w}_t + u_t\kappa_t\big)\big|\,\|E_{1,t}(x)\| \\ &+ \left|\frac{\alpha_t^3\kappa_t^2}{\beta_t^2}\right| \|E_{3,t}(x) - E_{1,t}(x)E_{2,t}(x)\| \\ \lesssim& \delta^{-\kappa_1}(1 + \|x\|).\end{aligned}$$

□

Lemma 19. *Fix $t \in (0,1)$ with $\alpha_t \neq 0$, $\beta_t > 0$. For any measurable $f : \mathbb{R}^D \to \mathbb{R}$ with $\mathbb{E}|f(X_1)| < \infty$, define $E_f(x,t) := \mathbb{E}\big(f(X_1) \mid X_t = x\big)$. Then, for every $x \in \mathbb{R}^D$,*

$$\nabla_x E_f(x,t) = \frac{\alpha_t}{\beta_t^2}\, \operatorname{Cov}\big(f(X_1), X_1 \mid X_t = x\big).$$

Proof. Write $N_f(x) := \varrho_t(x)\, E_f(x,t) = \int f(y)\, \phi(x; \alpha_t y, \beta_t^2 I_D)\, P_n^{\text{emp}}(\mathrm{d}y)$. Differentiating under the integral sign,

$$\begin{aligned}\nabla_x N_f(x) &= \int f(y) \left(-\frac{x - \alpha_t y}{\beta_t^2}\right) \phi(x; \alpha_t y, \beta_t^2 I_D)\, P_n^{\text{emp}}(\mathrm{d}y) \\ &= -\frac{x}{\beta_t^2} N_f(x) + \frac{\alpha_t}{\beta_t^2} \int f(y)\, y\, \phi(x; \alpha_t y, \beta_t^2 I_D)\, P_n^{\text{emp}}(\mathrm{d}y).\end{aligned}$$

Dividing by $\varrho_t(x)$ and applying the quotient rule together with $g(x) := \nabla_x \log \varrho_t(x) = -\frac{x - \alpha_t E_1(x,t)}{\beta_t^2}$ (Lemma 15),

$$\begin{aligned}\nabla_x E_f(x,t) &= \frac{\nabla_x N_f(x)}{\varrho_t(x)} - E_f(x,t)\, g(x) \\ &= -\frac{x}{\beta_t^2} E_f(x,t) + \frac{\alpha_t}{\beta_t^2}\mathbb{E}(f(X_1)X_1 \mid X_t = x) + \frac{x - \alpha_t E_1(x,t)}{\beta_t^2} E_f(x,t) \\ &= \frac{\alpha_t}{\beta_t^2}\Big(\mathbb{E}(f(X_1)X_1 \mid X_t = x) - E_f(x,t)\, E_1(x,t)\Big) \\ &= \frac{\alpha_t}{\beta_t^2}\operatorname{Cov}\big(f(X_1), X_1 \mid X_t = x\big).\end{aligned} \tag{70}$$

□

D.2 Results for the Wasserstein Distance

Lemma 20. *For any $d > d_p^*(P_{data})$, with probability at least $1-\xi$,*

$$\mathbb{W}_p(P_n^{emp}, P_{data}) \le n^{-1/d} + n^{-\frac{1}{2p}} D^{1/2} \log^{\frac{1}{2p}}(1/\xi).$$

Proof. Let $f_c(y) = \sup_{x\in[0,1]^D}(f(x) - \|x-y\|^p)$ denote the c-transform of f. It can be shown that $f(x) = \inf_{y\in[0,1]^D}(f_c(y) + \|x-y\|^p)$. From Kantorovich duality, it is well known that, for any $p \ge 1$,

$$\mathbb{W}_p^p(\mu,\nu) = \sup_{f\in\mathcal{C}_b([0,1]^D)} \left(\int f \mathrm{d}\mu - \int f_c \mathrm{d}\nu\right),$$

where $\mathcal{C}_b$ denotes the set of all bounded continuous functions on $[0,1]^D$. By noting that $(f+a)_c = f_c + a$ for any constant a, we can restrict this function class by shifting f such that $\inf_{x\in[0,1]^D} f(x) = 0$. Since f satisfies $f(x) - f(y) \le \|x-y\|^p \le D^{p/2}$ under the optimal transform, this shift guarantees $0 \le f(x) \le D^{p/2}$. Consequently, for any $y \in [0,1]^D$, we have

$$f_c(y) = \sup_{x\in[0,1]^D}(f(x) - \|x-y\|^p) \ge f(y) - \|y-y\|^p = f(y) \ge 0.$$

Hence, we can take the supremum on the set of all bounded continuous functions from $[0,1]^D \to [0, D^{p/2}]$. Thus,

$$\mathbb{W}_p^p(\mu,\nu) = \sup_{f\in\mathcal{C}_b([0,1]^D)\cap\{0\le f\le D^{p/2}\}} \left(\int f \mathrm{d}\mu - \int f_c \mathrm{d}\nu\right).$$

Let

$$\begin{aligned} W(x_1,\dots,x_n) =& \mathbb{W}_p^p(P_{\text{data}}, P_n^{\text{emp}}) \\ =& \sup_{f\in\mathcal{C}_b([0,1]^D)\cap\{0\le f\le D^{p/2}\}} \left(\frac{1}{n}\sum_{i=1}^n f(x_i) - \int f_c \mathrm{d}P_{\text{data}}\right). \end{aligned}$$

Hence,

$$\begin{aligned} & W(x_1,\dots,x_i,\dots,x_n) - W(x_1,\dots,x_i',\dots,x_n) \\ \le & \sup_{f\in\mathcal{C}_b([0,1]^D)\cap\{0\le f\le D^{p/2}\}} \left(\frac{1}{n}\sum_{j=1}^n f(x_j) - \int f_c \mathrm{d}P_{\text{data}}\right) \\ & - \sup_{f\in\mathcal{C}_b([0,1]^D)\cap\{0\le f\le D^{p/2}\}} \left(\frac{1}{n}\sum_{j\ne i} f(x_j) + \frac{f(x_i')}{n} - \int f_c \mathrm{d}P_{\text{data}}\right) \\ \le & \frac{1}{n}\sup_{0\le f\le D^{p/2}} (f(x_i) - f(x_i')) \\ \le & \frac{D^{p/2}}{n}. \end{aligned}$$

Applying the bounded difference inequality, we get that with probability at least $1-\xi$,

$$\begin{aligned} \mathbb{W}_p^p(P_n^{\text{emp}}, P_{\text{data}}) &\le \mathbb{E}\mathbb{W}_p^p(P_n^{\text{emp}}, P_{\text{data}}) + \sqrt{\frac{D^p \log(1/\xi)}{2n}} \\ &\le n^{-p/d} + \sqrt{\frac{D^p \log(1/\xi)}{2n}}, \end{aligned}$$

where the last inequality follows from Chakraborty et al. (2026, Theorem 10). The desired inequality now follows from taking the p-th root of both sides and applying the subadditivity inequality $(a+b)^{1/p} \le a^{1/p} + b^{1/p}$ (valid for $a, b \ge 0$ and $p \ge 1$). □

Lemma 21. *For any $M > 0$ and any estimator $\hat{v}$ of $v_\star$,*

$$\mathbb{W}_p(P_{data}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}))) \lesssim \mathbb{W}_p(P_{data}, P_n^{emp}) + \delta^{\kappa_3} + M^{-\frac{1}{2p}} + M\left(\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\right)\right)^{\frac{1}{2p}}.$$

Proof.

$$\begin{aligned}
&\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}))) \\
\leq &\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(P_{1-\delta})) + \mathbb{W}_p(\mathcal{T}_M(P_{1-\delta}), \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}))) \\
\leq &\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(P_{1-\delta})) + 2M\left(\mathrm{TV}\left(\mathcal{T}_M(P_{1-\delta}), \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}))\right)\right)^{1/p} \\
\leq &\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(P_{1-\delta})) + 2M\left(\mathrm{TV}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\right)\right)^{1/p} \\
\lesssim &\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(P_{1-\delta})) + M\left(\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\right)\right)^{\frac{1}{2p}} \\
\leq &\mathbb{W}_p(P_{\text{data}}, P_n^{\text{emp}}) + \mathbb{W}_p(P_n^{\text{emp}}, P_{1-\delta}) + \mathbb{W}_p(P_{1-\delta}, \mathcal{T}_M(P_{1-\delta})) + M\left(\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\right)\right)^{\frac{1}{2p}}
\end{aligned}$$

$$\begin{aligned}
\mathbb{W}_p^p(P_n^{\text{emp}}, P_{1-\delta}) \leq &\mathbb{E}_{X_1 \sim P_n^{\text{emp}}, X_0 \sim \mathcal{N}(0, I_D)} \|X_1 - (\alpha_{1-\delta} X_1 + \beta_{1-\delta} X_0)\|^p \\
\lesssim &(1 - \alpha_{1-\delta})^p \mathbb{E}_{X_1 \sim P_n^{\text{emp}}} \|X_1\|^p + \beta_{1-\delta}^p \mathbb{E}_{X_1 \sim P_n^{\text{emp}}, X_0 \sim \mathcal{N}(0, I_D)} \|X_0\|^p \\
\lesssim &\delta^{p\kappa_3}.
\end{aligned}$$

Taking the p-th root, we get, $\mathbb{W}_p(P_n^{\text{emp}}, P_{1-\delta}) \lesssim \delta^{\kappa_3}$. Further,

$$\begin{aligned}
\mathbb{W}_p^p(P_{1-\delta}, \mathcal{T}_M(P_{1-\delta})) \leq &\mathbb{E}\|X_{1-\delta}\|^p \mathbb{1}\{\|X_{1-\delta}\|_2 \geq M\} \\
\leq &\sqrt{\mathbb{E}\|X_{1-\delta}\|^{2p}} \sqrt{\mathbb{P}\left(\|X_{1-\delta}\|_2 \geq M\right)} \\
\lesssim &\sqrt{\frac{\mathbb{E}\|X_{1-\delta}\|_2}{M}} \\
\lesssim &M^{-1/2},
\end{aligned}$$

so that, taking p-th roots, $\mathbb{W}_p(P_{1-\delta}, \mathcal{T}_M(P_{1-\delta})) \lesssim M^{-1/(2p)}$. Hence,

$$\mathbb{W}_p(P_{\text{data}}, \mathcal{T}_M(\hat{P}_{1-\delta}(\hat{v}))) \lesssim \mathbb{W}_p(P_{\text{data}}, P_n^{\text{emp}}) + \delta^{\kappa_3} + M^{-\frac{1}{2p}} + M\left(\mathrm{KL}\left(P_{1-\delta}, \hat{P}_{1-\delta}(\hat{v})\right)\right)^{\frac{1}{2p}}.$$

□

E Supporting Results from the Literature

Lemma 22. *For any $\varepsilon \in (0,1)$, there exists a ReLU network architecture $\mathcal{RN}(W, L, B, K)$, such that for any target 1-Lipschitz function f defined on $[0,1]^D$ with $f(0) = 0$, the architecture yields an approximation $\hat{f}$ satisfying $\|f - \hat{f}\|_\infty \leq \varepsilon$.*

Moreover, the Lipschitz continuity of $\hat{f}$ is bounded by

$$\|\hat{f}(x) - \hat{f}(y)\| \leq 10D \|x - y\|_1, \qquad \forall\, x, y \in [0,1]^D.$$

The configuration of the network architecture is $K = \sqrt{D}, B = O(1), L = O(\log(1/\varepsilon) + D)$, and $W = O\big(\varepsilon^{-D}(\log(1/\varepsilon) + D)\big)$.

Lemma 23 (Lemma 21 of Nakada & Imaizumi (2020))**.** *Let $\mathcal{F} = \mathcal{RN}(W, L, B)$ be a space of ReLU networks with the number of weights, the number of layers, and the maximum absolute value of weights bounded by W, L, and B, respectively. Then,*

$$\log \mathcal{N}(\epsilon; \mathcal{F}, \ell_\infty) \leq W \log\left(2LB^L(W+1)^L \frac{1}{\epsilon}\right).$$