FLOW MATCHING UNDER NOISY LATENT STRUCTURE: BEYOND EXACT LOW-DIMENSIONAL SUPPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow Matching (FM) learns a velocity field whose ODE transports a simple source distribution to a target law. Existing finite-sample theory largely treats ambient-space regularity or data supported exactly on low-dimensional sets. We study linear FM under a noisy latent-generator model, where a low-dimensional Hölder map is perturbed by nondegenerate ambient Gaussian noise, so the target law is full-dimensional despite its latent structure. We construct a spatially regular ReLU velocity class and establish non-asymptotic high-probability approximation and estimation bounds whose leading sample-size exponent is governed by the latent dimension rather than the ambient dimension, with ambient and noise dependence kept explicit. Fixed positive target noise keeps the interpolation nondegenerate over the full time interval. The same spatial regularity propagates the learned velocity error through the transport ODE, yielding a corresponding Wasserstein convergence guarantee. These results show that exact low-dimensional support is not necessary for Flow Matching to retain latent-dimensional statistical behavior.

1 INTRODUCTION

Flow Matching (FM) has emerged as a simulation-free framework for continuous-time generative modeling, learning a time-dependent velocity field through regression and generating samples by integrating the resulting ordinary differential equation (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023). Its simple training objective and deterministic sampling dynamics have led to successful applications in high-dimensional domains including image, audio, video, and molecular generation (Esser et al., 2024; Le et al., 2023; Bose et al., 2024). A central theoretical question is therefore how accurately the population velocity can be learned from finitely many observations and how this statistical error propagates to the generated distribution.

Recent work has established increasingly complete statistical guarantees for FM. Stability analyses relate velocity-field error to distributional error of the induced flow (Benton et al., 2024; Gao et al., 2024), while finite-sample analyses derive neural velocity-estimation and Wasserstein convergence rates under regularity assumptions on full-dimensional target distributions or ambient velocity fields (Zhou & Liu, 2025; Fukumizu et al., 2025; Kunkel, 2026). In these formulations, the statistical complexity is primarily described through regularity in the observed state space.

A complementary line of work asks whether generative models can exploit lower-dimensional structure. For diffusion models, intrinsic-dimensional guarantees have been established under linear-subspace, singular-support, and smooth-manifold assumptions (Chen et al., 2023; De Bortoli, 2022; Tang & Yang, 2024; Azangulov et al., 2024). Recent FM theory similarly establishes intrinsic-dimensional behavior for targets supported on low-dimensional subspaces or smooth manifolds (Kunkel & Trabs, 2025; Kumar et al., 2026; Pi et al., 2026; Roy et al., 2026). Thus, the question is no longer whether FM can exploit low-dimensional structure, but how far the exact-support assumptions underlying these guarantees can be relaxed.

Low-dimensional structure need not imply that the observed distribution is itself supported on a low-dimensional set. Motivated by noisy-manifold and latent-variable models (Genovese et al., 2012; Fefferman et al., 2018; Tipping & Bishop, 1999; Lawrence, 2005), we consider

$$X_1 = g^*(U) + \sigma\xi, \qquad U \sim \mathrm{Unif}([0,1]^d), \qquad \xi \sim \mathcal{N}(0, I_D), \qquad d \ll D,$$

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.

where g^* is Hölder regular and $0 < \sigma < 1$ is fixed. The target law is full-dimensional, while its systematic variation is parameterized by a d-dimensional latent variable; the image of g^* need not form a regular embedded manifold. Under the same noisy latent-generator model, Yakovlev & Puchkin (2025) established latent-dimensional approximation and generalization guarantees for denoising score matching. We investigate whether the same statistical benefit persists for Flow Matching. For fixed $\sigma > 0$, the interpolation variance remains uniformly positive on $t \in [0, 1]$, so the present analysis naturally covers the full time interval.

Our results show that FM retains latent-dimensional statistical behavior in this full-dimensional noisy setting. Using the latent posterior structure, we construct a spatially regular ReLU velocity class and establish a non-asymptotic high-probability bound for its empirical risk minimizer, with leading sample-size factor $n^{-2\beta/(2\beta+d)}$. The remaining dependence on the ambient dimension D, the Hölder radius H, the noise level σ, the confidence parameter, and logarithmic factors is kept explicit in our main theorem. We then propagate this velocity-estimation error through the learned ODE and obtain a corresponding non-asymptotic W_2 bound with leading sample-size factor $n^{-\beta/(2\beta+d)}$. Thus, the algebraic dependence on the sample size is governed by the latent dimension d, even though the observed target distribution has full ambient support.

1.1 Main Contributions

Our main contributions are summarized as follows.

- **Latent-dimensional velocity approximation.** Under the noisy Hölder latent-generator model, we construct a spatially regular ReLU approximation of the population FM velocity whose leading resolution complexity scales as ε^{-d}, even though the target distribution has full ambient support.
- **Non-asymptotic velocity estimation.** For empirical risk minimization over the corresponding spatially regular velocity class, we establish a high-probability excess-risk bound with leading sample-size factor $n^{-2\beta/(2\beta+d)}$, while keeping the dependence on D, H, σ, and the confidence level explicit.
- **Wasserstein recovery.** The same spatial regularity propagates the velocity-estimation guarantee through the continuous learned flow, yielding a non-asymptotic W_2 bound with leading sample-size factor $n^{-\beta/(2\beta+d)}$. For fixed positive target noise, the continuous-flow analysis applies on the full interval $t \in [0, 1]$.

1.2 Other Related Work

Generative learning with noisy observations. A related line of work studies generative modeling when training observations are corrupted or noisy. Ambient Diffusion and subsequent developments learn an underlying clean distribution from corrupted observations using diffusion-based objectives (Daras et al., 2023; 2024); related approaches consider alternative estimators and training procedures under observation corruption (Kawar et al., 2024; Bai et al., 2024). These works treat corruption as an observation mechanism to be removed or inverted, whereas in our setting the nondegenerate ambient perturbation is part of the target distribution itself. Noise smoothing also appears in continuous generative modeling through dequantization, where discrete observations are perturbed to obtain a continuous density suitable for flow-based models (Ho et al., 2019; Theis et al., 2016; Nielsen & Winther, 2020). This provides a separate algorithmic precedent for working with full-dimensional perturbed observations, rather than a statistical assumption underlying our analysis. Flow Matching has also been extended to prescribed non-Euclidean state spaces through Riemannian and geometric constructions (Chen & Lipman, 2024); our setting remains Euclidean and instead concerns statistical complexity induced by a low-dimensional latent generating mechanism.

2 Problem Setup and Preliminaries

2.1 Notation and conventions

For vectors, $\|\cdot\|$ and $\|\cdot\|_\infty$ denote the Euclidean and coordinatewise maximum norms, respectively. For a vector-valued function f, write $\|f\|_{L^\infty(\Omega)} := \sup_{u\in\Omega} \|f(u)\|$. We use $a \vee b := \max\{a, b\}$ and

$a \wedge b := \min\{a, b\}$. For $\beta, H > 0$, $\mathcal{H}^\beta(\Omega, \mathbb{R}^m, H)$ denotes the coordinatewise Hölder class used throughout the paper; its precise multi-index definition is given in Appendix A.1. For nonnegative quantities, $A \lesssim_\Theta B$ means $A \le C_\Theta B$ for a constant depending only on the parameters in Θ; when the subscript is omitted, the constant is universal. We write $A \asymp_\Theta B$ when both $A \lesssim_\Theta B$ and $B \lesssim_\Theta A$ hold. Throughout, $\log$ denotes the natural logarithm and e Euler's number.

2.2 Noisy latent Flow Matching

Target model and interpolation. Let d and D denote the latent and ambient dimensions, respectively. We consider the target model

$$X_1 = g^*(U) + \sigma\xi, \qquad U \sim \mathrm{Unif}([0,1]^d), \qquad \xi \sim \mathcal{N}(0, I_D), \tag{1}$$

where $0 < \sigma < 1$ is fixed and

$$g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H), \qquad \|g^*\|_{L^\infty([0,1]^d)} \le 1. \tag{2}$$

The variables U and ξ are independent. For every fixed $\sigma > 0$, the target law π_1 is full-dimensional; the low-dimensional structure enters through the latent generator g^*, rather than through exact low-dimensional support. Let $X_0 \sim \mathcal{N}(0, I_D)$ be independent of (U, ξ) and consider the linear interpolation

$$X_t = (1-t)X_0 + tX_1, \qquad t \in [0,1]. \tag{3}$$

We write $\pi_t = \mathrm{Law}(X_t)$ and $\|h(\cdot,t)\|^2_{L^2(\pi_t)} := \mathbb{E}\|h(X_t,t)\|^2$. Conditional on $U = u$, $X_t \mid U = u \sim \mathcal{N}(tg^*(u), q_t I_D)$, $q_t := (1-t)^2 + \sigma^2 t^2$. Since $q_t \ge \sigma^2/(1+\sigma^2) > 0$, the interpolation remains nondegenerate throughout $t \in [0,1]$.

Population objective. The population Flow Matching risk is

$$\mathcal{L}(v) := \int_0^1 \mathbb{E}_{X_0, X_1}\left[\|X_1 - X_0 - v(X_t, t)\|^2\right] dt,$$

whose population velocity is

$$v^*(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x].$$

Our statistical experiment observes $X_{1,1}, \ldots, X_{1,n} \overset{\text{iid}}{\sim} \pi_1$. The Gaussian source and interpolation time are auxiliary variables and are integrated exactly in the idealized empirical loss introduced in Section 4.

Posterior representation. Define

$$a_t := \frac{t\sigma^2 - (1-t)}{q_t}, \qquad b_t := \frac{1-t}{q_t}, \qquad m_t(x) := \mathbb{E}[g^*(U) \mid X_t = x].$$

The following representation reduces the nonlinear part of the population velocity to the latent posterior mean m_t.

Lemma 2.1 (Posterior representation of the population velocity)**.** *Under the preceding model, the conditional means admit versions such that, for every $(x,t) \in \mathbb{R}^D \times [0,1]$,*

$$v^*(x,t) = a_t x + b_t m_t(x), \qquad \|m_t(x)\| \le 1,$$

where

$$m_t(x) = \frac{\displaystyle\int_{[0,1]^d} g^*(u) \exp\left(-\frac{\|x - tg^*(u)\|^2}{2q_t}\right) du}{\displaystyle\int_{[0,1]^d} \exp\left(-\frac{\|x - tg^*(u)\|^2}{2q_t}\right) du}.$$

The proof is given in Appendix A. The decomposition separates the known affine component $a_t x$ from the posterior component $b_t m_t(x)$ that is approximated by the neural network in Section 3.

2.3 NEURAL NETWORK CLASS

Throughout the approximation and statistical analysis, we use sparse feed-forward ReLU networks. Let $\rho(z) = \max\{0, z\}$ act componentwise. For $L, W, S \in \mathbb{N}$ and $B \geq 1$, denoting depth, maximal width, sparsity, and the parameter-magnitude bound, respectively, define

$$\mathrm{NN}(L, W, S, B) := \left\{ \begin{array}{l} (A_L\rho(\cdot) + b_L) \circ \cdots \circ (A_2\rho(\cdot) + b_2) \circ (A_1 x + b_1) : \\ A_\ell \in \mathbb{R}^{W_\ell \times W_{\ell-1}},\ b_\ell \in \mathbb{R}^{W_\ell},\ \max_{0\leq \ell \leq L} W_\ell \leq W, \\ \sum_{\ell=1}^{L} (\|A_\ell\|_0 + \|b_\ell\|_0) \leq S, \quad \max_{1\leq \ell \leq L} (\|A_\ell\|_\infty \vee \|b_\ell\|_\infty) \leq B \end{array} \right\}.$$

Here $\theta = \{(A_\ell, b_\ell)\}_{\ell=1}^{L}$ denotes the trainable parameters, $\|\cdot\|_0$ counts nonzero entries, and $\|\cdot\|_\infty$ is the largest absolute entry. For the velocity network, $W_0 = D + 1$ and $W_L = D$; auxiliary networks use their stated dimensions.

3 NEURAL APPROXIMATION OF THE POPULATION VELOCITY

This section establishes a neural approximation theorem for the population FM velocity under the noisy latent-generator model. By the posterior representation in Section 2, $v^*(x,t) = a_t x + b_t m_t(x)$, the nonlinear approximation problem reduces to the posterior mean m_t. The main difficulty is that m_t is defined on the unbounded ambient state space, while the exploitable structure is inherited from the d-dimensional latent generator; the approximation must also retain the spatial regularity required for the subsequent transport analysis.

Let clip_2 be Euclidean projection onto $\mathcal{B}(0,2)$. For a deterministic spatial envelope $\gamma_x : [0,1] \to [0,\infty)$, define

$$\begin{aligned} \mathcal{V}(L, W, S, B, \gamma_x) := \big\{ v_\theta : v_\theta(x,t) = a_t x + b_t\, \mathrm{clip}_2(f_\theta(x,t)), \quad f_\theta \in \mathrm{NN}(L, W, S, B), \\ \mathrm{Lip}_x(v_\theta(\cdot, t)) \leq \gamma_x(t) \quad (t \in [0,1]) \big\}. \end{aligned} \tag{4}$$

The network approximates the nonlinear posterior component, while a_t and b_t remain the exact coefficients of the linear path. No separate temporal regularity constraint is imposed on this class. Let $\varepsilon \in (0, 1/2]$ satisfy the scale conditions

$$\varepsilon^{-2\beta} > D, \qquad H\sqrt{D}\,\varepsilon^{\beta\wedge 1} \leq c_{d,\beta}\sigma, \tag{5}$$

$$\varepsilon\left[1 + \frac{1+\sigma^2}{\sigma^2} D(H \vee 1)\left\{2 + \sqrt{D} + 16\left(\sqrt{D\log(\varepsilon^{-2\beta}/D)} \vee \log(\varepsilon^{-2\beta}/D)\right)\right\}\right] \leq c_{d,\beta}, \tag{6}$$

where $c_{d,\beta} > 0$ is sufficiently small.

Theorem 3.1 (Approximation of the population velocity)**.** *Assume* (1)–(2)*, with* $0 < \sigma < 1$*, and let* $\varepsilon \in (0, 1/2]$ *satisfy* (5)– (6)*. Set* $\Gamma_\varepsilon := 1 + D + \log(1/\varepsilon) + \log(H \vee e) + \log(1/\sigma)$*. For a sufficiently large fixed constant* C *depending only on* (d, β)*, choose* $\gamma_x(t) := C\left[|a_t| + |b_t|\left(1 + \sigma^{-1} + \frac{t}{q_t}\right)\right]$*. Fix one finite exponent* $c_{\mathrm{arch}} = c_{\mathrm{arch}}(d, \beta)$*, large enough to dominate the first-order realization and the covering-number calculation below. Then network parameters can be chosen so that*

$$L, \log B \lesssim_{d,\beta} \Gamma_\varepsilon^{c_{\mathrm{arch}}}, \qquad W, S \lesssim_{d,\beta} D^2 \varepsilon^{-d} \Gamma_\varepsilon^{c_{\mathrm{arch}}},$$

and $\mathcal{V}(L, W, S, B, \gamma_x)$ *contains a velocity field* v_θ *satisfying*

$$\int_0^1 \|v_\theta(\cdot, t) - v^*(\cdot, t)\|^2_{L^2(\pi_t)}\, dt \lesssim_{d,\beta} \frac{D(1 + H^2)}{\sigma}\, \varepsilon^{2\beta}. \tag{7}$$

Moreover, $\|\gamma_x\|_{L^1(0,1)} \lesssim_{d,\beta} 1 + \sigma^{-2}$ *and* $\|\gamma_x\|_{L^\infty(0,1)} \lesssim_{d,\beta} 1 + \sigma^{-3}$.

The proof is given in Appendix B.2. The leading dependence on the approximation scale is ε^{-d}, inherited from the d-dimensional latent parameterization. The ambient dimension D remains in the explicit polynomial factors and in Γ_ε. The same comparator also satisfies the spatial profile γ_x used later in the transport analysis.

Low-dimensional approximation. The approximation complexity of FM depends strongly on the structural description of the target distribution. Existing results treat bounded-support assumptions (Zhou & Liu, 2025) and B-spline approximation under Besov regularity of full-dimensional target densities (Fukumizu et al., 2025). Intrinsic-dimensional guarantees have also been obtained from exact low-dimensional support, including smooth manifolds (Kumar et al., 2026; Kunkel & Trabs, 2025) and linear subspaces (Pi et al., 2026). Related diffusion analyses exploit manifold geometry and regularity (Tang & Yang, 2024; Azangulov et al., 2024), while Yakovlev & Puchkin (2025) study a noisy latent generator and approximate its Gaussian-mixture posterior through local approximation on the latent domain. Under the present model, the FM posterior mean is defined on the ambient state space but is generated by the d-dimensional map g^*.

Spatially regular approximation. Spatial regularity is a separate requirement because the learned velocity is later used as the drift of a deterministic ODE. Neural approximation with regularity control has been developed through norm constraints on network parameters (Jiao et al., 2023) and through approximation results that control derivatives together with function values (Gühring et al., 2020; Opschoor et al., 2022). In FM, such regularity is directly tied to the stability of the learned ODE and is incorporated into the velocity classes considered in Zhou & Liu (2025); Kunkel (2026). In the present setting, the spatial control follows from the same posterior representation used for value approximation. Gaussian smoothing gives an explicit bound on the spatial Jacobian of the posterior mean, and the neural realization is refined to control the corresponding first derivatives. The resulting comparator satisfies both the integrated velocity approximation bound and the profile γ_x.

Proof sketch. Let $N = \lceil \varepsilon^{-1} \rceil$ and partition $[0,1]^d$ into N^d latent cells. On each cell, approximate g^* by its Taylor polynomial of degree $r_\beta := \lceil \beta \rceil - 1$. The β-Hölder regularity of g^* gives

$$\|g^* - g^\circ\|_{L^\infty([0,1]^d)} \lesssim_{d,\beta} H\sqrt{D}\,\varepsilon^\beta.$$

Let v° denote the population FM velocity induced by the surrogate target $g^\circ(U)+\sigma\xi$. The following proposition transfers this latent approximation to the corresponding velocity error.

Proposition 3.2 (Generator-to-velocity stability)**.** *Let $g^*, g^\circ : [0,1]^d \to \mathbb{R}^D$ be bounded and measurable, and let v^*, v° be the population velocities generated by the targets $g^*(U) + \sigma\xi$ and $g^\circ(U) + \sigma\xi$, respectively, under the same linear interpolation with $\sigma > 0$. Then*

$$\int_0^1 \|v^*(\cdot,t) - v^\circ(\cdot,t)\|^2_{L^2(\pi_t)}\,dt \le \frac{\pi}{4\sigma}\|g^* - g^\circ\|^2_{L^\infty([0,1]^d)},$$

where π_t is the marginal law of the true interpolation.

The proof is given in Appendix B.3.2 and is based on a Gaussian-mixture relative-entropy comparison. It remains to approximate v°. Since $v^\circ(x,t) = a_t x + b_t m_t^\circ(x)$, the nonlinear part reduces to the surrogate posterior mean m_t°. Gaussian concentration localizes the approximation to a high-probability spatial region, on which the cellwise polynomial structure of g° reduces the posterior numerator and denominator to low-dimensional integral terms. This concentration-based localization also avoids imposing a bounded ambient support assumption. Lemma B.3 realizes these terms and their ratio while controlling the spatial Jacobian. Combining this posterior realization with Proposition 3.2 gives the approximation rate, network complexity, and spatial profile stated in Theorem 3.1.

4 GENERALIZATION

Section 3 shows that the structured velocity class contains a comparator with small population approximation error. We now study whether this accuracy can be recovered from finitely many target observations. The starting point is the following population identity, which identifies the statistical target with the integrated velocity error.

Lemma 4.1 (Population risk identity)**.** *For every measurable v with finite population risk,*

$$\mathcal{L}(v) - \mathcal{L}(v^*) = \int_0^1 \|v(\cdot,t) - v^*(\cdot,t)\|^2_{L^2(\pi_t)}\,dt. \tag{8}$$

Thus, controlling the excess Flow Matching risk is equivalent to controlling the time-integrated $L^2(\pi_t)$ error of the learned velocity field. The proof is given in Appendix C.2.

Empirical evaluation. We observe $X_{1,1}, \dots, X_{1,n} \overset{\text{iid}}{\sim} \pi_1$. For $x_1 \in \mathbb{R}^D$ and a measurable velocity field v, define

$$\ell(x_1, v) := \int_0^1 \mathbb{E}_{X_0} \left\| x_1 - X_0 - v((1-t)X_0 + tx_1, t) \right\|^2 dt, \qquad \widehat{\mathcal{L}}_n(v) := \frac{1}{n} \sum_{i=1}^n \ell(X_{1,i}, v). \tag{9}$$

Then $\mathcal{L}(v) = P\ell(\cdot, v)$, where P denotes expectation with respect to $X_1 \sim \pi_1$. As in Fukumizu et al. (2025), the auxiliary Gaussian source and the time variable are integrated exactly inside ℓ, so the finite-sample randomness comes only from the target observations $X_{1,1:n}$. Finite auxiliary Monte Carlo sampling would introduce an additional error term; see, e.g., Zhou & Liu (2025). For fixed L, W, S, B, define the empirical risk minimizer

$$\widehat{v} \in \arg \min_{v \in \mathcal{V}(L,W,S,B,\gamma_x)} \widehat{\mathcal{L}}_n(v). \tag{10}$$

4.1 Error Decomposition

The excess risk of $\widehat{v}$ decomposes as

$$\mathcal{L}(\widehat{v}) - \mathcal{L}(v^*) = \underbrace{\mathcal{L}(\widehat{v}) - \inf_{v \in \mathcal{V}(L,W,S,B,\gamma_x)} \mathcal{L}(v)}_{\text{estimation}} + \underbrace{\inf_{v \in \mathcal{V}(L,W,S,B,\gamma_x)} \left\{ \mathcal{L}(v) - \mathcal{L}(v^*) \right\}}_{\text{approximation}}. \tag{11}$$

Section 3 controls the approximation term. To analyze the estimation term, define the signed excess loss $\Delta_v(x_1) := \ell(x_1, v) - \ell(x_1, v^*)$, and write $Ph := \mathbb{E}[h(X_1)]$ and $P_n h := n^{-1} \sum_{i=1}^n h(X_{1,i})$. By Lemma 4.1, $P\Delta_v = \mathcal{L}(v) - \mathcal{L}(v^*) \geq 0$, although $\Delta_v(x_1)$ need not be pointwise nonnegative. The squared Flow Matching loss is unbounded. For a real random variable Z, write $\|Z\|_{\psi_1} := \inf\{c > 0 : \mathbb{E} \exp(|Z|/c) \leq 2\}$. The following lemma gives the localized moment and tail bounds needed for an unbounded Bernstein inequality.

Lemma 4.2 (Localized Flow Matching excess loss)**.** *For every $v \in \mathcal{V}(L, W, S, B, \gamma_x)$,*

$$P\Delta_v^2 \lesssim \frac{1}{\sigma^2} P\Delta_v, \qquad \|\Delta_v - P\Delta_v\|_{\psi_1} \lesssim \frac{1}{\sigma} + \sqrt{D}, \tag{12}$$

where the hidden constants are universal.

Consequently, for every fixed deterministic $v \in \mathcal{V}(L, W, S, B, \gamma_x)$ and every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$|(P - P_n)\Delta_v| \lesssim \sqrt{\frac{P\Delta_v \, \log(4/\delta)}{\sigma^2 n}} + \frac{(\sigma^{-1} + \sqrt{D}) \log(en) \log(4/\delta)}{n}. \tag{13}$$

4.2 Covering and Generalization

Definition 4.3 (Covering number)**.** Let $(\mathcal{M}, d)$ be a pseudo-metric space and let $\mathcal{F} \subseteq \mathcal{M}$. For $\eta > 0$, a set $\mathcal{A} \subseteq \mathcal{M}$ is called an η-cover of $\mathcal{F}$ if for every $f \in \mathcal{F}$ there exists $\widetilde{f} \in \mathcal{A}$ such that $d(f, \widetilde{f}) \leq \eta$. The covering number $\mathcal{N}(\eta, \mathcal{F}, d)$ is the minimum cardinality of an η-cover of $\mathcal{F}$.

To make the fixed-field bound uniform over the excess-loss class, we first control the complexity of the underlying neural networks. By the standard sparse-ReLU covering estimate of Suzuki (2019, Lemma 3), adapted to D-dimensional outputs, for every $R \geq 1$, $\eta \in (0, 1)$, and $W \geq D + 1$,

$$\log \mathcal{N}\big(\eta, \mathrm{NN}(L, W, S, B), \|\cdot\|_{L^\infty(\mathcal{B}(0,R) \times [0,1]; \ell_2)}\big) \lesssim SL \log \left(\frac{\sqrt{D}\, L(W+1)(B \vee 1)(R+1)}{\eta} \right). \tag{14}$$

The network covering estimate is local to a bounded state–time domain. Gaussian concentration and the local stability of the conditional Flow Matching loss allow the same network representatives to control both the population and empirical excess losses.

Lemma 4.4 (Finite reduction of the Flow Matching excess-loss class)**.** *For every $\tau, \delta \in (0,1)$, there exists a deterministic family $v_1, \dots, v_{M_{\tau,\delta}} \in \mathcal{V}(L, W, S, B, \gamma_x)$ such that, with probability at least $1-\delta$, every $v \in \mathcal{V}(L, W, S, B, \gamma_x)$ admits a representative v_j satisfying*

$$|P(\Delta_v - \Delta_{v_j})| + |P_n(\Delta_v - \Delta_{v_j})| \le \tau. \tag{15}$$

Moreover,

$$\log M_{\tau,\delta} \lesssim SL \log\left[\frac{\sqrt{D}\, L(W+1)(B \vee 1)}{\tau}\left(1 + \frac{1}{\sigma} + \sqrt{D} + \sqrt{\log\frac{en}{\delta}} + \sqrt{\log\frac{e}{\tau}}\right)^2\right]. \tag{16}$$

The proof is given in Appendix C.3. Applying (13) to the finite representatives, taking a union bound, and using the ERM comparison yield, with probability at least $1-\delta$,

$$\begin{aligned} P\Delta_{\widehat{v}} \lesssim & \inf_{v \in \mathcal{V}(L,W,S,B,\gamma_x)} P\Delta_v + \tau \\ & + \left[\frac{1}{\sigma^2} + \left(\frac{1}{\sigma} + \sqrt{D}\right)\log(en)\right]\frac{\log(8M_{\tau,\delta/2}/\delta)}{n}. \end{aligned} \tag{17}$$

Combining (17) with Theorem 3.1 yields the finite-sample rate below.

Theorem 4.5 (Finite-sample velocity learning)**.** *Assume* (1)*,* (2)*, and* (3)*, with $0 < \sigma < 1$. Given $X_{1,1}, \dots, X_{1,n} \overset{\text{iid}}{\sim} \pi_1$, set $\varepsilon := n^{-1/(2\beta+d)}$ and suppose that ε satisfies* (5)–(6)*. Choose $\mathcal{V}(L, W, S, B, \gamma_x)$ according to Theorem 3.1 at resolution ε, and let $\widehat{v}$ be the ERM in* (10)*. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the target observations,*

$$\begin{aligned} \int_0^1 \|\widehat{v}(\cdot,t) - v^*(\cdot,t)\|^2_{L^2(\pi_t)}\, dt \lesssim_{d,\beta} & n^{-\frac{2\beta}{2\beta+d}}\Bigg[\frac{D(1+H^2)}{\sigma} \\ & + D^2\big[\sigma^{-2} + (\sigma^{-1} + \sqrt{D})\log(en)\big]\big[\Gamma_\varepsilon + \log(e/\delta)\big]^{c_{\text{arch}}}\Bigg]. \end{aligned} \tag{18}$$

The algebraic sample-size exponent $2\beta/(2\beta+d)$ is determined by the latent dimension d and the Hölder smoothness β of the generator. The ambient dimension D, noise level σ, Hölder radius H, and confidence level δ remain explicit in the prefactor and logarithmic complexity term.

Proof sketch. For a fixed velocity field, the conditional Flow Matching excess loss is unbounded, but its squared-loss structure yields a localized moment bound. The conditional covariance of the Flow Matching label, together with the discrepancy bound obtained from affine cancellation and radius-two clipping, gives $P\Delta_v^2 \lesssim \sigma^{-2} P\Delta_v$, while the Gaussian tail of the target observation gives a uniform sub-exponential envelope for $\Delta_v - P\Delta_v$. These estimates yield the fixed-field Bernstein inequality (13).

To pass from a fixed field to the data-dependent ERM, we reduce the velocity class to finitely many deterministic representatives. The sparse-ReLU covering estimate is available on bounded state–time domains; Gaussian localization and the local stability of the conditional loss allow the same representatives to control both P and P_n. A union bound and the ERM comparison then give (17). Finally, the Section 3 comparator has approximation scale $\varepsilon^{2\beta}$, while the corresponding entropy contribution has scale ε^{-d}/n, which yields the stated choice of ε and the final rate. The complete proof is given in Appendix C.1.

5 Distributional Convergence of the Learned Flow

We now translate this statistical error into distributional error of the induced flow. For a measurable map T, let $T_{\#}\mu$ denote the pushforward of μ, and let W_2 denote the 2-Wasserstein distance.

5.1 Continuous-time transport

Let Φ_t denote the flow generated by the population velocity, $\partial_t \Phi_t(x) = v^*(\Phi_t(x), t)$ with $\Phi_0(x) = x$.

Lemma 5.1 (Population transport)**.** *The population flow is globally well posed and satisfies* $(\Phi_t)_{\#}\pi_0 = \pi_t$ *for every* $t \in [0,1]$.

Proof. Let $G = g^*(U)$. Differentiating the posterior mean gives

$$\nabla_x m_t(x) = \frac{t}{q_t}\operatorname{Cov}(G \mid X_t = x),$$

and hence

$$\nabla_x v^*(x,t) = a_t I_D + \frac{t(1-t)}{q_t^2}\operatorname{Cov}(G \mid X_t = x).$$

Since $\|G\| \le 1$, the conditional covariance has operator norm at most one. Therefore $\operatorname{Lip}_x(v^*(\cdot,t)) \le |a_t| + t(1-t)/q_t^2$, while the posterior representation also gives $\|v^*(x,t)\| \le |a_t|\|x\| + |b_t|$. Because q_t is uniformly bounded away from zero on $[0,1]$, these bounds yield a finite spatial Lipschitz profile and uniform linear growth. The characteristic ODE therefore admits a unique global flow.

For every $\varphi \in C_c^\infty(\mathbb{R}^D)$,

$$\frac{d}{dt}\mathbb{E}\varphi(X_t) = \mathbb{E}\big[\nabla\varphi(X_t)^\top(X_1 - X_0)\big] = \mathbb{E}\big[\nabla\varphi(X_t)^\top v^*(X_t,t)\big].$$

Thus $(\pi_t)_{t\in[0,1]}$ solves the continuity equation associated with v^*. Since v^* is globally Lipschitz in space with an integrable Lipschitz profile, the characteristic solution is unique, and hence $(\Phi_t)_{\#}\pi_0 = \pi_t$ for all $t \in [0,1]$. □

Couple the population and learned flows through the common initial condition $Z_0 = \widehat{Z}_0 \sim \pi_0$, with $\dot{Z}_t = v^*(Z_t,t)$ and $\dot{\widehat{Z}}_t = \widehat{v}(\widehat{Z}_t,t)$, and write $\widehat{\pi}_t := \operatorname{Law}(\widehat{Z}_t)$. By Lemma 5.1, $Z_t \sim \pi_t$.

Theorem 5.2 (Continuous-flow Wasserstein convergence)**.** *Under the conditions of Theorem 4.5, the learned ODE is globally well posed and*

$$W_2^2(\widehat{\pi}_1,\pi_1) \le \exp\big\{1 + 2\|\gamma_x\|_{L^1(0,1)}\big\}\int_0^1 \|\widehat{v}(\cdot,t) - v^*(\cdot,t)\|_{L^2(\pi_t)}^2\,dt. \tag{19}$$

Consequently, with probability at least $1-\delta$ *over the target observations, the resulting nonasymptotic* W_2 *bound has leading sample-size factor* $n^{-\beta/(2\beta+d)}$.

Proof. Condition on the target observations. Since $\widehat{v} \in \mathcal{V}(L,W,S,B,\gamma_x)$, the spatial class constraint and Lemma D.1 give global spatial Lipschitz continuity and uniform linear growth. Hence the learned ODE admits a unique global solution.

Set $R_t := \mathbb{E}\|Z_t - \widehat{Z}_t\|^2$, where the expectation is with respect to the common initial condition. Since $(Z_t,\widehat{Z}_t)$ is a coupling of $(\pi_t,\widehat{\pi}_t)$, $W_2^2(\widehat{\pi}_t,\pi_t) \le R_t$. For almost every t,

$$\begin{aligned} R_t' &= 2\mathbb{E}\left\langle Z_t - \widehat{Z}_t,\, v^*(Z_t,t) - \widehat{v}(\widehat{Z}_t,t)\right\rangle \\ &= 2\mathbb{E}\left\langle Z_t - \widehat{Z}_t,\, v^*(Z_t,t) - \widehat{v}(Z_t,t)\right\rangle \\ &\quad + 2\mathbb{E}\left\langle Z_t - \widehat{Z}_t,\, \widehat{v}(Z_t,t) - \widehat{v}(\widehat{Z}_t,t)\right\rangle. \end{aligned}$$

Using $2\langle a,b\rangle \le \|a\|^2 + \|b\|^2$ for the first term, $Z_t \sim \pi_t$, and $\operatorname{Lip}_x(\widehat{v}(\cdot,t)) \le \gamma_x(t)$ for the second gives

$$R_t' \le \{1 + 2\gamma_x(t)\}R_t + \|\widehat{v}(\cdot,t) - v^*(\cdot,t)\|_{L^2(\pi_t)}^2, \qquad R_0 = 0.$$

The time-dependent Grönwall inequality therefore yields, for every $t \in [0,1]$,

$$R_t \le \int_0^t \exp\left\{(t-s) + 2\int_s^t \gamma_x(u)\,du\right\}\|\widehat{v}(\cdot,s) - v^*(\cdot,s)\|_{L^2(\pi_s)}^2\,ds.$$

Setting $t = 1$ and using $W_2^2(\widehat{\pi}_1,\pi_1) \le R_1$ proves (19).

Finally, Theorem 3.1 gives $\|\gamma_x\|_{L^1(0,1)} \lesssim_{d,\beta} 1 + \sigma^{-2}$. Combining (19) with Theorem 4.5 and taking square roots yields the stated nonasymptotic Wasserstein consequence. □

Because the fixed positive target noise keeps q_t uniformly bounded away from zero, the continuous-flow bound applies directly at the target endpoint $t = 1$.

5.2 EULER DISCRETIZATION

The continuous-flow result above uses the spatial regularity built into the learning class. To control the additional numerical error introduced by a time discretization, we impose the following temporal regularity condition only in this subsection.

Assumption 5.3 (Temporal regularity of the learned velocity)**.** There exists a deterministic constant $\gamma_t < \infty$, independent of n, such that, almost surely with respect to the target observations,

$$\|\widehat{v}(x,t) - \widehat{v}(x,s)\| \le \gamma_t(1 + \|x\|)|t-s|, \qquad x \in \mathbb{R}^D, \quad s,t \in [0,1].$$

Let $0 = t_0 < t_1 < \cdots < t_M = 1, h := \max_{0\le k<M}(t_{k+1} - t_k)$, be a deterministic time grid. Starting from $\widetilde{Z}_{t_0} \sim \pi_0$, define the explicit Euler scheme

$$\widetilde{Z}_{t_{k+1}} = \widetilde{Z}_{t_k} + (t_{k+1} - t_k)\widehat{v}(\widetilde{Z}_{t_k}, t_k), 0 \le k < M, \tag{20}$$

and let $\widetilde{\pi}_1 := \mathrm{Law}(\widetilde{Z}_{t_M})$. Set $\bar{\gamma}_x := \|\gamma_x\|_{L^\infty(0,1)}, K_\sigma^{\mathrm{sam}} := \max\left\{\sup_{t\in[0,1]}|a_t|, 2\sup_{t\in[0,1]}|b_t|\right\}$.

Proposition 5.4 (Euler discretization error)**.** *Under the conditions of Theorem 4.5 and Assumption 5.3, with probability at least* $1-\delta$ *over the target observations,*

$$\begin{aligned} W_2(\widetilde{\pi}_1, \pi_1) \le{}& \frac{1}{2}(1+\sqrt{D})e^{K_\sigma^{\mathrm{sam}} + \bar{\gamma}_x}\left(\bar{\gamma}_x K_\sigma^{\mathrm{sam}} + \gamma_t\right)h \\ &+ \exp\left\{\frac{1}{2} + \|\gamma_x\|_{L^1(0,1)}\right\}\left(\int_0^1 \|\widehat{v}(\cdot,t) - v^*(\cdot,t)\|^2_{L^2(\pi_t)}\,dt\right)^{1/2}. \end{aligned} \tag{21}$$

The first term in (21) is the numerical discretization error, whereas the second is the continuous-flow statistical error controlled by Theorem 4.5. Consequently, the leading algebraic contributions in the time step and sample size are h and $n^{-\beta/(2\beta+d)}$, respectively. For a uniform grid, $h = M^{-1}$, so choosing M at least of order $n^{\beta/(2\beta+d)}$ balances these two leading scales. The proof of Proposition 5.4 is given in Appendix D.2.

6 DISCUSSION AND CONCLUSION

We studied linear Flow Matching when high-dimensional observations are generated by a low-dimensional Hölder mechanism with nondegenerate ambient noise. Although the resulting target distribution has full-dimensional support, its population velocity retains a latent posterior structure that can be approximated and learned with a leading complexity governed by the latent dimension. This yields a nonasymptotic high-probability velocity bound with leading sample-size factor $n^{-2\beta/(2\beta+d)}$, while the dependence on the ambient dimension and noise level remains explicit. By carrying the required spatial regularity through the same learning class, the velocity guarantee further propagates to the learned continuous flow, giving a W_2 bound with leading sample-size factor $n^{-\beta/(2\beta+d)}$. For fixed $\sigma > 0$, the interpolation remains nondegenerate on the full interval $[0,1]$. Overall, the results show that exact low-dimensional support is not necessary for Flow Matching to retain latent-dimensional statistical behavior.

The analysis also identifies several boundaries of this conclusion. The ambient dimension D remains in the prefactors and network complexity, and the constants deteriorate as σ decreases, so the present theory does not provide a uniform passage to the exact-support limit. Moreover, the established rates are upper bounds rather than a minimax characterization: proving minimax optimality would require matching lower bounds for the same noisy latent-generator class and integrated velocity risk, and a separate lower-bound analysis is needed for the induced Wasserstein error. Extending the theory to vanishing noise, unknown latent dimension or smoothness, and finite auxiliary sampling remains an important direction for future work. These questions would further clarify when the statistical complexity of Flow Matching is determined by the latent generating mechanism rather than by the ambient observation space.

AI USE STATEMENT

Generative AI tools, including ChatGPT, were used to assist with adversarial proof checking and language editing. In particular, these tools were used to help examine intermediate mathematical claims, identify potential gaps in proof interfaces, and suggest revisions to the exposition. All mathematical statements, assumptions, proofs, citations, and final manuscript text were reviewed by the authors, who take full responsibility for the correctness and originality of the work.

REPRODUCIBILITY STATEMENT

All assumptions, probability models, learning objectives, and neural-network classes used in the theoretical analysis are stated explicitly in the main text. Complete proofs of the approximation, finite-sample generalization, continuous-flow Wasserstein convergence, and auxiliary neural-network construction results are provided in the appendix, together with the technical lemmas needed to verify their assumptions and parameter dependence. The paper does not rely on empirical datasets or numerical experiments.

REFERENCES

Michael S. 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. No publisher DOI assigned; arXiv:2209.15571.

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions, 2024. URL https://arxiv.org/abs/2409.18804. Preprint.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0613. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/22912080c32a244c20bfc101033a6ac7-Abstract-Conference.html.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=uqQPyWFDhY. No publisher DOI assigned; arXiv:2305.16860.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/618c95f4557c15b253fb0e6f548ea0c0-Abstract-Conference.html. No publisher DOI assigned; arXiv:2310.02391.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23o.html. No publisher DOI assigned.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL. No publisher DOI assigned; arXiv:2302.03660.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0015. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/012af729c5d14d279581fc8a5db975a1-Abstract-Conference.html.

Giannis Daras, Alex Dimakis, and Constantinos Costis Daskalakis. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10091–10108. PMLR, 2024. URL https://proceedings.mlr.press/v235/daras24a.html. No publisher DOI assigned.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=MhK5aXo3gB. No publisher DOI assigned; arXiv:2208.05314.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html. No publisher DOI assigned.

Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Fitting a putative manifold to noisy data. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 688–720. PMLR, 2018. URL https://proceedings.mlr.press/v75/fefferman18a.html. No publisher DOI assigned.

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. No publisher DOI assigned; arXiv:2405.20879.

Yuan Gao, Jian Huang, and Yuling Jiao. Gaussian interpolation flows. *Journal of Machine Learning Research*, 25(253):1–52, 2024. URL https://www.jmlr.org/papers/v25/23-1515.html. No publisher DOI assigned; arXiv:2311.11475.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012. doi: 10.1214/12-AOS994. URL https://doi.org/10.1214/12-AOS994.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020. doi: 10.1142/S0219530519410021. URL https://doi.org/10.1142/S0219530519410021.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2722–2730. PMLR, 2019. URL https://proceedings.mlr.press/v97/ho19a.html. No publisher DOI assigned.

Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and GANs. *Applied and Computational Harmonic Analysis*, 65:249–278, 2023. doi: 10.1016/j.acha.2023.03.004. URL https://doi.org/10.1016/j.acha.2023.03.004.

Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-based diffusion model training with corrupted data. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=BRl7fqMwaJ. No publisher DOI assigned; arXiv:2305.13128.

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

Lea Kunkel. Distribution estimation via flow matching with lipschitz guarantees. *Journal of Multivariate Analysis*, 215:105657, 2026. doi: 10.1016/j.jmva.2026.105657. URL https://doi.org/10.1016/j.jmva.2026.105657.

Lea Kunkel and Mathias Trabs. On the minimax optimality of flow matching through the connection to kernel density estimation, 2025. URL https://arxiv.org/abs/2504.13336. Preprint; revised in 2026.

Neil D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005. URL https://www.jmlr.org/papers/v6/lawrence05a.html. No publisher DOI assigned.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0618. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2d8911db9ecedf866015091b28946e15-Abstract-Conference.html.

Guillaume Lecué and Shahar Mendelson. General nonexact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40(2):832–860, 2012. doi: 10.1214/11-AOS965. URL https://doi.org/10.1214/11-AOS965.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. No publisher DOI assigned; arXiv:2210.02747.

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. No publisher DOI assigned; arXiv:2209.03003.

Didrik Nielsen and Ole Winther. Closing the dequantization gap: PixelCNN as a single-layer flow. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/26ed695e9b7b9f6463ef4bc1fd74fc87-Abstract.html. No confirmed publisher DOI.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582. PMLR, 2023. URL https://proceedings.mlr.press/v202/oko23a.html.

Joost A. A. Opschoor, Christoph Schwab, and Jakob Zech. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constructive Approximation*, 55:537–582, 2022. doi: 10.1007/s00365-021-09542-5.

Sophia Pi, Mingcheng Lu, Maojiang Su, Weimin Wu, Jerry Yao-Chieh Hu, and Han Liu. Learning manifold data with flow matching. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/pdf?id=Q97luuOnqy. No publisher DOI assigned.

Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. Low-dimensional adaptation of rectified flow: A new perspective through the lens of diffusion and stochastic localization, 2026. URL https://arxiv.org/abs/2601.15500. Preprint.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1648–1656. PMLR, 2024. URL https://proceedings.mlr.press/v238/tang24a.html. No publisher DOI assigned.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *The Fourth International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.01844. No publisher DOI assigned.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196. URL https://doi.org/10.1111/1467-9868.00196.

Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5824–5891. PMLR, 2025. URL https://proceedings.mlr.press/v291/yakovlev25a.html. No publisher DOI assigned; arXiv:2502.13662.

Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78903–78932. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25l.html. No publisher DOI assigned.

A Population Identities and Structural Properties

Proof of Lemma 2.1. Write $G = g^*(U)$, $W = \sigma\xi - X_0$ and $\eta_t = (1-t)X_0 + t\sigma\xi$, so $X_t = tG + \eta_t$ and $X_1 - X_0 = G + W$. The pair (W, η_t) is centered Gaussian, independent of U, with $\mathrm{Var}(\eta_t) = q_t I_D$ and $\mathrm{Cov}(W, \eta_t) = c_t I_D$, where $c_t = t\sigma^2 - (1-t)$. Since $q_t > 0$, Gaussian conditioning gives

$$\begin{aligned}\mathbb{E}[X_1 - X_0 \mid X_t, U] &= G + \frac{c_t}{q_t}(X_t - tG) \\ &= a_t X_t + b_t G,\end{aligned}$$

using $q_t - tc_t = 1 - t$. Taking conditional expectation given X_t proves the velocity representation almost surely.

Conditional on $U = u$, X_t has Gaussian density $(2\pi q_t)^{-D/2}\exp(-\|x - tg^*(u)\|^2/(2q_t))$. Bayes' formula with the uniform latent prior gives the displayed posterior representation of m_t. Its denominator is positive for every x and t, so the ratio defines a version everywhere; we take v^* to be the version given in Lemma 2.1. The posterior is a probability measure and $\|g^*(u)\| \le 1$, whence $\|m_t(x)\| \le 1$ by the triangle inequality. At $t = 0$, the posterior equals the prior and $v^*(x, 0) = \mathbb{E}G - x$; at $t = 1$, $q_1 = \sigma^2$, $a_1 = 1$, $b_1 = 0$, and $v^*(x, 1) = x$. Both agree with the defining conditional expectations. No division by t or $1 - t$ is used. □

The same Gaussian calculation applies to any bounded surrogate generator g°: the posterior bound then becomes $\|m_t^\circ(x)\| \le \|g^\circ\|_{L^\infty}$.

A.1 Smoothness and network conventions

We use the global notation introduced in Section 2. This subsection records the multi-index, smoothness, and network-architecture conventions used in the proofs.

Let $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$. For $\mathbf{k} = (k_1, \ldots, k_p) \in \mathbb{Z}_+^p$, define

$$|\mathbf{k}| := \sum_{j=1}^p k_j, \qquad \mathbf{k}! := \prod_{j=1}^p k_j!, \qquad z^{\mathbf{k}} := \prod_{j=1}^p z_j^{k_j},$$

and

$$\partial^{\mathbf{k}} := \partial_1^{k_1} \cdots \partial_p^{k_p}, \qquad \partial^{\mathbf{0}} f := f.$$

Let $\Omega \subseteq \mathbb{R}^p$. For $\beta, H > 0$, the scalar Hölder class $\mathcal{H}^\beta(\Omega, \mathbb{R}, H)$ consists of functions $f : \Omega \to \mathbb{R}$ satisfying

$$\max_{1 \le |\mathbf{k}| \le \lfloor \beta \rfloor} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} \le H, \qquad \max_{|\mathbf{k}| = \lfloor \beta \rfloor} \sup_{\substack{x,y \in \Omega \\ x \ne y}} \frac{|\partial^{\mathbf{k}} f(x) - \partial^{\mathbf{k}} f(y)|}{\min\{1, \|x - y\|_\infty\}^{\beta - \lfloor \beta \rfloor}} \le H.$$

The first condition is understood as vacuous when $\lfloor \beta \rfloor = 0$. A vector-valued function $f = (f_1, \dots, f_m)^\top$ belongs to $\mathcal{H}^\beta(\Omega, \mathbb{R}^m, H)$ if each coordinate f_j belongs to $\mathcal{H}^\beta(\Omega, \mathbb{R}, H)$. The Euclidean supremum bound of a vector-valued function is imposed separately when required, as in (2).

For an architecture vector $\mathbf{W} = (W_0, \dots, W_L) \in \mathbb{N}^{L+1}$, write

$$\|\mathbf{W}\|_\infty := \max_{0 \le \ell \le L} W_\ell.$$

For fixed input and output dimensions, let $\mathrm{NN}(L, \mathbf{W}, S, B)$ denote the subclass of $\mathrm{NN}(L, W, S, B)$ whose layer widths are exactly $\mathbf{W}$. Equivalently,

$$\mathrm{NN}(L, W, S, B) = \bigcup_{\substack{\mathbf{W} = (W_0, \dots, W_L) \\ \|\mathbf{W}\|_\infty \le W}} \mathrm{NN}(L, \mathbf{W}, S, B),$$

where W_0 and W_L are held fixed by context. For velocity networks, $W_0 = D + 1$ and $W_L = D$; auxiliary networks use the input and output dimensions stated locally.

Throughout, L counts affine layers, and no activation is applied after the final affine layer. Width padding is performed by adding zero neurons, which preserves the realized function and does not increase the sparsity or parameter-magnitude bounds. When networks of different depths must be combined, depth alignment is carried out using exact identity padding as in Lemma B.13, with the corresponding complexity cost accounted for there.

B Proofs for Neural Approximation

B.1 Consequences of the scale conditions

The conditions (5)– (6) are the scale assumptions used throughout this proof. For every fixed (d, β, D, H, σ) with $\sigma > 0$, the corresponding range is nonempty: the first two left-hand sides have the required limits as $\varepsilon \downarrow 0$, while the last follows from $\varepsilon \log(1/\varepsilon) \to 0$ and $\varepsilon \sqrt{\log(1/\varepsilon)} \to 0$. The same conditions imply $H\sqrt{D}\, N^{-(\beta \wedge 1)} \lesssim_{d,\beta} \sigma$ for $N = \lceil \varepsilon^{-1} \rceil$ and provide the small-cell bound used in the posterior construction. All conclusions remain pointwise in the fixed noise level $\sigma > 0$.

B.2 Proof of the main approximation theorem

Proof of Theorem 3.1. **Step 1: Local polynomial surrogate.** The structural smoothness lies in the latent generator, rather than in a generic function on $\mathbb{R}^D$. We therefore begin in latent space: partition $[0,1]^d$ into N^d cells and replace g^* on each cell by its local Taylor polynomial. Set $r_\beta := \lceil \beta \rceil - 1$ and $N = \lceil \varepsilon^{-1} \rceil$. For $j = 1, \dots, N-1$, let $I_j = [(j-1)/N, j/N)$ and let $I_N = [(N-1)/N, 1]$. For $\mathbf{j} = (j_1, \dots, j_d) \in \{1, \dots, N\}^d$, define

$$\mathcal{U}_{\mathbf{j}} := \prod_{r=1}^{d} I_{j_r}, \qquad u_{\mathbf{j}} := \frac{\mathbf{j}}{N}. \tag{22}$$

These N^d half-open cells form a disjoint partition of $[0,1]^d$. On $\overline{\mathcal{U}_{\mathbf{j}}}$, let

$$g^\circ_{\mathbf{j}}(u) := \sum_{|\mathbf{k}| \le r_\beta} \frac{\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})}{\mathbf{k}!} (u - u_{\mathbf{j}})^{\mathbf{k}}, \qquad g^\circ(u) := \sum_{\mathbf{j}} g^\circ_{\mathbf{j}}(u) \mathbf{1}_{\{u \in \mathcal{U}_{\mathbf{j}}\}}. \tag{23}$$

The partition is the source of the factor $N^d \lesssim_d \varepsilon^{-d}$ in the network size. Lemma B.1 gives $\|g^* - g^\circ\|_{L^\infty} \lesssim_{d,\beta} H\sqrt{D}\,\varepsilon^\beta$ and $\|g^\circ\|_{L^\infty} \le 2$. The next step transfers this latent error to the FM velocity.

Step 2: Transfer generator error to velocity error. Proposition 3.2 controls the induced velocities directly in the integrated norm under the true path law. Let $X_1^\circ = g^\circ(U) + \sigma\xi$, and denote its posterior mean and population velocity by m_t° and $v^\circ(x,t) = a_t x + b_t m_t^\circ(x)$. For any candidate v_θ,

$$v^* - v_\theta = (v^* - v^\circ) + (v^\circ - v_\theta).$$

The first difference is the change in the conditional velocity caused by replacing g^* with g°. Proposition 3.2 and Lemma B.1 give

$$\int_0^1 \|v^*(\cdot,t) - v^\circ(\cdot,t)\|^2_{L^2(\pi_t)} dt \lesssim_{d,\beta} \frac{DH^2}{\sigma}\,\varepsilon^{2\beta}. \tag{24}$$

It remains to realize v°, equivalently its nonlinear component m_t°.

Step 3: Localize the surrogate posterior. The state variable x is unbounded, whereas the posterior circuit is constructed uniformly on a controlled state region. We therefore split the velocity error into uniform posterior approximation on a time-dependent region $\mathcal{K}_t$ and a Gaussian-tail contribution outside that region. The surrogate posterior has the explicit representation

$$m_t^\circ(x) = \frac{\int_{[0,1]^d} g^\circ(u)\exp\{-\|x - tg^\circ(u)\|^2/(2q_t)\}\,du}{\int_{[0,1]^d}\exp\{-\|x - tg^\circ(u)\|^2/(2q_t)\}\,du}. \tag{25}$$

For a measurable f, put $v_f(x,t) = a_t x + b_t\,\mathrm{clip}_2(f(x,t))$. Since $\|m_t^\circ(x)\| \le 2$,

$$\|v^\circ(x,t) - v_f(x,t)\|^2 \le b_t^2\big(16 \wedge \|m_t^\circ(x) - f(x,t)\|^2\big).$$

Let $L_\varepsilon = \log(\varepsilon^{-2\beta}/D)$,

$$R_t := \sqrt{q_t D} + 16\sqrt{q_t}\big(\sqrt{DL_\varepsilon} \vee L_\varepsilon\big), \qquad \mathcal{K}_t := \left\{x : \min_{u\in[0,1]^d} \|x - tg^*(u)\| \le R_t\right\},$$

and $\mathcal{C}^*_{[0,1]} := \{(x,t) : x \in \mathcal{K}_t,\ t \in [0,1]\}$. Splitting the integral over $\mathcal{K}_t$ and its complement isolates the compact-domain approximation from the Gaussian tail. Lemma B.2 makes this decomposition quantitative: a uniform value approximation on $\mathcal{C}^*_{[0,1]}$ is sufficient, while the complement contributes at most order $D\varepsilon^{2\beta}/\sigma$ to the velocity risk.

Step 4: Neural realization of the surrogate posterior. On the localized region, the posterior is a ratio of sums over the same latent cells used in Step 1. We first describe the value realization; the spatial constraint is then obtained from first-order realizations of its scalar modules. The remaining problem is uniform approximation of the ratio in (25) on $\mathcal{C}^*_{[0,1]}$. Using the same cells and Taylor polynomials as in (22)–(23), write

$$m_t^\circ(x) = \frac{P^\circ(x,t)}{Q^\circ(x,t)}, \quad P^\circ := \sum_{\mathbf{j}} P^\circ_{\mathbf{j}}, \quad Q^\circ := \sum_{\mathbf{j}} Q^\circ_{\mathbf{j}}, \tag{26}$$

where

$$P^\circ_{\mathbf{j}} := \int_{\mathcal{U}_{\mathbf{j}}} g^\circ_{\mathbf{j}}(u) e^{-\|x - tg^\circ_{\mathbf{j}}(u)\|^2/(2q_t)} du, \qquad Q^\circ_{\mathbf{j}} := \int_{\mathcal{U}_{\mathbf{j}}} e^{-\|x - tg^\circ_{\mathbf{j}}(u)\|^2/(2q_t)} du.$$

For a fixed cell, set $h_{\mathbf{j}}(u) = g^\circ_{\mathbf{j}}(u) - g^*(u_{\mathbf{j}})$. Expanding the Gaussian exponent gives

$$\frac{\|x - tg^\circ_{\mathbf{j}}(u)\|^2}{2q_t} = \frac{\|x - tg^*(u_{\mathbf{j}})\|^2}{2q_t} + \frac{t^2}{2q_t}\|h_{\mathbf{j}}(u)\|^2 - \frac{t}{q_t}(x - tg^*(u_{\mathbf{j}}))^\top h_{\mathbf{j}}(u).$$

After the change of variables from $\mathcal{U}_{\mathbf{j}}$ to the unit cube, $h_{\mathbf{j}}$ is a polynomial in d variables with coefficients given by the derivatives of g^* at $u_{\mathbf{j}}$. Each cell integral therefore depends on (x,t) through finitely many scalar coefficients: the local Taylor coefficients paired with $x - tg^*(u_{\mathbf{j}})$, their quadratic

products, and $1/q_t, t/q_t, t^2/q_t$. This is the low-dimensional exponential-integral map approximated by the network: its nonlinear input dimension depends on (d, β), even though $x \in \mathbb{R}^D$.

The network is assembled in the same order as this formula. Subnetworks first approximate the time coefficients and scalar coordinates, followed by the cellwise exponential integrals. Their outputs are summed over the N^d cells. A uniform lower bound for Q° permits reciprocal and multiplication networks to form P°/Q°, and the D scalar coordinates are parallelized.

Compatibility with the velocity class. The class in Theorem 3.1 also requires a spatial Lipschitz envelope. Gaussian smoothing makes m_t° smooth in x, and first-order realizations of the scalar cell modules preserve the value approximation. Lemma B.3 provides the resulting network, its global spatial bound, and its architecture.

Step 5: Assemble the comparator. Steps 1–4 provide the latent surrogate error, its velocity transfer, the localization reduction, and a spatially regular posterior network. Insert that network into the structured velocity formula and decompose the final error as

$$v^* - v_\theta = (v^* - v^\circ) + (v^\circ - v_\theta).$$

Set $v_\theta(x,t) = a_t x + b_t \,\mathrm{clip}_2(f_\theta(x,t))$. The factor $\sqrt{D}$ in (38) converts the coordinatewise approximation into Euclidean output error; the coefficients a_t, b_t remain exact. Since Euclidean projection is nonexpansive, Lemmas B.2 and B.3 yield

$$\int_0^1 \|v^\circ(\cdot,t) - v_\theta(\cdot,t)\|^2_{L^2(\pi_t)} dt \lesssim_{d,\beta} \frac{D}{\sigma}\varepsilon^{2\beta}. \tag{27}$$

Combining this with (24) gives

$$\int_0^1 \|v^*(\cdot,t) - v_\theta(\cdot,t)\|^2_{L^2(\pi_t)} dt \lesssim_{d,\beta} \frac{D(1+H^2)}{\sigma}\,\varepsilon^{2\beta},$$

which is (7). The construction above has $N^d \asymp_d \varepsilon^{-d}$ parallel cell modules, which determine the algebraic dependence on the approximation scale. Time-coefficient networks, scalar exponential-integral primitives, stable division, the first-order refinement, and the global gate satisfy the explicit Γ_ε bounds in Lemma B.3. Euclidean clipping is nonexpansive, so the same field also obeys

$$\mathrm{Lip}_x(v_\theta(\cdot,t)) \le |a_t| + |b_t|\,\mathrm{Lip}_x(f_\theta(\cdot,t)) \le \gamma_x(t).$$

Taking the fixed constant in the definition of γ_x sufficiently large, depending only on (d,β), makes the right-hand side no larger than $\gamma_x(t)$. The elementary bounds $|a_t|, |b_t| \le 2/((1-t)+\sigma t)$ and $q_t \ge ((1-t)+\sigma t)^2/2$ give the supremum and integral estimates stated in Theorem 3.1; hence the risk, profile, and architecture bounds all hold for this one $v_\theta \in \mathcal{V}(L, W, S, B, \gamma_x)$. Thus the spatially regular comparator retains the same latent-dimensional leading approximation complexity. □

B.3 Latent surrogate and localization

The results below supply the latent approximation, its FM stability transfer, and the reduction from the unbounded state space to $\mathcal{C}^*_{[0,1]}$.

B.3.1 Latent Taylor approximation

Lemma B.1 (Latent Taylor surrogate)**.** *The surrogate defined above satisfies*

$$\|g^* - g^\circ\|_{L^\infty([0,1]^d)} \lesssim_{d,\beta} H\sqrt{D}\,\varepsilon^\beta. \tag{28}$$

Under the scale assumptions of Theorem 3.1, $\|g^\circ\|_{L^\infty([0,1]^d)} \le 2$.

Proof. Fix a cell and write $h = u - u_{\mathbf{j}}$, so $\|h\|_\infty \le N^{-1}$. We first bound one coordinate g_l^*. If $\beta \notin \mathbb{N}$, write $\beta = r_\beta + \alpha$ with $\alpha \in (0,1)$. For $r_\beta = 0$, the defining Hölder condition gives $|g_l^*(u) - g_l^*(u_{\mathbf{j}})| \le H\|h\|_\infty^\beta$. For $r_\beta \ge 1$, apply the one-dimensional integral Taylor formula to $\varphi(s) = g_l^*(u_{\mathbf{j}} + sh)$. After subtracting the order r_β Taylor polynomial, the remainder is

$$\frac{1}{(r_\beta - 1)!}\int_0^1 (1-s)^{r_\beta - 1}\{\varphi^{(r_\beta)}(s) - \varphi^{(r_\beta)}(0)\}\,ds.$$

The coordinatewise Hölder condition and the multinomial formula imply that its absolute value is at most $C_{d,\beta}H\|h\|_\infty^\beta$.

If $\beta = m \in \mathbb{N}$, then $r_\beta = m-1$. The integral remainder after the degree-$(m-1)$ Taylor polynomial is

$$\frac{1}{(m-1)!}\int_0^1 (1-s)^{m-1}\varphi^{(m)}(s)\,ds.$$

All order-m partial derivatives are bounded by H under the stated Hölder-class convention, so the multinomial formula bounds this remainder by $C_{d,m}H\|h\|_\infty^m$. Thus, in both cases,

$$\sup_{u\in\overline{\mathcal{U}_\mathbf{j}}} |g_l^*(u) - g_{\mathbf{j},l}^\circ(u)| \le C_{d,\beta}HN^{-\beta}, \qquad 1 \le l \le D.$$

Converting these coordinatewise bounds once to the Euclidean norm yields

$$\sup_{u\in\overline{\mathcal{U}_\mathbf{j}}} \|g^*(u) - g_\mathbf{j}^\circ(u)\| \le C_{d,\beta}H\sqrt{D}\,N^{-\beta}. \tag{29}$$

The half-open cells are disjoint and cover $[0,1]^d$, while all anchors $u_\mathbf{j}$ belong to this cube. Hence

$$\|g^* - g^\circ\|_{L^\infty([0,1]^d)} \le C_{d,\beta}H\sqrt{D}\,N^{-\beta} \le C_{d,\beta}H\sqrt{D}\,\varepsilon^\beta. \tag{30}$$

The second condition in (5), decreased through $c_{d,\beta}$ if necessary, makes this error at most one because $\varepsilon^\beta \le \varepsilon^{\beta\wedge 1}$. Together with $\|g^*\|_{L^\infty} \le 1$, this gives $\|g^\circ\|_{L^\infty} \le 2$. □

B.3.2 Generator-to-velocity stability

Proof of Proposition 3.2. This proposition connects latent-generator approximation to the integrated FM velocity risk. The score representation used below has an apparent singularity at $t = 0$, while relative entropy supplies the compensating quadratic behavior in time.

The true and surrogate interpolants have the Gaussian-mixture representations

$$X_t \overset{d}{=} tg^*(U) + \sqrt{q_t}\,Z, \qquad X_t^\circ \overset{d}{=} tg^\circ(U) + \sqrt{q_t}\,Z, \qquad Z \sim \mathcal{N}(0, I_D).$$

The lower bound on q_t from Section 2 implies that their densities p_t^* and p_t° are smooth and positive throughout $[0,1]$. Let $K(t) := D_{\mathrm{KL}}(p_t^* \,\|\, p_t^\circ)$. Introduce the joint laws R_t^* and R_t° of (U, X_t) under g^* and g°, respectively. Data processing under the projection $(U, X_t) \mapsto X_t$, followed by conditioning on the common uniform variable U, gives

$$\begin{aligned} K(t) &\le D_{\mathrm{KL}}(\mathsf{R}_t^* \,\|\, \mathsf{R}_t^\circ) \\ &= \mathbb{E}_U D_{\mathrm{KL}}(\mathcal{N}(tg^*(U), q_t I_D) \,\|\, \mathcal{N}(tg^\circ(U), q_t I_D)) \\ &\le \frac{t^2}{2q_t}\|g^* - g^\circ\|^2_{L^\infty([0,1]^d)}. \end{aligned} \tag{31}$$

In particular, $K(0) = 0$ and $K(t) = O(t^2)$ as $t \downarrow 0$.

For $t \in (0,1)$, write $s_t^* = \nabla \log p_t^*$ and $s_t^\circ = \nabla \log p_t^\circ$. Gaussian conditioning and Lemma 2.1 give

$$v^*(x,t) = \frac{x}{t} + \frac{1-t}{t}s_t^*(x), \qquad v^\circ(x,t) = \frac{x}{t} + \frac{1-t}{t}s_t^\circ(x),$$

and hence

$$v^*(x,t) - v^\circ(x,t) = \frac{1-t}{t}\big(s_t^*(x) - s_t^\circ(x)\big). \tag{32}$$

The two marginal paths satisfy their continuity equations with velocities v^* and v°. Bounded generators and the uniform positive lower bound on q_t give smooth positive Gaussian-mixture densities with Gaussian envelopes sufficient for differentiation under the integral and spatial integration by parts. Differentiating $K(t)$ along the two continuity equations first gives

$$K'(t) = \int_{\mathbb{R}^D} p_t^*(x)\big(v^*(x,t) - v^\circ(x,t)\big)\cdot\big(s_t^*(x) - s_t^\circ(x)\big)\,dx.$$

Substituting (32) yields

$$K'(t) = \frac{1-t}{t} \int_{\mathbb{R}^D} \|s_t^*(x) - s_t^\circ(x)\|^2 p_t^*(x)\, dx$$

and therefore

$$\int_{\mathbb{R}^D} \|v^*(x,t) - v^\circ(x,t)\|^2 p_t^*(x)\, dx = \frac{1-t}{t} K'(t).$$

For $\delta \in (0,1)$, integration by parts gives

$$\int_\delta^1 \frac{1-t}{t} K'(t)\, dt = -\frac{1-\delta}{\delta} K(\delta) + \int_\delta^1 \frac{K(t)}{t^2}\, dt.$$

The boundary term vanishes as $\delta \downarrow 0$ because $K(\delta) = O(\delta^2)$. Combining the resulting identity with that bound gives

$$\begin{aligned}\int_0^1 \int_{\mathbb{R}^D} \|v^*(x,t) - v^\circ(x,t)\|^2 p_t^*(x)\, dx\, dt &\le \frac{1}{2}\|g^* - g^\circ\|^2_{L^\infty([0,1]^d)} \int_0^1 \frac{dt}{q_t} \\ &= \frac{\pi}{4\sigma}\|g^* - g^\circ\|^2_{L^\infty([0,1]^d)},\end{aligned}$$

where $\int_0^1 ((1-t)^2 + \sigma^2 t^2)^{-1} dt = \pi/(2\sigma)$. Thus the $O(t^2)$ relative-entropy behavior exactly compensates for the source-endpoint weight. The target endpoint is harmless because the weight $(1-t)/t$ vanishes there. □

B.3.3 Localization of the surrogate posterior

Lemma B.2 (Compact localization of the surrogate posterior)**.** *The set $\mathcal{C}^*_{[0,1]}$ defined above is compact and, for every measurable $f : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$,*

$$\begin{aligned}\int_0^1 \|v^\circ - v_f\|^2_{L^2(\pi_t)} dt \lesssim \frac{D\varepsilon^{2\beta}}{\sigma} \\ + \int_0^1 \int_{\mathcal{K}_t} b_t^2 \|m_t^\circ(x) - f(x,t)\|^2\, d\pi_t(x)\, dt.\end{aligned} \tag{33}$$

Proof. The representation (25) and (30) imply

$$\|g^\circ\|_{L^\infty([0,1]^d)} \le 2. \tag{34}$$

Thus $\|m_t^\circ(x)\| \le 2$ and $m_t^\circ(x) = \mathrm{clip}_2(m_t^\circ(x))$ on $\mathbb{R}^D \times [0,1]$.

Let $f_\theta : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$ be measurable and let v_θ be its structured velocity.

Since the Euclidean projection onto the ball $\mathcal{B}(0,2)$ is nonexpansive and both clipped vectors have norm at most two,

$$\begin{aligned}\|v^\circ(x,t) - v_\theta(x,t)\|^2 &= b_t^2 \left\|\mathrm{clip}_2(m_t^\circ(x)) - \mathrm{clip}_2(f_\theta(x,t))\right\|^2 \\ &\le b_t^2 \left(16 \wedge \|m_t^\circ(x) - f_\theta(x,t)\|^2\right).\end{aligned}$$

Splitting the integral over the set $\mathcal{K}_t$ defined in the main proof and its complement gives

$$\begin{aligned}&\int_{\mathbb{R}^D} \|v^\circ(x,t) - v_\theta(x,t)\|^2 p_t^*(x)\, dx \\ &\quad\le b_t^2 \int_{\mathcal{K}_t} \|m_t^\circ(x) - f_\theta(x,t)\|^2 p_t^*(x)\, dx \\ &\qquad + 16 b_t^2 \int_{\mathbb{R}^D \setminus \mathcal{K}_t} p_t^*(x)\, dx.\end{aligned} \tag{35}$$

The map clip_2 is deterministic post-processing; the ReLU-network complexity is that of the underlying map f_θ.

The second term is controlled by Gaussian concentration. Indeed, the true interpolation satisfies

$$X_t \overset{d}{=} tg^*(U) + \sqrt{q_t}\, Z.$$

Therefore, by the same chi-squared concentration argument as in (Yakovlev & Puchkin, 2025, Lemma A.2), with $\widetilde{\sigma}_t^2$ replaced by q_t and m_t replaced by t, one obtains, whenever $R_t^2 \geq Dq_t$,

$$\int_{\mathbb{R}^D \setminus \mathcal{K}_t} p_t^*(x)\, dx \leq \exp\left\{ -\frac{1}{16}\left(\frac{R_t^2 - Dq_t}{Dq_t} \wedge \frac{\sqrt{R_t^2 - Dq_t}}{\sqrt{q_t}} \right) \right\}. \tag{36}$$

The choice of R_t in the main proof and the assumption $\varepsilon^{-2\beta} > D$ imply from (36) that

$$\int_{\mathbb{R}^D \setminus \mathcal{K}_t} p_t^*(x)\, dx \lesssim D\varepsilon^{2\beta}, \qquad t \in [0,1].$$

Integrating (35) over time therefore yields

$$\begin{aligned} &\int_0^1 \int_{\mathbb{R}^D} \|v^\circ(x,t) - v_\theta(x,t)\|^2 p_t^*(x)\, dx\, dt \\ &\quad \lesssim D\varepsilon^{2\beta} \int_0^1 b_t^2\, dt \\ &\qquad + \int_0^1 \int_{\mathcal{K}_t} b_t^2 \|m_t^\circ(x) - f_\theta(x,t)\|^2 p_t^*(x)\, dx\, dt. \end{aligned}$$

Since

$$\int_0^1 b_t^2\, dt = \int_0^1 \frac{(1-t)^2}{\left((1-t)^2 + t^2\sigma^2\right)^2}\, dt = \frac{\pi}{4\sigma},$$

we further obtain

$$\begin{aligned} &\int_0^1 \int_{\mathbb{R}^D} \|v^\circ(x,t) - v_\theta(x,t)\|^2 p_t^*(x)\, dx\, dt \\ &\quad \lesssim \frac{D\varepsilon^{2\beta}}{\sigma} + \int_0^1 \int_{\mathcal{K}_t} b_t^2 \|m_t^\circ(x) - f_\theta(x,t)\|^2 p_t^*(x)\, dx\, dt. \end{aligned} \tag{37}$$

Since g^* is continuous on the compact set $[0,1]^d$ and $t \mapsto R_t$ is continuous, the map

$$(x,t) \longmapsto \min_{u \in [0,1]^d} \|x - tg^*(u)\| - R_t$$

is continuous. Hence $\mathcal{C}^*_{[0,1]}$ is closed. Moreover, the definition of $\mathcal{K}_t$ and (2) imply a uniform bound on $\|x\|$ over $\mathcal{C}^*_{[0,1]}$. Therefore $\mathcal{C}^*_{[0,1]}$ is compact.

Notice that, since $\sigma > 0$,

$$q_t \geq \frac{\sigma^2}{1+\sigma^2} > 0, \qquad t \in [0,1],$$

so the above construction is uniformly nondegenerate on the entire time interval, including $t = 0$. This allows us to carry out the subsequent neural network approximation directly on $\mathcal{C}^*_{[0,1]}$ without introducing a positive stopping time. This proves the compactness and reduction asserted in Lemma B.2. □

B.4 Neural realization of the surrogate posterior

Lemma B.3 (Neural approximation of the surrogate posterior)**.** *Under the assumptions of Theorem 3.1, there is a D-output ReLU network $f_\theta \in \mathrm{NN}(L, W, S, B)$ such that*

$$\sup_{(x,t) \in \mathcal{C}^*_{[0,1]}} \|f_\theta(x,t) - m_t^\circ(x)\|_2 \lesssim_{d,\beta} \sqrt{D}\, \varepsilon^\beta. \tag{38}$$

For every $t \in [0,1]$, the same network satisfies

$$\mathrm{Lip}_x(f_\theta(\cdot, t)) \leq C_{d,\beta}(1 + \sigma^{-1} + t/q_t). \tag{39}$$

Its architecture satisfies

$$L, \log B \lesssim_{d,\beta} \Gamma_\varepsilon^{c_{\mathrm{arch}}}, \qquad W, S \lesssim_{d,\beta} D^2 \varepsilon^{-d} \Gamma_\varepsilon^{c_{\mathrm{arch}}}. \tag{40}$$

The construction proceeds from the latent-cell value approximation to a first-order realization on an enlarged tube, and then to a global network.

B.4.1 VALUE REALIZATION OF THE LATENT-CELL POSTERIOR

The decomposition (26) separates the dependence on (x,t) from the polynomial variable on each latent cell. We now construct uniform approximations of its numerator and denominator on $\mathcal{C}^*_{[0,1]}$.

Fix $\mathbf{j} \in \{1,\dots,N\}^d$. The exponent expansion derived in the main proof will now be encoded by explicit scalar coordinates.

Retain the Taylor degree $r := r_\beta = \lceil \beta \rceil - 1$ from the main proof and set

$$m_\beta^{\mathrm{Tay}} := \binom{d+r_\beta}{d}, \qquad m_\beta := \binom{d+\lfloor \beta \rfloor}{d}.$$

Thus $m_\beta^{\mathrm{Tay}} \le m_\beta$. The first quantity is the actual number of Taylor monomials, whereas the second is a convenient (d,β)-dependent upper bound used in the fixed-dimensional cell construction. From the definition of the local Taylor polynomial,

$$g_{\mathbf{j}}^\circ(u) - g^*(u_{\mathbf{j}}) = \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} \frac{\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})}{\mathbf{k}!}(u-u_{\mathbf{j}})^{\mathbf{k}}, \qquad u \in \mathcal{U}_{\mathbf{j}}. \tag{41}$$

Define

$$V(t) := \frac{t^2}{2q_t}, \qquad V_{\mathbf{j},0}(x,t) := \frac{\|x - t g^*(u_{\mathbf{j}})\|^2}{2q_t}, \tag{42}$$

and, for $\mathbf{k} \in \mathbb{Z}_+^d$ with $1 \le |\mathbf{k}| \le r$,

$$\begin{aligned} V_{\mathbf{j},\mathbf{k}}(x,t) &:= -\frac{t}{q_t}\,\partial_u^{\mathbf{k}}\left[\big(x - t g^*(u_{\mathbf{j}})\big)^\top g^*(u)\right]\Big|_{u=u_{\mathbf{j}}} \\ &= -\frac{t}{q_t}\big(x - t g^*(u_{\mathbf{j}})\big)^\top \partial^{\mathbf{k}} g^*(u_{\mathbf{j}}). \end{aligned} \tag{43}$$

With this notation,

$$\begin{aligned} \frac{\|x - t g_{\mathbf{j}}^\circ(u)\|^2}{2q_t} &= V_{\mathbf{j},0}(x,t) + V(t)\left\|g_{\mathbf{j}}^\circ(u) - g^*(u_{\mathbf{j}})\right\|^2 \\ &\quad + \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} V_{\mathbf{j},\mathbf{k}}(x,t)\frac{(u-u_{\mathbf{j}})^{\mathbf{k}}}{\mathbf{k}!}. \end{aligned} \tag{44}$$

For each fixed $\mathbf{j}$, collect the functions $V_{\mathbf{j},\mathbf{k}}$ together with V into

$$\mathcal{V}_{\mathbf{j}}(x,t) := \left((V_{\mathbf{j},\mathbf{k}}(x,t))_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}}, V(t) \right)^\top.$$

Since

$$\#\left\{\mathbf{k} \in \mathbb{Z}_+^d : |\mathbf{k}| \le r\right\} = \binom{d+r}{d} = m_\beta^{\mathrm{Tay}},$$

the vector $\mathcal{V}_{\mathbf{j}}$ has dimension

$$\dim(\mathcal{V}_{\mathbf{j}}) = \binom{d+r}{d} = m_\beta^{\mathrm{Tay}}.$$

For each fixed $\mathbf{j}$, the representation (44) shows that the dependence of the $\mathbf{j}$-th cell integrals on (x,t) is entirely captured by $V_{\mathbf{j},0}(x,t)$ and $\mathcal{V}_{\mathbf{j}}(x,t)$. Indeed, after factoring out $\exp\{-V_{\mathbf{j},0}(x,t)\}$, the remaining integral is a function of $\mathcal{V}_{\mathbf{j}}(x,t)$ only.

Since

$$\#\left\{\mathbf{k}\in\mathbb{Z}_+^d : 1\le|\mathbf{k}|\le r\right\} = \binom{d+r}{d} - 1 = m_\beta^{\mathrm{Tay}} - 1,$$

the vector $\mathcal{V}_\mathbf{j}$ has dimension

$$\binom{d+r}{d} = m_\beta^{\mathrm{Tay}},$$

and hence each cell integral is a composition involving only $m_\beta^{\mathrm{Tay}}+1 \le m_\beta+1$ scalar intermediate variables, including $V_{\mathbf{j},0}$. This dimension depends only on (d,β) and not on the ambient input dimension D.

The required neural components are $V_{\mathbf{j},0}$, $V_{\mathbf{j},\mathbf{k}}$, and V.

Scalar coordinate networks. We now approximate the functions $V_{\mathbf{j},0}(x,t)$, $V_{\mathbf{j},\mathbf{k}}(x,t)$, $1\le|\mathbf{k}|\le r$, and $V(t)$ introduced in (42)–(43). Throughout this step, let

$$M := 1 \vee \sup_{(x,t)\in\mathcal{C}^*_{[0,1]}} \|x\|_\infty.$$

By the construction of $\mathcal{C}^*_{[0,1]}$ above,

$$M \le 1 \vee \sup_{t\in[0,1]} (R_t+1),$$

and hence $M<\infty$. We first consider $V_{\mathbf{j},0}$. Since $g^\circ_\mathbf{j}(u_\mathbf{j}) = g^*(u_\mathbf{j})$, we have

$$\begin{aligned} V_{\mathbf{j},0}(x,t) &= \frac{\|x - t g^*(u_\mathbf{j})\|^2}{2q_t} \\ &= \frac{\|x\|^2}{2q_t} - \frac{t\,x^\top g^*(u_\mathbf{j})}{q_t} + \frac{\|g^*(u_\mathbf{j})\|^2}{2}\frac{t^2}{q_t}. \end{aligned}$$

The three terms on the right-hand side are precisely of the forms treated in Lemmas B.21, B.22, and B.20, respectively. Since $\|g^*\|_{L^\infty([0,1]^d)} \le 1$, applying these lemmas with accuracy $\varepsilon'/3$ yields the ReLU network

$$\widetilde{V}_{\mathbf{j},0}(x,t) := \rho_{\varepsilon'/3}(x,t) - \omega_{\varepsilon'/3}(x,t) + \frac{\|g^*(u_\mathbf{j})\|^2}{2}\chi_{2,\varepsilon'/3}(t),$$

where Lemma B.22 is used with $a = g^*(u_\mathbf{j})$. Consequently,

$$\left\|\widetilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \le \varepsilon'.$$

We next consider $V_{\mathbf{j},\mathbf{k}}$, $1\le|\mathbf{k}|\le r$. By (43),

$$\begin{aligned} V_{\mathbf{j},\mathbf{k}}(x,t) &= -\frac{t}{q_t}\big(x - t g^*(u_\mathbf{j})\big)^\top \partial^\mathbf{k} g^*(u_\mathbf{j}) \\ &= -\frac{t\,x^\top \partial^\mathbf{k} g^*(u_\mathbf{j})}{q_t} + g^*(u_\mathbf{j})^\top \partial^\mathbf{k} g^*(u_\mathbf{j})\frac{t^2}{q_t}. \end{aligned}$$

The first term is approximated by Lemma B.22 with $a = \partial^\mathbf{k} g^*(u_\mathbf{j})$. For the second term, set

$$\delta_{\mathbf{j},\mathbf{k}} := \frac{\varepsilon'}{2\big(|g^*(u_\mathbf{j})^\top \partial^\mathbf{k} g^*(u_\mathbf{j})| \vee 1\big)}.$$

Using Lemma B.22 with accuracy $\varepsilon'/2$ and Lemma B.20 with $\gamma=2$ and accuracy $\delta_{\mathbf{j},\mathbf{k}}$, define

$$\begin{aligned} \widetilde{V}_{\mathbf{j},\mathbf{k}}(x,t) := &-\omega_{\varepsilon'/2}(x,t) \\ &+ g^*(u_\mathbf{j})^\top \partial^\mathbf{k} g^*(u_\mathbf{j})\,\chi_{2,\delta_{\mathbf{j},\mathbf{k}}}(t), \end{aligned}$$

where $\omega_{\varepsilon'/2}$ is constructed with $a = \partial^\mathbf{k} g^*(u_\mathbf{j})$. Then

$$\left\|\widetilde{V}_{\mathbf{j},\mathbf{k}} - V_{\mathbf{j},\mathbf{k}}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \le \varepsilon', \qquad 1\le|\mathbf{k}|\le r.$$

Finally, $V(t) = t^2/(2q_t)$ is treated directly by Lemma B.20. Namely, setting

$$\widetilde{V}(t) := \frac{1}{2}\chi_{2,\varepsilon'}(t),$$

we have

$$\left\|\widetilde{V} - V\right\|_{L^\infty([0,1])} \le \varepsilon'.$$

We now collect the corresponding network configurations. Since $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$,

$$\|\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})\|_\infty \le H, \qquad 1 \le |\mathbf{k}| \le r,$$

and

$$\left|g^*(u_{\mathbf{j}})^\top \partial^{\mathbf{k}} g^*(u_{\mathbf{j}})\right| \lesssim \sqrt{D}\, H.$$

Thus, uniformly over $\mathbf{j} \in \{1, \dots, N\}^d$ and $1 \le |\mathbf{k}| \le r$, all the networks constructed above may be embedded into a common class

$$\mathrm{NN}\big(\widetilde{L}, \widetilde{\mathbf{W}}, \widetilde{S}, \widetilde{B}\big),$$

with

$$\begin{aligned}
\widetilde{L} \vee \log \widetilde{B} &\lesssim_{d,\beta} \log^2(1/\varepsilon') + \log^2\big(MD(H \vee 1)\big) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right), \\
\|\widetilde{\mathbf{W}}\|_\infty &\lesssim_{d,\beta} D\left[\log^3(1/\varepsilon') + \log^3\big(MD(H \vee 1)\big) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right)\right], \\
\widetilde{S} &\lesssim_{d,\beta} D\left[\log^4(1/\varepsilon') + \log^4\big(MD(H \vee 1)\big) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right)\right].
\end{aligned} \tag{45}$$

The dependence on H remains explicit in this logarithmic factor and is retained in the final envelope Γ_ε.

Then, uniformly over $\mathbf{j} \in \{1, \dots, N\}^d$,

$$\max\left\{ \|\widetilde{V} - V\|_{L^\infty([0,1])},\ \max_{\substack{\mathbf{k} \in \mathbb{Z}_+^d \\ 1 \le |\mathbf{k}| \le r}} \|\widetilde{V}_{\mathbf{j},\mathbf{k}} - V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \right\} \le \varepsilon',$$

and

$$\|\widetilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \le \varepsilon'.$$

Low-dimensional cell integrals. We use the scalar coordinate networks above to approximate the local numerator and denominator integrals in (26). Fix $\mathbf{j} \in \{1, \dots, N\}^d$ and consider the change of variables

$$u = u_{\mathbf{j}} - \frac{w}{N}, \qquad w \in [0,1]^d.$$

By the definition of the partition in (22), this map sends $[0,1]^d$ onto $\overline{\mathcal{U}_{\mathbf{j}}}$ and satisfies $du = N^{-d} dw$. Since $\overline{\mathcal{U}_{\mathbf{j}}} \setminus \mathcal{U}_{\mathbf{j}}$ has Lebesgue measure zero, for every integrable function F,

$$N^d \int_{\mathcal{U}_{\mathbf{j}}} F(u)\, du = \int_{[0,1]^d} F\left(u_{\mathbf{j}} - \frac{w}{N}\right) dw. \tag{46}$$

By (44),

$$\begin{aligned}
&\frac{\|x - t g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - N^{-1}w)\|^2}{2q_t} \\
&\quad = V_{\mathbf{j},0}(x,t) + V(t)\left\|g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - N^{-1}w) - g^*(u_{\mathbf{j}})\right\|^2 \\
&\qquad + \sum_{\substack{\mathbf{k} \in \mathbb{Z}_+^d \\ 1 \le |\mathbf{k}| \le r}} V_{\mathbf{j},\mathbf{k}}(x,t) \frac{(-N^{-1}w)^{\mathbf{k}}}{\mathbf{k}!}.
\end{aligned}$$

When $r = 0$, all families and sums indexed by $\mathbf{k} \in \mathbb{Z}_+^d$ with $1 \le |\mathbf{k}| \le r$ are understood to be absent. In this case, the constructions below contain only the coordinate corresponding to $V(t)$.

For every $\mathbf{k} \in \mathbb{Z}_+^d$ with $1 \le |\mathbf{k}| \le r$, set

$$C_{\mathbf{j},\mathbf{k}} := 1 \vee \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[0,1]})},$$

and define

$$R_{\mathbf{j}}(x,t) := \left(\left(\frac{V_{\mathbf{j},\mathbf{k}}(x,t)}{2C_{\mathbf{j},\mathbf{k}}} + \frac{1}{2} \right)_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}}, \frac{V(t)}{\|V\|_{L^\infty([0,1])}} \right)^\top. \tag{47}$$

By construction,

$$R_{\mathbf{j}}(x,t) \in [0,1]^{\binom{d+r}{d}}.$$

For $w \in [0,1]^d$, define

$$a_{\mathbf{j}}(w) := \left(\left(\frac{2C_{\mathbf{j},\mathbf{k}}(-N^{-1}w)^{\mathbf{k}}}{\mathbf{k}!} \right)_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}}, \|V\|_{L^\infty([0,1])} \left\| g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - N^{-1}w) - g^*(u_{\mathbf{j}}) \right\|^2 \right)^\top \tag{48}$$

and

$$b_{\mathbf{j}}(w) := -\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} C_{\mathbf{j},\mathbf{k}} \frac{(-N^{-1}w)^{\mathbf{k}}}{\mathbf{k}!}. \tag{49}$$

Then the decomposition is algebraically exact:

$$\frac{\|x - tg^\circ_{\mathbf{j}}(u_{\mathbf{j}} - N^{-1}w)\|^2}{2q_t} = V_{\mathbf{j},0}(x,t) + R_{\mathbf{j}}(x,t)^\top a_{\mathbf{j}}(w) + b_{\mathbf{j}}(w).$$

Auxiliary estimates for the local-integral construction. The following results are used in the proof of Lemma B.5. The first result controls the normalized intermediate coordinates, the second is the low-dimensional integral approximation result of Yakovlev & Puchkin (2025), and the third provides the FM-specific bounds required to verify its assumptions.

Normalization bounds. Let $F, \widetilde{F} : \Omega \to \mathbb{R}$ satisfy

$$\|F\|_{L^\infty(\Omega)} \le C, \qquad C \ge 1, \qquad \|\widetilde{F} - F\|_{L^\infty(\Omega)} \le \eta.$$

For $z \in \mathbb{R}$, let

$$\mathrm{clip}_{[0,1]}(z) := \min\{1, \max\{0, z\}\}.$$

Define

$$R := \frac{F}{2C} + \frac{1}{2}, \qquad \widetilde{R} := \mathrm{clip}_{[0,1]}\left(\frac{\widetilde{F}}{2C} + \frac{1}{2} \right).$$

Then $R, \widetilde{R} \in [0,1]$ and

$$\|\widetilde{R} - R\|_{L^\infty(\Omega)} \le \frac{\eta}{2C}.$$

If, in addition, $F \ge 0$ and $\|F\|_{L^\infty(\Omega)} > 0$, then

$$R_+ := \frac{F}{\|F\|_{L^\infty(\Omega)}}, \qquad \widetilde{R}_+ := \mathrm{clip}_{[0,1]}\left(\frac{\widetilde{F}}{\|F\|_{L^\infty(\Omega)}} \right)$$

satisfy

$$\|\widetilde{R}_+ - R_+\|_{L^\infty(\Omega)} \le \frac{\|\widetilde{F} - F\|_{L^\infty(\Omega)}}{\|F\|_{L^\infty(\Omega)}}.$$

Proof. The map $\mathrm{clip}_{[0,1]}$ is one-Lipschitz and equals the identity on $[0,1]$. Since

$$\frac{F}{2C} + \frac{1}{2} \in [0,1],$$

we obtain

$$\begin{aligned} |\widetilde{R} - R| &= \left| \mathrm{clip}_{[0,1]} \left(\frac{\widetilde{F}}{2C} + \frac{1}{2} \right) - \mathrm{clip}_{[0,1]} \left(\frac{F}{2C} + \frac{1}{2} \right) \right| \\ &\leq \frac{|\widetilde{F} - F|}{2C}. \end{aligned}$$

Taking the supremum over Ω proves the first claim.

If $F \geq 0$, then

$$\frac{F}{\|F\|_{L^\infty(\Omega)}} \in [0,1].$$

The same one-Lipschitz argument gives

$$\left| \widetilde{R}_+ - R_+ \right| \leq \frac{|\widetilde{F} - F|}{\|F\|_{L^\infty(\Omega)}},$$

which proves the second claim. □

Lemma B.4 (Low-dimensional exponential-integral approximation; Lemma A.6 of Yakovlev & Puchkin (2025))**.** *Let $m \in \mathbb{N}$ and let*

$$\varphi : [0,1]^d \to \mathbb{R}, \qquad a : [0,1]^d \to \mathbb{R}^m, \qquad b : [0,1]^d \to \mathbb{R}.$$

Consider

$$\Phi(v) := \int_{[0,1]^d} \varphi(w) \exp\{-v^\top a(w) - b(w)\}\, dw, \qquad v \in [0,1]^m.$$

Suppose that, for some $\varphi_{\max} \geq 1$, $a_{\max} \geq 1$, and $A \geq 0$,

$$|\varphi(w)| \leq \varphi_{\max}, \qquad \|a(w)\|_\infty \leq a_{\max},$$

and

$$-v^\top a(w) - b(w) \leq A$$

for every $v \in [0,1]^m$ and $w \in [0,1]^d$. Then, for every $\delta \in (0,1)$, there exists a ReLU network

$$\widetilde{\Phi} \in \mathrm{NN}(L_\Phi, \mathbf{W}_\Phi, S_\Phi, 1)$$

such that

$$\|\widetilde{\Phi} - \Phi\|_{L^\infty([0,1]^m)} \leq \delta.$$

Moreover,

$$L_\Phi \lesssim_{\varphi_{\max}, a_{\max}, A} \left(m + 1 + \log \frac{1}{\delta} \right) \log \left(e + m + \log \frac{1}{\delta} \right) \log \left(e + \log \frac{1}{\delta} \right),$$

$$\|\mathbf{W}_\Phi\|_\infty \lesssim_{\varphi_{\max}, a_{\max}, A} \left(m + 1 + \log \frac{1}{\delta} \right)^{m+1}, \qquad S_\Phi \lesssim_{\varphi_{\max}, a_{\max}, A} \left(m + 1 + \log \frac{1}{\delta} \right)^{2m+5}.$$

The hidden constants depend only on $\varphi_{\max}$, $a_{\max}$, and A.

Bounds for the FM intermediate coefficients. With the notation introduced in Steps 3–5, uniformly over $\mathbf{j} \in \{1, \ldots, N\}^d$,

$$\|V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim \frac{1+\sigma^2}{\sigma^2} D(M+1)^2,$$

$$\|V\|_{L^\infty([0,1])} = \frac{1}{2\sigma^2},$$

and, for every $\mathbf{k} \in \mathbb{Z}_+^d$ with $1 \le |\mathbf{k}| \le r$,

$$\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim \frac{1+\sigma^2}{\sigma^2} DH(M+1).$$

Furthermore,

$$\sup_{w\in[0,1]^d} \left\| g_{\mathbf{j}}^{\circ}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}}) \right\| \lesssim \sqrt{D}\, H\left(e^{d/N} - 1\right). \tag{50}$$

Consequently, if (54) holds, then

$$\sup_{w\in[0,1]^d} \|a_{\mathbf{j}}(w)\|_\infty \le 1 \tag{51}$$

and

$$\sup_{\substack{v\in[0,1]^{\binom{d+r}{d}} \\ w\in[0,1]^d}} \left\{-v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)\right\} \le 1, \tag{52}$$

provided the constant c_0 in (54) is chosen sufficiently small.

Proof. We first bound $V_{\mathbf{j},0}$. By (42),

$$V_{\mathbf{j},0}(x,t) = \frac{\|x - tg^*(u_{\mathbf{j}})\|^2}{2q_t}.$$

On $\mathcal{C}^*_{[0,1]}$, $\|x\|_2 \le \sqrt{D}\, M$, while $\|g^*(u_{\mathbf{j}})\|_2 \le 1$. Since

$$q_t \ge \frac{\sigma^2}{1+\sigma^2},$$

we have

$$\begin{aligned} V_{\mathbf{j},0}(x,t) &\le \frac{1+\sigma^2}{2\sigma^2}\left(\sqrt{D}\, M + 1\right)^2 \\ &\lesssim \frac{1+\sigma^2}{\sigma^2} D(M+1)^2. \end{aligned}$$

Next,

$$V(t) = \frac{t^2}{2q_t}.$$

For $t > 0$,

$$\frac{t^2}{q_t} = \frac{1}{((1-t)/t)^2 + \sigma^2} \le \frac{1}{\sigma^2},$$

and equality is attained at $t = 1$. Since $V(0) = 0$,

$$\|V\|_{L^\infty([0,1])} = \frac{1}{2\sigma^2}.$$

For $\mathbf{k} \in \mathbb{Z}_+^d$ with $1 \le |\mathbf{k}| \le r$, (43) gives

$$\begin{aligned} |V_{\mathbf{j},\mathbf{k}}(x,t)| &\le \frac{t}{q_t}\|x - tg^*(u_{\mathbf{j}})\|_2 \|\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})\|_2 \\ &\le \frac{1+\sigma^2}{\sigma^2}(\sqrt{D}\, M + 1)\sqrt{D}\, H \\ &\lesssim \frac{1+\sigma^2}{\sigma^2} DH(M+1). \end{aligned}$$

We next control the local Taylor increment. By (41),

$$g_{\mathbf{j}}^{\circ}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}}) = \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} \frac{\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})}{\mathbf{k}!}\left(-\frac{w}{N}\right)^{\mathbf{k}}.$$

Since

$$\|\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})\|_2 \le \sqrt{D}\, H,$$

the multinomial theorem yields

$$\begin{aligned}
&\left\| g_{\mathbf{j}}^{\circ}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}}) \right\| \\
&\quad \le \sqrt{D}\, H \sum_{m=1}^{r} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ |\mathbf{k}|=m}} \frac{N^{-m} w^{\mathbf{k}}}{\mathbf{k}!} \\
&\quad = \sqrt{D}\, H \sum_{m=1}^{r} \frac{(N^{-1}\|w\|_1)^m}{m!} \\
&\quad \le \sqrt{D}\, H \left(e^{d/N} - 1\right),
\end{aligned}$$

which proves (50).

We now verify the two bounds required by Lemma B.4. For every Taylor coordinate of $a_{\mathbf{j}}$,

$$\left| \frac{2C_{\mathbf{j},\mathbf{k}}(-N^{-1}w)^{\mathbf{k}}}{\mathbf{k}!} \right| \le \frac{2C_{\mathbf{j},\mathbf{k}}}{N^{|\mathbf{k}|}} \le \frac{2C_{\mathbf{j},\mathbf{k}}}{N}.$$

By the bound on $V_{\mathbf{j},\mathbf{k}}$ above,

$$\frac{C_{\mathbf{j},\mathbf{k}}}{N} \lesssim \frac{1}{N} + \frac{1}{N}\frac{1+\sigma^2}{\sigma^2} D(H \vee 1)(M+1).$$

Hence (54) implies, after reducing c_0 if necessary,

$$\sup_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} \sup_{w\in[0,1]^d} \left| \frac{2C_{\mathbf{j},\mathbf{k}}(-N^{-1}w)^{\mathbf{k}}}{\mathbf{k}!} \right| \le 1.$$

For the last coordinate of $a_{\mathbf{j}}$,

$$\begin{aligned}
&\|V\|_{L^\infty([0,1])} \left\| g_{\mathbf{j}}^{\circ}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}}) \right\|^2 \\
&\quad \le \frac{DH^2}{2\sigma^2} \left(e^{d/N} - 1\right)^2 \le 1,
\end{aligned}$$

again under (54), after decreasing c_0 if necessary. This proves (51).

Finally, write

$$v = \left((v_{\mathbf{k}})_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}}, v_V \right) \in [0,1]^{\binom{d+r}{d}}.$$

Using (48) and (49),

$$\begin{aligned}
&- v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w) \\
&= \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} (1 - 2v_{\mathbf{k}}) C_{\mathbf{j},\mathbf{k}} \frac{(-N^{-1}w)^{\mathbf{k}}}{\mathbf{k}!} \\
&\qquad - v_V \|V\|_{L^\infty([0,1])} \left\| g_{\mathbf{j}}^{\circ}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}}) \right\|^2.
\end{aligned}$$

The last term is non-positive. Therefore

$$\begin{aligned}
-v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w) &\le \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} C_{\mathbf{j},\mathbf{k}} \sum_{m=1}^{r} \frac{(d/N)^m}{m!} \\
&\le \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\le|\mathbf{k}|\le r}} C_{\mathbf{j},\mathbf{k}} \left(e^{d/N} - 1\right).
\end{aligned}$$

Since

$$e^{d/N} - 1 \lesssim_d \frac{1}{N},$$

and

$$C_{\mathbf{j},\mathbf{k}} = 1 \vee \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim 1 + \frac{1+\sigma^2}{\sigma^2} DH(M+1),$$

the small-cell condition (54) implies, after reducing c_0 if necessary, that

$$\max_{\substack{\mathbf{k}\in\mathbb{Z}^d_+ \\ 1\le|\mathbf{k}|\le r}} C_{\mathbf{j},\mathbf{k}} \left(e^{d/N} - 1\right) \le 1.$$

Therefore,

$$\sup_{\substack{v\in[0,1]^{\binom{d+r}{d}} \\ w\in[0,1]^d}} \left\{-v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)\right\} \le 1,$$

which proves (52). □

The following lemma combines these estimates into the neural approximation of a normalized local integral.

Lemma B.5 (Approximation of a normalized local FM integral)**.** *Let $\varepsilon' \in (0,1)$ and fix $\mathbf{j} \in \{1,\dots,N\}^d$. Let $\psi_{\mathbf{j}} : [0,1]^d \to \mathbb{R}$ satisfy*

$$\|\psi_{\mathbf{j}}\|_{L^\infty([0,1]^d)} \le 2,$$

and define

$$\Upsilon_{\mathbf{j}}(x,t) := e^{-V_{\mathbf{j},0}(x,t)} \int_{[0,1]^d} \psi_{\mathbf{j}}(w) \exp\left\{-R_{\mathbf{j}}(x,t)^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)\right\} dw. \tag{53}$$

Assume that

$$\frac{1}{N}\left[1 + \frac{1+\sigma^2}{\sigma^2} D(H\vee 1)(M+1)\right] \le c_0, \tag{54}$$

where $c_0 > 0$ is sufficiently small and depends only on d and β. Then there exists a ReLU network

$$\widetilde{\Upsilon}_{\mathbf{j}} \in \mathrm{NN}(L_\Upsilon, \mathbf{W}_\Upsilon, S_\Upsilon, B_\Upsilon)$$

such that

$$\left\|\widetilde{\Upsilon}_{\mathbf{j}} - \Upsilon_{\mathbf{j}}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \frac{\varepsilon'}{N}. \tag{55}$$

Moreover,

$$\begin{aligned}
L_\Upsilon &\lesssim_{d,\beta} \log^2\left(\frac{N}{\varepsilon'}\right) + \log^2\big(MD(H\vee 1)\big) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right) \\
&\quad + \left(\log N + \log\frac{1}{\varepsilon'}\right)\log^2\left(e + \log N + \log\frac{1}{\varepsilon'}\right), \\
\|\mathbf{W}_\Upsilon\|_\infty &\lesssim_{d,\beta} D\left[\log^3\left(\frac{N}{\varepsilon'}\right) + \log^3\big(MD(H\vee 1)\big) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right)\right] \\
&\quad \vee \left(1 + \log N + \log\frac{1}{\varepsilon'}\right)^{\binom{d+r}{d}+1}, \\
S_\Upsilon &\lesssim_{d,\beta} D\left[\log^4\left(\frac{N}{\varepsilon'}\right) + \log^4\big(MD(H\vee 1)\big) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right)\right] \\
&\quad + \left(1 + \log N + \log\frac{1}{\varepsilon'}\right)^{2\binom{d+r}{d}+5}, \\
\log B_\Upsilon &\lesssim_{d,\beta} \log^2\left(\frac{N}{\varepsilon'}\right) + \log^2\big(MD(H\vee 1)\big) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right).
\end{aligned} \tag{56}$$

The hidden constants depend only on d and β.

Proof. Choose the scalar-coordinate networks with the refined accuracy allocation

$$\begin{aligned} \|\widetilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} &\le \frac{\varepsilon'}{N}, \\ \|\widetilde{V} - V\|_{L^\infty([0,1])} &\le \frac{\varepsilon'}{N}, \end{aligned} \tag{57}$$

while, for every $\mathbf{k} \in \mathbb{Z}^d_+$ with $1 \le |\mathbf{k}| \le r$,

$$\|\widetilde{V}_{\mathbf{j},\mathbf{k}} - V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \le \varepsilon'. \tag{58}$$

Since these constructions admit arbitrary prescribed accuracies, the common configuration (45) remains valid after replacing $\log(1/\varepsilon')$ by $\log(N/\varepsilon')$.

Normalized-coordinate approximation.

For $\mathbf{k} \in \mathbb{Z}^d_+$ with $1 \le |\mathbf{k}| \le r$, define

$$\widetilde{R}_{\mathbf{j},\mathbf{k}}(x,t) := \mathrm{clip}_{[0,1]}\left(\frac{\widetilde{V}_{\mathbf{j},\mathbf{k}}(x,t)}{2C_{\mathbf{j},\mathbf{k}}} + \frac{1}{2}\right),$$

and

$$\widetilde{R}_{\mathbf{j},V}(t) := \mathrm{clip}_{[0,1]}\left(\frac{\widetilde{V}(t)}{\|V\|_{L^\infty([0,1])}}\right).$$

Let $\widetilde{R}_{\mathbf{j}}$ denote the vector obtained by concatenating these coordinates in the same order as in (47). By construction,

$$\widetilde{R}_{\mathbf{j}}(x,t) \in [0,1]^{\binom{d+r}{d}}.$$

The normalization bounds above together with (57) and (58) yields

$$\left|\widetilde{R}_{\mathbf{j},\mathbf{k}} - R_{\mathbf{j},\mathbf{k}}\right| \le \frac{\varepsilon'}{2C_{\mathbf{j},\mathbf{k}}}, \qquad 1 \le |\mathbf{k}| \le r,$$

and

$$\left|\widetilde{R}_{\mathbf{j},V} - R_{\mathbf{j},V}\right| \le \frac{\varepsilon'/N}{\|V\|_{L^\infty([0,1])}}.$$

Hence, by the definition of $a_{\mathbf{j}}$,

$$\begin{aligned} &\left|a_{\mathbf{j}}(w)^\top\left(\widetilde{R}_{\mathbf{j}}(x,t) - R_{\mathbf{j}}(x,t)\right)\right| \\ &\quad\le \varepsilon' \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d_+ \\ 1\le|\mathbf{k}|\le r}} \frac{|(-N^{-1}w)^{\mathbf{k}}|}{\mathbf{k}!} \\ &\qquad + \frac{\varepsilon'}{N}\left\|g^\circ_{\mathbf{j}}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}})\right\|^2. \end{aligned}$$

By the multinomial theorem,

$$\begin{aligned} \sum_{\substack{\mathbf{k}\in\mathbb{Z}^d_+ \\ 1\le|\mathbf{k}|\le r}} \frac{|(-N^{-1}w)^{\mathbf{k}}|}{\mathbf{k}!} &\le e^{d/N} - 1 \\ &\lesssim_d \frac{1}{N}. \end{aligned}$$

Moreover, by (34) and (2),

$$\left\|g^\circ_{\mathbf{j}}\left(u_{\mathbf{j}} - \frac{w}{N}\right) - g^*(u_{\mathbf{j}})\right\| \le \|g^\circ\|_{L^\infty([0,1]^d)} + \|g^*\|_{L^\infty([0,1]^d)} \le 3.$$

Consequently,

$$\sup_{w\in[0,1]^d}\left|a_{\mathbf{j}}(w)^\top\left(\widetilde{R}_{\mathbf{j}}-R_{\mathbf{j}}\right)\right|\lesssim_{d,\beta}\frac{\varepsilon'}{N}. \tag{59}$$

Low-dimensional integral map.

Define

$$\Phi_{\mathbf{j}}(v):=\int_{[0,1]^d}\psi_{\mathbf{j}}(w)\exp\left\{-v^\top a_{\mathbf{j}}(w)-b_{\mathbf{j}}(w)\right\}\,dw,\qquad v\in[0,1]^{\binom{d+r}{d}}.$$

Then

$$\Psi_{\mathbf{j}}(x,t):=\int_{[0,1]^d}\psi_{\mathbf{j}}(w)\exp\left\{-R_{\mathbf{j}}(x,t)^\top a_{\mathbf{j}}(w)-b_{\mathbf{j}}(w)\right\}\,dw=\Phi_{\mathbf{j}}(R_{\mathbf{j}}(x,t)).$$

By the intermediate-coefficient bounds above, the assumptions of Lemma B.4 hold with $\varphi_{\max}=2$, $a_{\max}=1$, and $A=1$. Applying that lemma with

$$\delta=\frac{\varepsilon'}{N}$$

gives a ReLU network $\widetilde{\Phi}_{\mathbf{j}}$ satisfying

$$\left\|\widetilde{\Phi}_{\mathbf{j}}-\Phi_{\mathbf{j}}\right\|_{L^\infty([0,1]^{\binom{d+r}{d}})}\le\frac{\varepsilon'}{N}. \tag{60}$$

Since the input dimension of $\Phi_{\mathbf{j}}$ is $\binom{d+r}{d}$, its configuration satisfies

$$L_\Phi\lesssim_{d,\beta}\left(1+\log N+\log\frac{1}{\varepsilon'}\right)\log^2\left(e+\log N+\log\frac{1}{\varepsilon'}\right),$$

$$\|\mathbf{W}_\Phi\|_\infty\lesssim_{d,\beta}\left(1+\log N+\log\frac{1}{\varepsilon'}\right)^{\binom{d+r}{d}+1},$$

and

$$S_\Phi\lesssim_{d,\beta}\left(1+\log N+\log\frac{1}{\varepsilon'}\right)^{2\binom{d+r}{d}+5},$$

where the dependence on d and β is absorbed into the hidden constants.

Composition with the normalized coordinates.

Set

$$\widetilde{\Psi}_{\mathbf{j}}(x,t):=\widetilde{\Phi}_{\mathbf{j}}\left(\widetilde{R}_{\mathbf{j}}(x,t)\right).$$

Since both $R_{\mathbf{j}}(x,t)$ and $\widetilde{R}_{\mathbf{j}}(x,t)$ belong to $[0,1]^{\binom{d+r}{d}}$,

$$\begin{aligned}&\left|\widetilde{\Psi}_{\mathbf{j}}(x,t)-\Psi_{\mathbf{j}}(x,t)\right|\\&\quad\le\left|\widetilde{\Phi}_{\mathbf{j}}(\widetilde{R}_{\mathbf{j}})-\Phi_{\mathbf{j}}(\widetilde{R}_{\mathbf{j}})\right|+\left|\Phi_{\mathbf{j}}(\widetilde{R}_{\mathbf{j}})-\Phi_{\mathbf{j}}(R_{\mathbf{j}})\right|.\end{aligned}$$

The first term is bounded by ε'/N by (60).

For the second term, differentiating under the integral gives

$$\nabla\Phi_{\mathbf{j}}(v)=-\int_{[0,1]^d}\psi_{\mathbf{j}}(w)a_{\mathbf{j}}(w)\exp\left\{-v^\top a_{\mathbf{j}}(w)-b_{\mathbf{j}}(w)\right\}\,dw.$$

By the Newton–Leibniz formula,

$$\begin{aligned}&\left|\Phi_{\mathbf{j}}(\widetilde{R}_{\mathbf{j}})-\Phi_{\mathbf{j}}(R_{\mathbf{j}})\right|\\&\quad\le 2\sup_{\substack{v\in[0,1]^{\binom{d+r}{d}}\\w\in[0,1]^d}}\exp\left\{-v^\top a_{\mathbf{j}}(w)-b_{\mathbf{j}}(w)\right\}\\&\qquad\times\sup_{w\in[0,1]^d}\left|a_{\mathbf{j}}(w)^\top\left(\widetilde{R}_{\mathbf{j}}-R_{\mathbf{j}}\right)\right|.\end{aligned} \tag{61}$$

Using (52) and (59),

$$\left|\Phi_{\mathbf{j}}(\widetilde{R}_{\mathbf{j}}) - \Phi_{\mathbf{j}}(R_{\mathbf{j}})\right| \lesssim_{d,\beta} \frac{\varepsilon'}{N}.$$

Consequently,

$$\left\|\widetilde{\Psi}_{\mathbf{j}} - \Psi_{\mathbf{j}}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \frac{\varepsilon'}{N}. \tag{62}$$

The network $\widetilde{R}_{\mathbf{j}}$ is obtained by parallelizing a fixed number $\binom{d+r}{d}$ of the coordinate subnetworks and appending exact ReLU clipping maps. Hence its depth is of the same order as their depth, while its width and sparsity increase only by a factor depending on d and β. Composing it with $\widetilde{\Phi}_{\mathbf{j}}$ therefore gives the first two complexity blocks in (56).

Outer exponential and final product.

By construction,

$$V_{\mathbf{j},0}(x,t) \geq 0.$$

For the local-integral construction, the coordinate networks are instantiated at accuracy ε'/N, and hence (57) gives

$$\left\|\widetilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \leq \frac{\varepsilon'}{N}.$$

Therefore

$$\widetilde{V}_{\mathbf{j},0}(x,t) \geq -\frac{\varepsilon'}{N}.$$

Applying Corollary B.19 with

$$a = \frac{\varepsilon'}{N}$$

and exponential approximation accuracy ε'/N, we obtain a ReLU network $\widetilde{E}_{\mathbf{j},0}$ satisfying

$$\left\|\widetilde{E}_{\mathbf{j},0} - e^{-V_{\mathbf{j},0}}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim \frac{\varepsilon'}{N}.$$

Moreover,

$$L_{\exp} \lesssim \log^2\left(\frac{N}{\varepsilon'}\right), \qquad \|\mathbf{W}_{\exp}\|_\infty \lesssim \log\left(\frac{N}{\varepsilon'}\right),$$
$$S_{\exp} \lesssim \log^2\left(\frac{N}{\varepsilon'}\right), \qquad \log B_{\exp} \lesssim \log^2\left(\frac{N}{\varepsilon'}\right).$$

By (52) and $\|\psi_{\mathbf{j}}\|_\infty \leq 2$,

$$|\Psi_{\mathbf{j}}(x,t)| \leq 2e,$$

and (62) therefore implies $|\widetilde{\Psi}_{\mathbf{j}}(x,t)| \lesssim 1$. Likewise,

$$e^{-V_{\mathbf{j},0}(x,t)} \leq 1, \qquad |\widetilde{E}_{\mathbf{j},0}(x,t)| \lesssim 1.$$

Apply the multiplication-network lemma from the neural-network toolkit on a fixed bounded interval with base accuracy ε'/N. For a sufficiently large universal constant C, its perturbation estimate gives

$$\begin{aligned}
&\left|\phi_{\mathrm{prod}}\left(\widetilde{E}_{\mathbf{j},0}, \widetilde{\Psi}_{\mathbf{j}}\right) - e^{-V_{\mathbf{j},0}}\Psi_{\mathbf{j}}\right| \\
&\quad\leq \frac{\varepsilon'}{N} + 2C\left(\left|\widetilde{E}_{\mathbf{j},0} - e^{-V_{\mathbf{j},0}}\right| \vee \left|\widetilde{\Psi}_{\mathbf{j}} - \Psi_{\mathbf{j}}\right|\right) \\
&\quad\lesssim_{d,\beta} \frac{\varepsilon'}{N}.
\end{aligned} \tag{63}$$

Thus

$$\widetilde{\Upsilon}_{\mathbf{j}} := \phi_{\mathrm{prod}}\left(\widetilde{E}_{\mathbf{j},0}, \widetilde{\Psi}_{\mathbf{j}}\right)$$

satisfies

$$\left\|\widetilde{\Upsilon}_{\mathbf{j}} - \Upsilon_{\mathbf{j}}\right\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \frac{\varepsilon'}{N}.$$

The networks used here belong to the common class (45) with $\log(1/\varepsilon')$ replaced by $\log(N/\varepsilon')$. The parallelization producing $\widetilde{R}_{\mathbf{j}}$ changes width and sparsity only by a factor depending on d and β. The low-dimensional network $\widetilde{\Phi}_{\mathbf{j}}$ contributes

$$\left(1 + \log N + \log \frac{1}{\varepsilon'}\right)^{\binom{d+r}{d}+1}$$

to the width and

$$\left(1 + \log N + \log \frac{1}{\varepsilon'}\right)^{2\binom{d+r}{d}+5}$$

to the sparsity. The exponential and multiplication networks contribute only lower polylogarithmic terms. Combining these bounds with the standard parallelization and composition rules yields precisely (56). □

We first apply Lemma B.5 with $\psi_{\mathbf{j}} \equiv 1$. By (46), there exists a scalar ReLU network $Q_{\mathbf{j}}$ such that

$$\left| N^d \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|x - t g^\circ_{\mathbf{j}}(u)\|^2}{2q_t} \right\} du - Q_{\mathbf{j}}(x,t) \right| \lesssim_{d,\beta} \frac{\varepsilon'}{N}. \tag{64}$$

Next, for each $1 \le l \le D$, take

$$\psi_{\mathbf{j}}(w) := g^\circ_{\mathbf{j},l}\left(u_{\mathbf{j}} - \frac{w}{N}\right).$$

By (29), (2), and the standing choice of ε for which the right-hand side of (30) is at most one,

$$\begin{aligned}\sup_{w\in[0,1]^d} \left\| g^\circ_{\mathbf{j}}\left(u_{\mathbf{j}} - \frac{w}{N}\right)\right\| &\le \sup_{u\in\overline{\mathcal{U}_{\mathbf{j}}}} \|g^*(u)\| \\ &\quad + \sup_{u\in\overline{\mathcal{U}_{\mathbf{j}}}} \|g^\circ_{\mathbf{j}}(u) - g^*(u)\| \\ &\le 2.\end{aligned}$$

Hence

$$\|\psi_{\mathbf{j}}\|_{L^\infty([0,1]^d)} \le 2,$$

and Lemma B.5 yields scalar ReLU networks $P_{\mathbf{j},l}$ satisfying

$$\max_{1\le l\le D} \left| N^d \int_{\mathcal{U}_{\mathbf{j}}} g^\circ_{\mathbf{j},l}(u) \exp\left\{ -\frac{\|x - t g^\circ_{\mathbf{j}}(u)\|^2}{2q_t} \right\} du - P_{\mathbf{j},l}(x,t) \right| \lesssim_{d,\beta} \frac{\varepsilon'}{N}. \tag{65}$$

Recall the exact numerator and denominator $P^\circ = (P^\circ_1, \dots, P^\circ_D)^\top$ and Q° from (26). We restore the cell volumes and assemble the local integral networks. Define

$$Q(x,t) := \sum_{\mathbf{j}\in\{1,\dots,N\}^d} N^{-d} Q_{\mathbf{j}}(x,t), \qquad P_l(x,t) := \sum_{\mathbf{j}\in\{1,\dots,N\}^d} N^{-d} P_{\mathbf{j},l}(x,t), \quad 1 \le l \le D,$$

and set

$$P(x,t) := (P_1(x,t), \dots, P_D(x,t))^\top.$$

Since there are exactly N^d cells, (64) and (65) imply

$$\|Q - Q^\circ\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \frac{\varepsilon'}{N}, \qquad \max_{1\le l\le D} \|P_l - P^\circ_l\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \frac{\varepsilon'}{N}. \tag{66}$$

We retain the numerator estimate in this coordinatewise form because the subsequent division is performed separately for each scalar ratio P°_l/Q°. Thus no Euclidean norm conversion, and hence no factor $\sqrt{D}$, is needed at this stage.

By Lemma B.17, the scalar denominator network Q and each scalar numerator-coordinate network P_l can be realized with

$$\begin{aligned} L_Q,\ L_{P_l} &\le L_\Upsilon + 1, \\ \|\mathbf{W}_Q\|_\infty,\ \|\mathbf{W}_{P_l}\|_\infty &\lesssim N^d \|\mathbf{W}_\Upsilon\|_\infty, \\ S_Q,\ S_{P_l} &\lesssim N^d (S_\Upsilon + 1), \\ B_Q,\ B_{P_l} &\lesssim B_\Upsilon \vee 1. \end{aligned}$$

The coefficients N^{-d} appearing in the final summation are implemented exactly by the output affine layer and have magnitude at most one.

Parallelizing the D numerator-coordinate networks gives the D-output ReLU network P satisfying

$$\begin{aligned} L_P &\le L_\Upsilon + 1, \\ \|\mathbf{W}_P\|_\infty &\lesssim D N^d \|\mathbf{W}_\Upsilon\|_\infty, \\ S_P &\lesssim D N^d (S_\Upsilon + 1), \\ B_P &\lesssim B_\Upsilon \vee 1. \end{aligned}$$

Thus the ambient dimension enters the network size through the parallelization of the D output coordinates, whereas the approximation tolerance for each numerator coordinate remains of order ε'/N.

Finally, parallelizing P and Q gives a $(D+1)$-output ReLU network (P, Q) with

$$\begin{aligned} L_{P,Q} &\le L_\Upsilon + 1, \\ \|\mathbf{W}_{P,Q}\|_\infty &\lesssim (D+1) N^d \|\mathbf{W}_\Upsilon\|_\infty, \\ S_{P,Q} &\lesssim (D+1) N^d (S_\Upsilon + 1), \\ B_{P,Q} &\lesssim B_\Upsilon \vee 1. \end{aligned}$$

Hence the local-integral construction reduces the remaining approximation problem to obtaining a uniform positive lower bound for Q° and approximating the scalar ratios P_l°/Q°, $1 \le l \le D$.

Denominator control and posterior assembly. We approximate the coordinatewise ratios P_l°/Q°, $1 \le l \le D$, by first establishing that the exact denominator is uniformly bounded away from zero on $\mathcal{C}^*_{[0,1]}$ and then applying the robust division network from (Yakovlev & Puchkin, 2025, Lemma A.4).

Fix $(x,t) \in \mathcal{C}^*_{[0,1]}$. By the definition of $\mathcal{K}_t$, there exists $u_* \in [0,1]^d$ such that $\|x - t g^*(u_*)\| \le R_t$. Let $\mathcal{U}_{\mathbf{j}_*}$ be the unique cell containing u_*. By the Hölder regularity of g^* and the local Taylor estimate in (30), for every $u \in \mathcal{U}_{\mathbf{j}_*}$,

$$\|g^\circ_{\mathbf{j}_*}(u) - g^*(u_*)\| \lesssim_{d,\beta} H\sqrt{D}\, N^{-(\beta\wedge 1)}.$$

Consequently,

$$\begin{aligned} \frac{\|x - t g^\circ_{\mathbf{j}_*}(u)\|^2}{2q_t} &\le \frac{\|x - t g^*(u_*)\|^2}{q_t} + \frac{t^2 \|g^\circ_{\mathbf{j}_*}(u) - g^*(u_*)\|^2}{q_t} \\ &\le \frac{R_t^2}{q_t} + C_{d,\beta} \frac{DH^2}{\sigma^2 N^{2(\beta\wedge 1)}}. \end{aligned}$$

Here we used $t^2/q_t \le \sigma^{-2}$ for $t \in [0,1]$. Therefore, using only the cell $\mathcal{U}_{\mathbf{j}_*}$ in (26),

$$Q^\circ(x,t) \ge N^{-d} \exp\left\{ -\frac{R_t^2}{q_t} - C_{d,\beta} \frac{DH^2}{\sigma^2 N^{2(\beta\wedge 1)}} \right\}. \tag{67}$$

Since the integrand defining Q° is bounded by one, $Q^\circ(x,t) \le 1$. Moreover, by the choice of R_t above, R_t^2/q_t is independent of t, so the lower bound (67) is uniform over $\mathcal{C}^*_{[0,1]}$.

Choose an integer $K \ge 4$ such that

$$2^{-K} \le N^{-d} \exp\left\{ -\frac{R_t^2}{q_t} - C_{d,\beta} \frac{DH^2}{\sigma^2 N^{2(\beta\wedge 1)}} \right\}. \tag{68}$$

Then $2^{-K} \le Q^\circ(x,t) \le 1$ uniformly on $\mathcal{C}^*_{[0,1]}$. In particular, one may take

$$K \lesssim_{d,\beta} 1 + d\log N + \frac{R_t^2}{q_t} + \frac{DH^2}{\sigma^2 N^{2(\beta\wedge 1)}}.$$

By $\|g^\circ\|_{L^\infty} \le 2$, each exact numerator coordinate satisfies

$$|P_l^\circ(x,t)| \le 2Q^\circ(x,t).$$

Hence

$$\left|\frac{1}{2}P_l^\circ(x,t)\right| \le Q^\circ(x,t),$$

which is precisely the numerator range required by Lemma B.16. Define its output-accuracy parameter by

$$\eta_{\mathrm{div}} := c_1 \frac{\varepsilon^\beta}{K^2 + \log^2(1/\varepsilon)},$$

where $c_1 = c_1(d,\beta) > 0$ is sufficiently small, and choose

$$\varepsilon' := c_2 N 2^{-2K} \eta_{\mathrm{div}}, \tag{69}$$

with $c_2 = c_2(d,\beta) > 0$ small enough to absorb the constants in (66). The approximations $(P_l/2, Q)$ then differ from $(P_l^\circ/2, Q^\circ)$ by at most $2^{-2K}\eta_{\mathrm{div}}$. Lemma B.16 therefore provides a scalar ReLU network $\mathcal{R}$ such that, for

$$\widetilde{m}_l(x,t) := 2\,\mathcal{R}\left(\frac{1}{2}P_l(x,t), Q(x,t)\right), \qquad 1 \le l \le D,$$

we have

$$\max_{1\le l\le D} \|\widetilde{m}_l - m_l^\circ\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \varepsilon^\beta.$$

Setting $\widetilde{m} = (\widetilde{m}_1, \dots, \widetilde{m}_D)^\top$ and performing the scalar-to-vector norm conversion gives

$$\|\widetilde{m} - m_t^\circ\|_{L^\infty(\mathcal{C}^*_{[0,1]})} \lesssim_{d,\beta} \sqrt{D}\,\varepsilon^\beta. \tag{70}$$

Since $K \ge 4$, $\log(1/\eta_{\mathrm{div}}) \lesssim_{d,\beta} K + \log(1/\varepsilon)$. Hence the complexity statement in Lemma B.16 gives

$$\begin{aligned}
L_R &\lesssim_{d,\beta} K^2 + \log^2\frac{1}{\varepsilon},\\
\|\mathbf{W}_R\|_\infty &\lesssim_{d,\beta} K^3 + K\log^2\frac{1}{\varepsilon},\\
S_R &\lesssim_{d,\beta} K^4 + K\log^3\frac{1}{\varepsilon},\\
B_R &\lesssim_{d,\beta} 16^K\left(K^2 + \log^2\frac{1}{\varepsilon}\right).
\end{aligned}$$

Starting from the $(D+1)$-output network (P,Q) constructed in the local-integral construction, one exact affine layer forms the D pairs $(P_l/2, Q)$, $1 \le l \le D$. Parallelizing D copies of $\mathcal{R}$ and applying the final exact scaling by 2 yields the D-output network $\widetilde{m}$ with

$$\begin{aligned}
L_m &\lesssim_{d,\beta} L_{P,Q} + L_R + 2,\\
\|\mathbf{W}_m\|_\infty &\lesssim_{d,\beta} \|\mathbf{W}_{P,Q}\|_\infty + D\|\mathbf{W}_R\|_\infty + D,\\
S_m &\lesssim_{d,\beta} S_{P,Q} + DS_R + D,\\
B_m &\lesssim_{d,\beta} B_{P,Q} \vee B_R \vee 2.
\end{aligned} \tag{71}$$

All hidden constants depend only on d and β. The value-level output is $f^{\mathrm{val}} := \widetilde{m}$. To record the complexity of this value-only circuit, use $N^d \lesssim_d \varepsilon^{-d}$, $\log N \lesssim 1 + \log(1/\varepsilon)$, and

$$\frac{R_t^2}{q_t} \lesssim_\beta \Gamma_\varepsilon^2, \qquad K \lesssim_{d,\beta} \Gamma_\varepsilon^2, \qquad \log\frac{N}{\varepsilon'} \lesssim_{d,\beta} \Gamma_\varepsilon^2.$$

Substitution into the component configurations and parallelization of the D outputs give $L \lesssim_{d,\beta} \Gamma_\varepsilon^4$, $W \lesssim_{d,\beta} D^2\varepsilon^{-d}\Gamma_\varepsilon^{2m_\beta+6}$, $S \lesssim_{d,\beta} D^2\varepsilon^{-d}\Gamma_\varepsilon^{4m_\beta+10}$, and $\log B \lesssim_{d,\beta} \Gamma_\varepsilon^4$. Together with (70), this establishes the value-level latent-cell construction. These are baseline bounds for the value modules, rather than the final architecture in Theorem 3.1. The spatial refinement below rebuilds the same latent-cell circuit with first-order modules and derives the final common envelope in (40).

B.4.2 First-order spatial refinement

We refine the cell factorization from (44)–(46) by replacing its scalar modules with first-order stable realizations. Two distinct variables play different roles. Hölder regularity of g^* controls approximation in the latent variable u, producing the ε^β error and the ε^{-d} cell count. Spatial differentiability in the ambient state x instead comes from the Gaussian kernel: differentiating the posterior weights requires no derivative of the bounded generator. Consequently, Gaussian smoothing supplies the covariance identity below. The resulting spatial control is the regularity used in deterministic FM perturbation analyses such as Zhou & Liu (2025) and Kunkel (2026).

The proof has four interfaces, recorded below in the order in which they are used. The covariance identity controls the exact posterior; the moving tube supplies both a denominator lower bound and an exact ReLU gate; fixed-box first-order modules realize the cell integrals and their quotient; and the final local lemma collects the error and architecture bookkeeping. Lemma B.3 then performs the global assembly. This separation keeps the main approximation proof focused on the statistical construction while making every first-order ingredient explicit here.

For fixed t, the Gaussian numerator and denominator are analytic in x; since $Q^\circ(x,t) > 0$ on the real domain, their ratio is real analytic there. After the latent-cell factorization, the nonlinear dependence is carried by fixed-dimensional analytic maps of coefficient vectors whose dimensions depend only on (d, β). General first-order ReLU approximation is studied, for example, by Gühring et al. (2020). Our fixed-box module follows the constructive polynomial architecture underlying Opschoor et al. (2022); because the entrywise parameter bound needed here is not quoted as a separate conclusion, it is derived below from the same polynomial realization. The citation therefore supplies the proof architecture, while the stated $W^{1,\infty}$ and parameter-magnitude bounds are verified in the lemma itself. This preserves the latent cell count.

Exact posterior geometry.

Lemma B.6 (Posterior covariance identity)**.** *Let $g : [0,1]^d \to \mathbb{R}^D$ be measurable with $\sup_u \|g(u)\| \le R$, and let $m_{g,t}$ be the posterior mean for $X_t = tg(U) + \sqrt{q_t}Z$. For every x and $t \in [0,1]$,*

$$\nabla_x m_{g,t}(x) = \frac{t}{q_t}\,\mathrm{Cov}(g(U) \mid X_t = x), \qquad \|\nabla_x m_{g,t}(x)\|_{\mathrm{op}} \le R^2\,\frac{t}{q_t}. \tag{72}$$

In particular, $\mathrm{Lip}_x(m_t^\circ) \le 4t/q_t$.

Proof. Let $w(u) = \exp\{-\|x - tg(u)\|^2/(2q_t)\}$ and normalize $w(u)du$ to a probability measure $\mu_{x,t}$. Dominated differentiation of the posterior quotient gives

$$\partial_{x_k} m_{g,t,l} = \frac{t}{q_t}\{\mathbb{E}_\mu[g_l g_k] - \mathbb{E}_\mu g_l\,\mathbb{E}_\mu g_k\}.$$

For any unit vector z, the quadratic form of this covariance is $\mathrm{Var}_\mu(z^\top g) \le \mathbb{E}_\mu(z^\top g)^2 \le R^2$. The surrogate bound (34) gives the last assertion. □

Put $\Lambda_\varepsilon := R_t/\sqrt{q_t} = \sqrt{D} + 16(\sqrt{DL_\varepsilon} \vee L_\varepsilon)$; the right side is independent of t. The original localization set $\mathcal{K}_t$ is unchanged. The following larger tube is only the domain on which the network ratio and its spatial derivative are controlled.

Lemma B.7 (Moving tube and denominator)**.** *Write $c_{\mathbf{j}} = g^*(u_{\mathbf{j}})$, $s_\sigma(t) = (1-t) + \sigma t$, and $d_N(x,t) = \min_{\mathbf{j}} \|x - tc_{\mathbf{j}}\|_\infty$. Choose $E_N = C_{d,\beta} H\sqrt{D}\,N^{-(\beta\wedge 1)}$ to dominate the cellwise Taylor variation, and a fixed $c_0 > 0$. Set*

$$r_{\mathrm{in}} = \Lambda_\varepsilon s_\sigma + E_N, \quad r_{\mathrm{out}} = r_{\mathrm{in}} + c_0\sigma, \quad \mathcal{T}_\varepsilon^+ = \{(x,t) : d_N(x,t) \le r_{\mathrm{out}}(t)\},$$

and $\chi(x,t) = \mathrm{clip}_{[0,1]}((r_{\mathrm{out}}(t) - d_N(x,t))/(c_0\sigma))$. Then $\chi = 1$ on $\mathcal{C}^*_{[0,1]}$, $\mathrm{Lip}_x\,\chi \le (c_0\sigma)^{-1}$, and throughout $\mathcal{T}_\varepsilon^+$,

$$Q^\circ(x,t) \ge N^{-d}\exp\{-CD(\Lambda_\varepsilon + 1)^2\} \ge 2^{-K_\varepsilon}, \qquad K_\varepsilon \lesssim_{d,\beta} \Gamma_\varepsilon^3. \tag{73}$$

The detector χ has an exact ReLU realization of width and sparsity $O(DN^d)$.

Proof. The inequalities $\sqrt{q_t} \le s_\sigma(t) \le \sqrt{2q_t}$ and $E_N \lesssim \sigma$ follow from the scale conditions. If $x \in \mathcal{K}_t$, choose a witnessing u and its cell. The Taylor bound gives $d_N(x,t) \le R_t + E_N \le r_{\mathrm{in}}(t)$. The minimum of the distances to fixed centers is one-Lipschitz in x, proving the gate bound. The absolute value, maximum, minimum, and scalar clip have exact ReLU formulas; balanced maximum and minimum trees yield the stated detector size.

For $(x,t) \in \mathcal{T}_\varepsilon^+$ choose a minimizing cell. On that entire cell, $\|x - t g_{\mathbf{j}}^\circ(u)\| \le \sqrt{D}\, r_{\mathrm{out}}(t) + E_N$. Moreover, $r_{\mathrm{out}}(t)/\sqrt{q_t} \lesssim \Lambda_\varepsilon + 1$. Integrating the positive Gaussian weight only over this cell, of volume N^{-d}, proves the first lower bound. The factor D comes from $\|z\| \le \sqrt{D}\|z\|_\infty$. Taking dyadic logarithms and using $d \log N + D(\Lambda_\varepsilon + 1)^2 \lesssim_{d,\beta} \Gamma_\varepsilon^3$ gives the second. □

Derivative-stable elementary modules. For $F = (F_1, \dots, F_m) : E \to \mathbb{R}^m$ on $E \subset \mathbb{R}^D \times [0,1]$, write

$$\|F\|_{\mathsf{W}_x(E)} := \max_{1 \le r \le m} \operatorname*{ess\,sup}_{(x,t) \in E} \big(|F_r(x,t)| + \|\nabla_x F_r(x,t)\|_2 \big).$$

For $m = 1$ this is the scalar norm used below. Only derivatives in x occur in W_x. The following three modules make the first-order construction explicit.

Lemma B.8 (Zero-preserving multiplication)**.** *For $M \ge 1$ and $0 < \delta < 1$, there is a two-input ReLU network $\Pi_{M,\delta}$ such that, on $[-M, M]^2$,*

$$\max\big\{ |\Pi_{M,\delta}(a,b) - ab|, |\partial_a \Pi_{M,\delta}(a,b) - b|, |\partial_b \Pi_{M,\delta}(a,b) - a| \big\} \le \delta \tag{74}$$

almost everywhere. Its depth and sparsity are $O(1 + \log(M/\delta))$, its width is bounded by a universal constant, and $\log B \lesssim 1 + \log M$. In addition,

$$\Pi_{M,\delta}(0,b) = 0 \qquad \text{for every } b \in \mathbb{R}, \tag{75}$$

and hence $\partial_b \Pi_{M,\delta}(0,b) = 0$ almost everywhere. If $A, B, \widetilde{A}, \widetilde{B}$ take values in $[-M, M]$, the ordinary chain rule and (74) imply

$$\|\Pi_{M,\delta}(\widetilde{A}, \widetilde{B}) - AB\|_{\mathsf{W}_x(E)} \le C_M \big(1 + \|A\|_{\mathsf{W}_x(E)} + \|B\|_{\mathsf{W}_x(E)}\big) \tag{76}$$

$$\times \big(\delta + \|\widetilde{A} - A\|_{\mathsf{W}_x(E)} + \|\widetilde{B} - B\|_{\mathsf{W}_x(E)}\big). \tag{77}$$

Proof. Let S_m be the standard deep ReLU realization of the dyadic linear interpolant of z^2 on $[0,1]$, extended by the same ReLU formula to $\mathbb{R}$. It satisfies $\|S_m - z^2\|_{W^{1,\infty}(0,1)} \le C 2^{-m}$ and $S_m(0) = 0$. Set

$$\Pi_{M,\delta}(a,b) := 2M^2 \left[S_m\left(\frac{|a+b|}{2M}\right) - S_m\left(\frac{|a|}{2M}\right) - S_m\left(\frac{|b|}{2M}\right) \right]$$

and take $m \ge C(1 + \log(M/\delta))$. Polarization gives (74). When $a = 0$, the first and third terms are identical for every real b, proving (75) without any restriction on the second input. The last assertion follows by differentiating the composition almost everywhere and adding and subtracting the exact product. □

Lemma B.9 (Fixed-box holomorphic maps)**.** *Let m be fixed and let $\mathcal{B} \subset \mathbb{R}^m$ be a fixed box. Suppose that F extends holomorphically to a fixed complex neighborhood $\mathcal{U}$ of $\mathcal{B}$ and $\sup_{z \in \mathcal{U}} |F(z)| \le M_0$. For $0 < \zeta < 1$, there is a ReLU network F_ζ satisfying*

$$\|F_\zeta - F\|_{W^{1,\infty}(\mathcal{B})} \le \zeta, \qquad L, W, S, \log B \le C_{m,\mathcal{U},M_0} \big(1 + \log(1/\zeta)\big)^{C_m}. \tag{78}$$

The same conclusion holds uniformly for a family sharing $(m, \mathcal{B}, \mathcal{U}, M_0)$, and every such family has uniformly bounded second derivatives on any smaller fixed complex neighborhood.

Proof. After an affine change of variables, truncate the tensor Chebyshev expansion at degree $s = C \log(1/\zeta)$. Holomorphy on $\mathcal{U}$ gives exponential decay of the coefficients and hence value and first-derivative truncation error at most $\zeta/2$. Realize the resulting polynomial with Lemma B.8, allocating accuracy $c\zeta C^{-s}(s+1)^{-m-2}$ to every product. The number of monomials is polynomial in s, and the sum of the absolute monomial coefficients is at most C^s. Consequently all depth, width, and sparsity costs are polynomial in s and $\log B \le C(s + \log(1/\zeta))$. This proves (78). The derivative statement is Cauchy's estimate on a smaller neighborhood. This is the fixed-dimensional constructive mechanism behind Opschoor et al. (2022, Theorem 3.6); the displayed parameter bound is extracted here from the polynomial construction rather than quoted from that theorem. □

In particular, if ψ is bounded and

$$\Phi(v) = \int_{[0,1]^d} \psi(w) \exp\{-v^\top a(w) - b(w)\}\, dw, \qquad \|a\|_\infty \le A_0, \quad \|b\|_\infty \le B_0, \tag{79}$$

where A_0, B_0 are fixed, then Φ is entire and the preceding lemma applies on every fixed real box. Indeed, on a fixed complex neighborhood the integrand is dominated by a constant depending only on that neighborhood, A_0, B_0, and $\|\psi\|_\infty$. This observation also supplies the common complex supremum and the second-derivative bound needed below. Unlike the integral-specific corollary, the full Lemma B.9 also applies to the fixed-box quotient map used in the division module.

Lemma B.10 (Tapered exponential under a spatial gradient envelope)**.** *Let $V \ge 0$ and $\widetilde{V} \ge -1$ on E, with $\|\nabla_x V\|_2 + \|\nabla_x \widetilde{V}\|_2 \le A$, $A \ge 1$, and $\|\widetilde{V} - V\|_{\mathsf{W}_x(E)} \le \rho$. For $0 < \kappa < 1$ there is a one-input ReLU network $\mathcal{E}_{A,\kappa}$ such that*

$$\|\mathcal{E}_{A,\kappa}(\widetilde{V}) - e^{-V}\|_{\mathsf{W}_x(E)} \le C(1+A)(\rho+\kappa). \tag{80}$$

It is identically zero when $\widetilde{V} \ge T+1$, where $T = C[1+\log(A/\kappa)]$, and its complexity is polynomial in $T + \log(1/\kappa)$, with $\log B$ bounded by the same quantity.

Proof. Let $g_T(s) = \mathrm{clip}_{[0,1]}(T+1-s)$ and $\vartheta_T(s) = g_T(s)^2(3-2g_T(s))$. Then $F_T(s) = e^{-s}\vartheta_T(s)$ equals e^{-s} on $[0,T]$, vanishes on $[T+1,\infty)$, and satisfies

$$\sup_{s \ge T} \big(|e^{-s} - F_T(s)| + |-e^{-s} - F_T'(s)|\big) \le Ce^{-T}.$$

Approximate e^{-s} in $W^{1,\infty}([-1, T+1])$ after affine rescaling. With a fixed input bound $M \ge 4$, realize the cutoff in the order

$$u_T = \Pi_{M,\kappa}(g_T, g_T), \qquad \widetilde{\vartheta}_T = \Pi_{M,\kappa}(u_T, 3 - 2g_T),$$

and use one final zero-preserving multiplier with $\widetilde{\vartheta}_T$ as its first input and the exponential approximant as its second input. The output is therefore exactly zero for every second input whenever $s \ge T+1$. The one-dimensional Chebyshev argument used in Lemma B.9 has cost polynomial in $T + \log(1/\kappa)$. Since $Ae^{-T} \le C\kappa$, the chain rule on $\{V \ge T\}$ and the Lipschitz bounds of F_T, F_T' on $[-1, T+1]$ yield (80). □

Principal local realization. The preceding ingredients are now assembled cell by cell. The proof is expanded into normalization, backward error allocation, stable division, and vector/architecture bookkeeping so that the origin of every dimension and accuracy factor remains visible.

Lemma B.11 (First-order cell and ratio realization)**.** *Let $\eta_* = c\min\{\varepsilon^\beta, D^{-1}\}$ with a sufficiently small $c = c(d,\beta)$. There is a D-output ReLU network f_θ^{loc} such that on $\mathcal{T}_\varepsilon^+$, almost everywhere in x,*

$$\|f_\theta^{\mathrm{loc}} - m_t^\circ\|_2 \le \sqrt{D}\,\eta_*, \qquad \|J_x f_\theta^{\mathrm{loc}} - J_x m_t^\circ\|_{\mathrm{op}} \le D\eta_* \le 1. \tag{81}$$

The network has the architecture envelope (40)*.*

Proof. **Step 1: normalization on the enlarged tube.** Retain the cells, Taylor polynomials, scalar coordinates, and the exact N^{-d} factor in (44)–(46). For the present first-order realization only, replace every old normalizing constant by

$$C_{\mathbf{j},\mathbf{k}}^+ := 1 \vee \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{T}_\varepsilon^+)}.$$

Define

$$R_{\mathbf{j}}^+ := \left(\left(\frac{V_{\mathbf{j},\mathbf{k}}}{2C_{\mathbf{j},\mathbf{k}}^+} + \frac{1}{2} \right)_{1 \le |\mathbf{k}| \le r}, \frac{V}{\|V\|_{L^\infty([0,1])}} \right)^\top.$$

Define $a_{\mathbf{j}}^+, b_{\mathbf{j}}^+$ by (48)–(49), with $C_{\mathbf{j},\mathbf{k}}$ replaced by $C_{\mathbf{j},\mathbf{k}}^+$. Thus the cell identity itself is unchanged and $R_{\mathbf{j}}^+ \in [0,1]^{m_\beta^{\mathrm{Tay}}}$ on the enlarged tube. The Taylor estimate, the scale condition, and $r_{\mathrm{out}}/\sqrt{q_t} \lesssim \Lambda_\varepsilon + 1$ give

$$\sup_{w \in [0,1]^d} \big(\|a_{\mathbf{j}}^+(w)\|_\infty + |b_{\mathbf{j}}^+(w)|\big) \le C_{d,\beta}, \qquad C_{\mathbf{j},\mathbf{k}}^+ \le [1 + D + H + \sigma^{-1} + \Lambda_\varepsilon]^{C_{d,\beta}}. \tag{82}$$

The nonlinear input dimension remains at most $m_\beta^{\mathrm{Tay}} + 1 \le m_\beta + 1$, independently of D. Fix $\rho_0 := 1/4$. We approximate $R_{\mathbf{j}}^+$ on the fixed box

$$\mathcal{B}_+ := [-\rho_0, 1+\rho_0]^{m_\beta^{\mathrm{Tay}}} \tag{83}$$

and do not clip its approximation; this avoids the derivative instability of hard clipping at an interior point of the exact range.

Step 2: backward first-order budget. Let $\lambda = 2^{-K_\varepsilon}$ and put

$$M_1 := C(1+\sigma^{-1}), \qquad \delta_{PQ} := c\eta_* M_1^{-1}\lambda^2.$$

All exact numerator and denominator cell sums and their spatial gradients are bounded by M_1. Indeed, $0 < Q^\circ \le 1$ and $|P_l^\circ| \le 2Q^\circ$, while

$$\|\nabla_x Q^\circ\|_2 \le \frac{1}{q_t}\int \|x - tg^\circ(u)\| e^{-\|x-tg^\circ(u)\|^2/(2q_t)} du \le C q_t^{-1/2} \le C\sigma^{-1},$$

and the same calculation with $|g_l^\circ| \le 2$ controls $\|\nabla_x P_l^\circ\|_2$. Put

$$M_\varepsilon := 1 + \sup_{(x,t)\in\mathcal{T}_\varepsilon^+} \|x\|_\infty.$$

The moving-tube definition gives $M_\varepsilon \le C(1+\Lambda_\varepsilon + H + \sigma^{-1})$. Choose a deterministic envelope

$$\mathfrak{A}_\varepsilon := [1 + D + H + \sigma^{-1} + \Lambda_\varepsilon + K_\varepsilon + M_\varepsilon + \log(1/\eta_*)]^{C_{d,\beta}} \tag{84}$$

large enough to dominate the value ranges, spatial gradients, and second-derivative constants of all exact cell coordinates and all intermediate approximations below. Set

$$\rho := \min\{\rho_0/4, c\delta_{PQ}\mathfrak{A}_\varepsilon^{-4}\}, \qquad \kappa := c\delta_{PQ}\mathfrak{A}_\varepsilon^{-4}. \tag{85}$$

Then

$$K_\varepsilon + \log(1/\eta_*) + \log(1/\rho) + \log(1/\kappa) \le \Gamma_\varepsilon^{C_{d,\beta}}. \tag{86}$$

Indeed, (73) bounds K_ε, while $\log \mathfrak{A}_\varepsilon$ and $\log M_1$ are logarithmic in the displayed polynomial parameters.

Affine maps in x are exact. Each remaining coordinate is a finite composition of such affine maps, the scalar coefficients $q_t^{-1}, tq_t^{-1}, t^2q_t^{-1}$, squares, and products. Because t is frozen in W_x, the coefficient networks require value accuracy only. The quadratic coordinate and the inner products in $V_{\mathbf{j},0}$ and $V_{\mathbf{j},\mathbf{k}}$ contain at most $C_{d,\beta}D$ arithmetic nodes. Allocate to every coefficient, square, and product module the accuracy

$$\rho_{\mathrm{ar}} := \frac{c_{d,\beta}\rho}{D(1+H+\sigma^{-2})(1+M_\varepsilon)^2\mathfrak{A}_\varepsilon^2}. \tag{87}$$

The reciprocal construction in Lemma B.20, followed by Lemma B.8 and the product rule, then gives $\widetilde{V}_{\mathbf{j},0}$ and $\widetilde{R}_{\mathbf{j}}^+$ such that

$$\|\widetilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\|_{\mathsf{W}_x} + \|\widetilde{R}_{\mathbf{j}}^+ - R_{\mathbf{j}}^+\|_{\mathsf{W}_x} \le \rho, \qquad \|\nabla_x \widetilde{V}_{\mathbf{j},0}\|_2 + \max_r \|\nabla_x \widetilde{R}_{\mathbf{j},r}^+\|_2 \le \mathfrak{A}_\varepsilon. \tag{88}$$

Because $R_{\mathbf{j}}^+ \in [0,1]^{m_\beta^{\mathrm{Tay}}}$ and $\rho \le \rho_0/4$, the approximate coordinate vector remains in $\mathcal{B}_+$, so the holomorphic network below is evaluated only on the box on which its first-order estimate is valid. Indeed, each value error is amplified by at most $C(1+H+\sigma^{-2})(1+M_\varepsilon)$, and each spatial derivative error by at most the same quantity times $\mathfrak{A}_\varepsilon^2$. Summing over at most $C_{d,\beta}D$ nodes and using (87) leaves the total below ρ. Moreover, $\log(1/\rho_{\mathrm{ar}}) \le \Gamma_\varepsilon^{C_{d,\beta}}$, so every cost here remains polynomial in the logarithmic budget (86).

For a cell primitive $\Phi_{\mathbf{j},\psi}$, the bounds (82) put it in the family (79) with fixed A_0, B_0. Lemma B.9 produces $\widetilde{\Phi}_{\mathbf{j},\psi}$ with first-order error ρ on the expanded box. Its uniform second-derivative bound and (88) give

$$\|\widetilde{\Phi}_{\mathbf{j},\psi}(\widetilde{R}_{\mathbf{j}}^+) - \Phi_{\mathbf{j},\psi}(R_{\mathbf{j}}^+)\|_{\mathsf{W}_x} \le C\mathfrak{A}_\varepsilon^2\rho. \tag{89}$$

For the outer Gaussian exponent, apply Lemma B.10 to $(V_{\mathbf{j},0}, \widetilde{V}_{\mathbf{j},0})$ with gradient envelope $\mathfrak{A}_\varepsilon$. Its truncation level is

$$T = C[1 + \log(\mathfrak{A}_\varepsilon/\kappa)],$$

and both its value and spatial derivative contribute at most $C\mathfrak{A}_\varepsilon(\rho + \kappa)$. Notice that this estimate uses the general factor $\mathfrak{A}_\varepsilon e^{-T}$; it does not rely on the special identity for an exact quadratic exponent after the exponent has been approximated.

Finally combine the tapered exponential, the composed primitive, and the bounded polynomial prefactors by repeated use of Lemma B.8. Allocating error κ to each of the finitely many products yields networks for the unit-cube-normalized exact cells

$$\bar{Q}^\circ_{\mathbf{j}} := N^d Q^\circ_{\mathbf{j}}, \qquad \bar{P}^\circ_{\mathbf{j},l} := N^d P^\circ_{\mathbf{j},l}.$$

Denote these networks by $\widetilde{\bar{Q}}_{\mathbf{j}}$ and $\widetilde{\bar{P}}_{\mathbf{j},l}$. For every cell and every numerator coordinate,

$$\|\widetilde{\bar{Q}}_{\mathbf{j}} - \bar{Q}^\circ_{\mathbf{j}}\|_{\mathsf{W}_x} + \max_{1\le l\le D} \|\widetilde{\bar{P}}_{\mathbf{j},l} - \bar{P}^\circ_{\mathbf{j},l}\|_{\mathsf{W}_x} \le C_{d,\beta}\mathfrak{A}^3_\varepsilon(\rho+\kappa) \le \delta_{PQ}. \tag{90}$$

Restoring the exact cell volume gives

$$\widetilde{Q} = N^{-d}\sum_{\mathbf{j}} \widetilde{\bar{Q}}_{\mathbf{j}}, \qquad \widetilde{P}_l = N^{-d}\sum_{\mathbf{j}} \widetilde{\bar{P}}_{\mathbf{j},l}.$$

Since $Q^\circ = N^{-d}\sum_{\mathbf{j}} \bar{Q}^\circ_{\mathbf{j}}$ and $P^\circ_l = N^{-d}\sum_{\mathbf{j}} \bar{P}^\circ_{\mathbf{j},l}$, the triangle inequality and (90) imply

$$\|\widetilde{Q} - Q^\circ\|_{\mathsf{W}_x} + \max_l \|\widetilde{P}_l - P^\circ_l\|_{\mathsf{W}_x} \le C_{d,\beta}\delta_{PQ}; \tag{91}$$

there is no factor N^d in the error.

Step 3: robust dyadic division. The exact inequalities $Q^\circ \in [\lambda, 1]$ and $|P^\circ_l| \le 2Q^\circ$, together with (91) and $\delta_{PQ} \le c\lambda$, place $(\widetilde{P}_l, \widetilde{Q})$ in the robust band $\widetilde{Q} \in [\lambda/2, 3/2]$ and $|\widetilde{P}_l| \le 3\widetilde{Q}$. In the assembled division network below, $p = \widetilde{P}_l$ and $q = \widetilde{Q}$; the symbols P°_l, Q° are reserved for the exact perturbation comparison. There are exact piecewise-linear hats h_j, $0 \le j \le K_\varepsilon$, forming a partition of unity on this band, with at most two active hats, $\mathrm{Lip}(h_j) \le C2^j$, and, on the support of h_j, $z = 2^j q \in [1/2, 2]$. Put $w = 2^j p/8$; then $w \in [-1, 1]$ on a fixed slightly enlarged shoulder. Apply Lemma B.9 to

$$F(w, z) = 8w/z, \qquad (w, z) \in [-1, 1] \times [1/2, 2],$$

and choose local first-order accuracy

$$\delta_j = c\eta_* M_1^{-1} 2^{-j}. \tag{92}$$

Allocate the same δ_j accuracy to the final multiplication by the hat. After composition with $p(x), q(x)$, the normalization error is bounded by $M_1 2^j \delta_j \le c\eta_*$, and the dangerous hat term is

$$M_1 |h_j'(q)|\, |F_j - F| \le C M_1 2^j \delta_j \le C\eta_*.$$

Inactive blocks have exactly zero value and zero derivative in their uncontrolled second input by (75); since at most two hats are active, the assembled division error contains no factor K_ε. This constructs a genuine ReLU network $\mathrm{Div}_{K_\varepsilon,\eta_*}$ with

$$\left\| \mathrm{Div}_{K_\varepsilon,\eta_*}(p, q) - \frac{p}{q} \right\|_{W^{1,\infty}} \le C\eta_*, \qquad \log B_{\mathrm{div}} \lesssim K_\varepsilon + \log(1/\eta_*) + \log M_1. \tag{93}$$

Independently, direct differentiation of the exact quotient gives

$$\left\| \frac{\widetilde{P}_l}{\widetilde{Q}} - \frac{P^\circ_l}{Q^\circ} \right\|_{\mathsf{W}_x} \le C\left(\frac{\delta_{PQ}}{\lambda} + \frac{M_1\delta_{PQ}}{\lambda^2} \right) \le C\eta_*, \tag{94}$$

which explains the necessary backward choice $\delta_{PQ} \asymp \eta_* M_1^{-1} 2^{-2K_\varepsilon}$, equivalently $\delta_{PQ} \lesssim \eta_* \sigma 2^{-2K_\varepsilon}$.

Step 4: vector and architecture bookkeeping. Applying the division network to every coordinate gives scalar value and spatial-gradient error at most $C\eta_*$. After decreasing the constant in the definition of η_*,

$$\|f_\theta^{\mathrm{loc}} - m_t^\circ\|_2 \le \sqrt{D}\,\eta_*, \qquad \|J_x f_\theta^{\mathrm{loc}} - J_x m_t^\circ\|_{\mathrm{op}} \le \|J_x f_\theta^{\mathrm{loc}} - J_x m_t^\circ\|_F \le D\eta_* \le 1.$$

All serial modules have depth, sparsity, and parameter logarithm bounded by a fixed power of the budget in (86). The only algebraic resolution cost is the N^d parallel cell count. Parallelize the denominator and the D numerator coordinates and add the $O(DN^d)$ detector. Since $N^d \lesssim_d \varepsilon^{-d}$, enlarge once, if necessary, the fixed exponent $c_{\mathrm{arch}} = c_{\mathrm{arch}}(d,\beta)$ from Theorem 3.1. Then

$$L, \log B \lesssim_{d,\beta} \Gamma_\varepsilon^{c_{\mathrm{arch}}}, \qquad W, S \lesssim_{d,\beta} D^2\varepsilon^{-d}\Gamma_\varepsilon^{c_{\mathrm{arch}}}. \tag{95}$$

Thus the refinement changes only powers of Γ_ε—in particular, only polylogarithmic factors in $1/\varepsilon$ when (D, H, σ) are fixed—and introduces no new algebraic factor ε^{-c}. □

Proof of Lemma B.3. **Step 1: Value realization.** The latent-cell construction in (44)–(71) identifies a network circuit approximating m_t° on $\mathcal{C}^*_{[0,1]}$ with coordinatewise error of order ε^β. The exact N^{-d} cell weights prevent the sum over the N^d cells from increasing this error. The N^d cells determine the algebraic dependence on the approximation scale. The next step realizes this circuit with first-order scalar modules.

Step 2: Spatial refinement. Lemma B.11 realizes the same posterior ratio on $\mathcal{T}_\varepsilon^+$ with Euclidean value error $\sqrt{D}\eta_*$ and Jacobian error at most one. The covariance identity in Lemma B.6 then gives the absolute Jacobian profile $4t/q_t + 1$ on this tube. The architecture calculation gives, for the finite exponent fixed in Theorem 3.1, $L, \log B \lesssim_{d,\beta} \Gamma_\varepsilon^{c_{\mathrm{arch}}}$ and $W, S \lesssim_{d,\beta} D^2\varepsilon^{-d}\Gamma_\varepsilon^{c_{\mathrm{arch}}}$.

Step 3: Global extension and assembly. Let f_θ^{loc} be the network of Lemma B.11 and clip its coordinates to $[-2, 2]$, obtaining y. Since $\|m_t^\circ\| \le 2$, coordinatewise clipping cannot increase the Euclidean value error. On the enlarged tube, $\|y\| \le 2 + \sqrt{D}\eta_* \le 3$ and $\|J_x y\|_{\mathrm{op}} \le 4t/q_t + 1$. Fix $M_{\mathrm{gate}} := 4$ and define

$$f_{\theta,l}(x,t) := \Pi_{M_{\mathrm{gate}},\delta}(\chi(x,t), y_l(x,t)), \qquad \delta = c\min\{\varepsilon^\beta, D^{-1/2}\}.$$

On $\mathcal{K}_t$, $\chi = 1$, so the local error and multiplier error give (38). Outside $\mathcal{T}_\varepsilon^+$, the output and its weak spatial derivative vanish away from the boundary. On the transition region, the chain rule gives

$$J_x f_\theta = A\nabla_x\chi^\top + \mathrm{diag}(B)J_x y,$$

where $A_l = \partial_1\Pi_{M_{\mathrm{gate}},\delta}(\chi, y_l)$ and $B_l = \partial_2\Pi_{M_{\mathrm{gate}},\delta}(\chi, y_l)$. Hence

$$\|A\|_2 \le \|y\|_2 + \sqrt{D}\delta \le 4, \quad \|\mathrm{diag}(B)\|_{\mathrm{op}} \le 1 + \delta.$$

Lemma B.7 therefore yields $\|J_x f_\theta\|_{\mathrm{op}} \lesssim_{d,\beta} 1 + \sigma^{-1} + t/q_t$. For each t, the ReLU realization is continuous and piecewise affine in x, hence absolutely continuous on line segments. Integrating the almost-everywhere Jacobian bound along such segments gives the global Lipschitz estimate. The detector costs $O(DN^d)$ and the zero-preserving multiplier contributes only logarithmic complexity, so the architecture envelope (40) is preserved. The W_x approximation to m_t° is used only on $\mathcal{T}_\varepsilon^+$. Outside this tube, exact zero preservation gives a global absolute Lipschitz bound; no global approximation of $J_x m_t^\circ$ on $\mathbb{R}^D$ is claimed. This proves all three conclusions. □

B.5 Auxiliary neural-network lemmas

Lemma B.12 (Composition of ReLU networks)**.** *Let*

$$\phi_1 \in \mathrm{NN}(L_1, \mathbf{W}^1, S_1, B_1), \qquad \phi_2 \in \mathrm{NN}(L_2, \mathbf{W}^2, S_2, B_2),$$

with compatible input and output dimensions. By Oko et al. (2023, Lemma F.1), $\phi_2 \circ \phi_1$ can be realized by a ReLU network satisfying

$$L_{\mathrm{comp}} \lesssim L_1 + L_2,$$
$$\|\mathbf{W}_{\mathrm{comp}}\|_\infty \lesssim \|\mathbf{W}^1\|_\infty + \|\mathbf{W}^2\|_\infty,$$
$$S_{\mathrm{comp}} \lesssim S_1 + S_2, \qquad B_{\mathrm{comp}} \lesssim B_1 \vee B_2 \vee 1.$$

Lemma B.13 (Identity padding)**.** *For every $m, L \in \mathbb{N}$, Oko et al. (2023, Lemma F.2) gives an exact ReLU realization of $\mathrm{id}_{\mathbb{R}^m}$ of depth L with*

$$\|\mathbf{W}\|_\infty \lesssim m, \qquad S \lesssim mL, \qquad B = 1.$$

Consequently, networks of unequal depths may be padded to a common depth before parallelization, with only the above additional sparsity cost.

This section collects the neural-network tools used in the posterior construction. In particular, we require ReLU approximations of

$$q_t^{-1}, \qquad tq_t^{-1}, \qquad t^2q_t^{-1}, \qquad t \in [0,1],$$

Under the standing assumption $0 < \sigma < 1$, q_t is uniformly bounded away from zero. We therefore combine standard multiplication and reciprocal approximation networks to construct the required time-dependent coefficients.

Lemma B.14 (Multiplication network)**.** *The multiplication primitive of Oko et al. (2023, Lemma F.6) states that, for $p \geq 2$, $C \geq 1$, $\varepsilon' \in (0,1]$, and $\varepsilon > 0$, there exists a ReLU network $\varphi(x_1, \ldots, x_p)$ with*

$$L \lesssim \log p(\log \varepsilon^{-1} + p \log C), \quad \|\mathbf{W}\|_\infty = 48p, \quad S \lesssim p \log \varepsilon^{-1} + p \log C, \quad B = C^p$$

such that

$$\left| \varphi(x_1', \ldots, x_p') - \prod_{k=1}^{p} x_k \right| \leq \varepsilon + pC^{p-1}\varepsilon',$$

for all $x \in [-C, C]^p$ and $x' \in \mathbb{R}^p$ with $\|x - x'\|_\infty \leq \varepsilon'$. Moreover, $\varphi(x_1', \ldots, x_p') = 0$ if at least one $x_i' = 0$ and $|\varphi(x_1', \ldots, x_p')| \leq C^p$. The same construction approximates $\prod_{i=1}^{I} x_i^{\alpha_i}$ when $\alpha_i \in \mathbb{Z}_+$ and $\sum_{i=1}^{I} \alpha_i = p$.

Lemma B.15 (Reciprocal network)**.** *The reciprocal primitive of Oko et al. (2023, Lemma F.7) gives, for any $0 < \varepsilon < 1$, a ReLU network $\phi_{\mathrm{rec}} \in \mathrm{NN}(L, \mathbf{W}, S, B)$ with*

$$L \lesssim \log^2(\varepsilon^{-1}), \qquad \|\mathbf{W}\|_\infty \lesssim \log^3(\varepsilon^{-1}), \qquad S \lesssim \log^4(\varepsilon^{-1}), \qquad B \lesssim \varepsilon^{-2},$$

such that, for all $x \in [\varepsilon, \varepsilon^{-1}]$ and $x' \in \mathbb{R}$,

$$\left| \phi_{\mathrm{rec}}(x') - \frac{1}{x} \right| \leq \varepsilon + \frac{|x' - x|}{\varepsilon^2}.$$

Lemma B.16 (Robust division)**.** *The following is Yakovlev & Puchkin (2025, Lemma A.4) in the notation used below. Let $K \geq 4$ be an integer. For every $\eta \in (0,1]$, there is a scalar-output ReLU network $\mathcal{R}_\eta$ such that*

$$\left| \mathcal{R}_\eta(x', y') - \frac{x}{y} \right| \leq 2049\big(4K^2 \log^2 2 + \log^2(1/\eta)\big)\eta \tag{96}$$

whenever $y \in [2^{-K}, 1]$, $|x| \leq y$, and $|x - x'| \vee |y - y'| \leq 2^{-2K}\eta$. Moreover,

$$\begin{aligned} L_R &\lesssim K^2 + \log^2(1/\eta), \\ \|\mathbf{W}_R\|_\infty &\lesssim K^3 + K \log^2(1/\eta), \\ S_R &\lesssim K^4 + K \log^3(1/\eta), \\ B_R &\lesssim 16^K\big(K^2 + \log^2(1/\eta)\big). \end{aligned} \tag{97}$$

The hidden constants are universal.

Lemma B.17 (Parallelization and summation of neural networks)**.** *Let $\phi^1, \ldots, \phi^K$ be ReLU networks of the same depth L with $\phi^i \in \mathrm{NN}(L, \mathbf{W}^i, S^i, B^i)$ for $i = 1, \ldots, K$. By Oko et al. (2023, Lemma F.3), their parallelization*

$$\phi_{\mathrm{par}} := \big((\phi^1)^\top, \ldots, (\phi^K)^\top\big)^\top$$

can be realized by a ReLU network satisfying

$$L_{\mathrm{par}} = L, \qquad \|\mathbf{W}_{\mathrm{par}}\|_\infty \lesssim \sum_{i=1}^{K} \|\mathbf{W}^i\|_\infty,$$

$$S_{\mathrm{par}} \lesssim \sum_{i=1}^{K} S^i, \qquad B_{\mathrm{par}} \lesssim \max_{1\le i\le K}(B^i \vee 1).$$

If all outputs are scalar, the sum $\sum_{i=1}^{K}\phi^i$ can be realized by appending one affine layer. In particular,

$$L_{\mathrm{sum}} \le L+1, \qquad \|\mathbf{W}_{\mathrm{sum}}\|_\infty \lesssim \sum_{i=1}^{K} \|\mathbf{W}^i\|_\infty,$$

$$S_{\mathrm{sum}} \lesssim \sum_{i=1}^{K} S^i + K, \qquad B_{\mathrm{sum}} \lesssim \max_{1\le i\le K}(B^i \vee 1).$$

Lemma B.18 (ReLU approximation of the exponential function)**.** *By Yakovlev & Puchkin (2025, Lemma E.2), for any $\varepsilon_0 \in (0,1)$, there exists a ReLU network $\phi_{\exp} \in \mathrm{NN}(L, \mathbf{W}, S, B)$ such that*

$$\sup_{x,x'\ge 0} \left| e^{-x'} - \phi_{\exp}(x) \right| \le \varepsilon_0 + |x - x'|.$$

Moreover,

$$L \lesssim \log^2\left(\frac{1}{\varepsilon_0}\right), \qquad \|\mathbf{W}\|_\infty \lesssim \log\left(\frac{1}{\varepsilon_0}\right),$$

$$S \lesssim \log^2\left(\frac{1}{\varepsilon_0}\right), \qquad \log B \lesssim \log^2\left(\frac{1}{\varepsilon_0}\right).$$

In addition, for every $x \ge \log(3/\varepsilon_0)$, it holds that

$$|\phi_{\exp}(x)| \le \varepsilon_0.$$

Corollary B.19 (Exponential approximation under perturbed inputs)**.** *By Yakovlev & Puchkin (2025, Corollary E.3), for any $\varepsilon_0 \in (0,1)$ and $a \ge 0$, there exists a ReLU network $\phi \in \mathrm{NN}(L, \mathbf{W}, S, B)$ satisfying*

$$L \lesssim \log^2\left(\frac{1}{\varepsilon_0}\right), \qquad \|\mathbf{W}\|_\infty \lesssim \log\left(\frac{1}{\varepsilon_0}\right),$$

$$S \lesssim \log^2\left(\frac{1}{\varepsilon_0}\right), \qquad \log B \lesssim \log^2\left(\frac{1}{\varepsilon_0}\right) + (a \vee 1),$$

such that

$$\sup_{\substack{x\ge 0\\ x'\ge -a}} \left|\phi(x') - e^{-x}\right| \le e^a\big(\varepsilon_0 + |x - x'|\big).$$

The following three lemmas provide the ReLU approximations of the time-dependent and space-time terms required in the posterior construction. The first approximates the coefficients t^γ/q_t, while the next two combine these approximations with multiplication networks to treat the quadratic and linear interaction terms.

Lemma B.20 (Approximation of the Flow Matching time coefficients)**.** *Under the standing assumption $0 < \sigma < 1$, for any $\gamma \in \{0,1,2\}$ and $\delta \in (0,1]$, there exists a ReLU network $\chi_{\gamma,\delta} \in \mathrm{NN}(L, \mathbf{W}, S, B)$ such that*

$$\sup_{t\in[0,1]} \left| \chi_{\gamma,\delta}(t) - \frac{t^\gamma}{q_t} \right| \le \delta.$$

Moreover,

$$L \lesssim \log^2(1/\delta) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$\|\mathbf{W}\|_\infty \lesssim \log^3(1/\delta) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$S \lesssim \log^4(1/\delta) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$\log B \lesssim \log(1/\delta) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Lemma B.21 (Approximation of the quadratic term)**.** *Let $x \in \mathbb{R}^D$ and $M \geq 1$. For any $\varepsilon \in (0,1]$, there exists a ReLU network $\rho_\varepsilon(x,t) \in \mathrm{NN}(L, \mathbf{W}, S, B)$ such that*

$$\sup_{\substack{\|x\|_\infty \leq M \\ t\in[0,1]}} \left|\rho_\varepsilon(x,t) - \frac{\|x\|^2}{2q_t}\right| \leq \varepsilon.$$

Moreover,

$$L \lesssim \log^2(1/\varepsilon) + \log^2(MD) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$\|\mathbf{W}\|_\infty \lesssim D\left[\log^3(1/\varepsilon) + \log^3(MD) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right)\right],$$
$$S \lesssim D\left[\log^4(1/\varepsilon) + \log^4(MD) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right)\right],$$
$$\log B \lesssim \log(1/\varepsilon) + \log(MD) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Lemma B.22 (Approximation of the linear interaction term)**.** *Let $x, a \in \mathbb{R}^D$ and $M \geq 1$. For any $\varepsilon \in (0,1]$, there exists a ReLU network $\omega_\varepsilon(x,t) \in \mathrm{NN}(L, \mathbf{W}, S, B)$ such that*

$$\sup_{\substack{\|x\|_\infty \leq M \\ t\in[0,1]}} \left|\omega_\varepsilon(x,t) - \frac{t\, x^\top a}{q_t}\right| \leq \varepsilon.$$

Moreover,

$$L \lesssim \log^2(1/\varepsilon) + \log^2(DM\|a\|_\infty \vee 1) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$\|\mathbf{W}\|_\infty \lesssim D + \log^3(1/\varepsilon) + \log^3(DM\|a\|_\infty \vee 1) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$S \lesssim D + \log^4(1/\varepsilon) + \log^4(DM\|a\|_\infty \vee 1) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$
$$\log B \lesssim \log(1/\varepsilon) + \log(DM\|a\|_\infty \vee 1) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

B.5.1 Constructions for the Flow Matching modules

Proof of Lemma B.20. We first approximate q_t^{-1}. The lower bound is stated in Section 2; convexity of q_t and $q_0 = 1$, $q_1 = \sigma^2 < 1$ give the upper bound. Thus

$$\frac{\sigma^2}{1+\sigma^2} \leq q_t \leq 1, \qquad t \in [0,1]. \tag{98}$$

Fix

$$\eta := \frac{\delta\sigma^2}{32(1+\sigma^2)}.$$

Since $\delta \le 1$, we have $0 < \eta \le \sigma^2/(1+\sigma^2)$, and therefore $q_t \in [\eta, \eta^{-1}]$ for all $t \in [0,1]$.

We first construct a neural approximation of q_t. Apply Lemma B.14 with $d = 2$, $C = 1$, $\varepsilon = \eta^3/4$, and $\varepsilon' = \eta^3/8$. Taking $x = x' = (t,t)$ gives a multiplication network φ_{sq} satisfying

$$\left|\varphi_{\mathrm{sq}}(t,t) - t^2\right| \le \frac{\eta^3}{4} + 2\frac{\eta^3}{8} = \frac{\eta^3}{2}.$$

Let $\vartheta_\eta(t) := \varphi_{\mathrm{sq}}(t,t)$ and define

$$\widetilde{q}_t := 1 - 2t + (1+\sigma^2)\vartheta_\eta(t).$$

Since $1 + \sigma^2 \le 2$,

$$|\widetilde{q}_t - q_t| \le (1+\sigma^2)|\vartheta_\eta(t) - t^2| \le \eta^3. \tag{99}$$

Let $\phi_{\mathrm{rec},\eta}$ be the reciprocal network from Lemma B.15. Since $q_t \in [\eta, \eta^{-1}]$, its sensitivity estimate together with (99) gives

$$\left|\phi_{\mathrm{rec},\eta}(\widetilde{q}_t) - \frac{1}{q_t}\right| \le \eta + \frac{|\widetilde{q}_t - q_t|}{\eta^2} \le 2\eta = \frac{\delta\sigma^2}{16(1+\sigma^2)}.$$

Hence, setting

$$\chi_{0,\delta}(t) := \phi_{\mathrm{rec},\eta}(\widetilde{q}_t),$$

we have

$$\sup_{t\in[0,1]} \left|\chi_{0,\delta}(t) - \frac{1}{q_t}\right| \le 2\eta \le \delta. \tag{100}$$

Moreover,

$$|\chi_{0,\delta}(t)| \le \frac{1}{q_t} + 2\eta \le \frac{1+\sigma^2}{\sigma^2} + 1 \le \frac{2(1+\sigma^2)}{\sigma^2}.$$

For $\gamma = 1$, apply Lemma B.14 with

$$d = 2, \qquad C = \frac{2(1+\sigma^2)}{\sigma^2}, \qquad \varepsilon = \frac{\delta}{4}, \qquad \varepsilon' = 2\eta,$$

and take

$$x = \left(t, \frac{1}{q_t}\right), \qquad x' = \left(t, \chi_{0,\delta}(t)\right).$$

By (100), $\|x - x'\|_\infty \le 2\eta$. Therefore there exists a multiplication network $\varphi_{1,\delta}$ satisfying

$$\left|\varphi_{1,\delta}\big(t, \chi_{0,\delta}(t)\big) - \frac{t}{q_t}\right| \le \frac{\delta}{4} + 2\left(\frac{2(1+\sigma^2)}{\sigma^2}\right)(2\eta) = \frac{\delta}{2}.$$

Thus

$$\chi_{1,\delta}(t) := \varphi_{1,\delta}\big(t, \chi_{0,\delta}(t)\big)$$

satisfies the required δ-accuracy.

For $\gamma = 2$, we directly approximate $t^2 q_t^{-1}$. Using the same multiplication-network parameters, take

$$x = \left(t^2, \frac{1}{q_t}\right), \qquad x' = \left(\vartheta_\eta(t), \chi_{0,\delta}(t)\right).$$

Since $\eta < 1$,

$$\|x - x'\|_\infty \le \max\left\{\frac{\eta^3}{2}, 2\eta\right\} \le 2\eta.$$

Hence there exists a multiplication network $\varphi_{2,\delta}$ such that

$$\left|\varphi_{2,\delta}\big(\vartheta_\eta(t), \chi_{0,\delta}(t)\big) - \frac{t^2}{q_t}\right| \le \frac{\delta}{4} + 2\left(\frac{2(1+\sigma^2)}{\sigma^2}\right)(2\eta) = \frac{\delta}{2}.$$

Defining

$$\chi_{2,\delta}(t) := \varphi_{2,\delta}\big(\vartheta_\eta(t), \chi_{0,\delta}(t)\big)$$

proves the approximation claim for all $\gamma \in \{0, 1, 2\}$.

Since

$$\eta^{-1} = \frac{32(1+\sigma^2)}{\delta\sigma^2},$$

we have

$$\log(\eta^{-1}) \lesssim \log(1/\delta) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Hence Lemma B.15 yields

$$L_{\mathrm{rec}} \lesssim \log^2(1/\delta) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}_{\mathrm{rec}}\|_\infty \lesssim \log^3(1/\delta) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$S_{\mathrm{rec}} \lesssim \log^4(1/\delta) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right), \qquad \log B_{\mathrm{rec}} \lesssim \log(1/\delta) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

The square network and the additional multiplication network used for $\gamma = 1$ or $\gamma = 2$ have only first-order logarithmic depth and sparsity, constant hidden width, and logarithmic weight complexity. Hence, under composition, their contributions are absorbed by the reciprocal-network bounds. □

Proof of Lemma B.21. Set

$$C := M \vee \frac{DM^2}{2} \vee \frac{1+\sigma^2}{\sigma^2}.$$

We first approximate $\|x\|^2/2$. Apply Lemma B.14 with $p = 2$, domain $[-C, C]^2$, approximation accuracy $\varepsilon/(4DC)$, and perturbation level $\varepsilon/(8DC^2)$. For every $j \in \{1, \dots, D\}$, using the exact inputs (x_j, x_j) gives

$$\left|\phi_{\mathrm{sq}}(x_j, x_j) - x_j^2\right| \le \frac{\varepsilon}{4DC} + 2C\frac{\varepsilon}{8DC^2} = \frac{\varepsilon}{2DC}.$$

Parallel stacking of the D square networks, followed by the affine map $(z_1, \dots, z_D) \mapsto \frac{1}{2}\sum_{j=1}^D z_j$, yields

$$\left|\frac{1}{2}\sum_{j=1}^D \phi_{\mathrm{sq}}(x_j, x_j) - \frac{\|x\|^2}{2}\right| \le \frac{\varepsilon}{4C}. \tag{101}$$

Lemma B.20 with $\gamma = 0$ and accuracy $\varepsilon/(4C)$ gives

$$\sup_{t\in[0,1]} \left|\chi_{0,\varepsilon/(4C)}(t) - \frac{1}{q_t}\right| \le \frac{\varepsilon}{4C}. \tag{102}$$

By the definition of C and (98),

$$\frac{\|x\|^2}{2} \le \frac{DM^2}{2} \le C, \qquad \frac{1}{q_t} \le \frac{1+\sigma^2}{\sigma^2} \le C.$$

Thus both exact factors lie in $[-C, C]$.

Applying Lemma B.14 again with $d = 2$, domain $[-C, C]^2$, approximation accuracy $\varepsilon/2$, and perturbation level $\varepsilon/(4C)$, define

$$\rho_\varepsilon(x, t) := \phi_{\mathrm{mult}}\left(\chi_{0,\varepsilon/(4C)}(t), \frac{1}{2}\sum_{j=1}^D \phi_{\mathrm{sq}}(x_j, x_j)\right).$$

By (101) and (102), the perturbation of both inputs is at most $\varepsilon/(4C)$. Hence

$$\left|\rho_\varepsilon(x, t) - \frac{\|x\|^2}{2q_t}\right| \le \frac{\varepsilon}{2} + 2C\frac{\varepsilon}{4C} = \varepsilon.$$

We now determine the network configuration. Each square network has

$$L_{\mathrm{sq}} \vee S_{\mathrm{sq}} \lesssim \log(1/\varepsilon) + \log D + \log C, \qquad \|\mathbf{W}_{\mathrm{sq}}\|_\infty \lesssim 1, \qquad \log B_{\mathrm{sq}} \lesssim \log C.$$

By Lemma B.17, the D square networks can be evaluated in parallel with

$$L_{\mathrm{par}} = L_{\mathrm{sq}}, \qquad \|\mathbf{W}_{\mathrm{par}}\|_\infty \lesssim D\|\mathbf{W}_{\mathrm{sq}}\|_\infty \lesssim D,$$

and

$$S_{\mathrm{par}} \lesssim DS_{\mathrm{sq}} \lesssim D\big[\log(1/\varepsilon) + \log D + \log C\big].$$

The subsequent affine summation layer adds one layer and only $O(D)$ non-zero weights, and therefore does not change the above width and sparsity orders.

Lemma B.20, used with accuracy $\varepsilon/(4C)$, gives a time-coefficient network satisfying

$$L_{\mathrm{time}} \lesssim \log^2(1/\varepsilon) + \log^2 C + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}_{\mathrm{time}}\|_\infty \lesssim \log^3(1/\varepsilon) + \log^3 C + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$S_{\mathrm{time}} \lesssim \log^4(1/\varepsilon) + \log^4 C + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

and

$$\log B_{\mathrm{time}} \lesssim \log(1/\varepsilon) + \log C + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

The outer multiplication network has only first-order logarithmic depth and sparsity and constant hidden width.

If the spatial and time branches have different depths, the scalar output of the shallower branch is propagated through identity layers until both branches have the same depth. Since only a scalar output is padded, this adds at most $O(L_{\mathrm{time}})$ non-zero parameters, which is absorbed by S_{time}.

Consequently, before simplification,

$$L \lesssim \log^2(1/\varepsilon) + \log^2 C + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}\|_\infty \lesssim D + \log^3(1/\varepsilon) + \log^3 C + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$S \lesssim D\big[\log(1/\varepsilon) + \log D + \log C\big] + \log^4(1/\varepsilon) + \log^4 C + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

and

$$\log B \lesssim \log(1/\varepsilon) + \log C + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Since $M \geq 1$,

$$\log C \lesssim \log(MD) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Moreover, $D \geq 1$ and $(1+\sigma^2)/\sigma^2 > 2$, so the lower-order additive terms can be absorbed into the corresponding higher-order ones. Hence

$$L \lesssim \log^2(1/\varepsilon) + \log^2(MD) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}\|_\infty \lesssim D\left[\log^3(1/\varepsilon) + \log^3(MD) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right)\right],$$

$$S \lesssim D\left[\log^4(1/\varepsilon) + \log^4(MD) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right)\right],$$

and

$$\log B \lesssim \log(1/\varepsilon) + \log(MD) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

This proves the stated configuration. □

Proof of Lemma B.22. Set

$$C := DM\|a\|_\infty \vee \frac{1+\sigma^2}{\sigma^2}.$$

Since $M \geq 1$ and $0 < \sigma < 1$, we have $C \geq 1$. For $\|x\|_\infty \leq M$,

$$|x^\top a| \leq D\|x\|_\infty \|a\|_\infty \leq DM\|a\|_\infty \leq C.$$

Moreover, by (98) and $t \in [0,1]$,

$$0 \leq \frac{t}{q_t} \leq \frac{1}{q_t} \leq \frac{1+\sigma^2}{\sigma^2} \leq C.$$

Hence the two exact factors

$$x^\top a, \qquad \frac{t}{q_t}$$

belong to $[-C, C]$.

Lemma B.20, with $\gamma = 1$ and accuracy $\varepsilon/(4C)$, provides a network $\chi_{1,\varepsilon/(4C)}$ satisfying

$$\sup_{t\in[0,1]} \left| \chi_{1,\varepsilon/(4C)}(t) - \frac{t}{q_t} \right| \leq \frac{\varepsilon}{4C}.$$

The map

$$\ell_a(x) := x^\top a$$

is affine and is therefore represented exactly. Applying Lemma B.14 with $p = 2$, domain $[-C, C]^2$, approximation accuracy $\varepsilon/2$, and perturbation level $\varepsilon/(4C)$, define

$$\omega_\varepsilon(x,t) := \psi_{\mathrm{mult}}\left(\ell_a(x), \chi_{1,\varepsilon/(4C)}(t)\right).$$

The first input is exact, while the second satisfies

$$\left| \chi_{1,\varepsilon/(4C)}(t) - \frac{t}{q_t} \right| \leq \frac{\varepsilon}{4C}.$$

Therefore the perturbation of the two-dimensional input is bounded by $\varepsilon/(4C)$ in the ℓ_∞ norm, and Lemma B.14 yields

$$\left| \omega_\varepsilon(x,t) - \frac{t\, x^\top a}{q_t} \right| \leq \frac{\varepsilon}{2} + 2C\frac{\varepsilon}{4C} \quad = \varepsilon.$$

This proves the approximation claim.

The resulting network configuration is obtained as follows. The exact affine map $\ell_a(x) = x^\top a$ uses D coefficients and hence contributes

$$S_{\mathrm{lin}} \lesssim D, \qquad B_{\mathrm{lin}} \lesssim \|a\|_\infty \vee 1.$$

Moreover, since the full network takes $(x,t) \in \mathbb{R}^{D+1}$ as input, its width parameter satisfies a contribution of order $D+1$.

For the time branch, Lemma B.20 with accuracy $\varepsilon/(4C)$ gives

$$L_{\mathrm{time}} \lesssim \log^2(1/\varepsilon) + \log^2 C + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}_{\mathrm{time}}\|_\infty \lesssim \log^3(1/\varepsilon) + \log^3 C + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$S_{\mathrm{time}} \lesssim \log^4(1/\varepsilon) + \log^4 C + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

and

$$\log B_{\mathrm{time}} \lesssim \log(1/\varepsilon) + \log C + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

The linear branch has scalar output. If its depth is smaller than that of the time branch, we propagate this scalar through identity layers until the two branches have the same depth. This requires only

$O(L_{\text{time}})$ additional non-zero parameters and constant additional width, and is therefore absorbed by S_{time} and $\|\mathbf{W}_{\text{time}}\|_\infty$.

The outer multiplication network has constant hidden width and only first-order logarithmic depth and sparsity. Hence its contribution is also absorbed by the time-coefficient network. Combining all components first gives

$$L \lesssim \log^2(1/\varepsilon) + \log^2 C + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}\|_\infty \lesssim D + \log^3(1/\varepsilon) + \log^3 C + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$S \lesssim D + \log^4(1/\varepsilon) + \log^4 C + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

and

$$\log B \lesssim \log(1/\varepsilon) + \log C + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Finally, since

$$C = DM\|a\|_\infty \vee \frac{1+\sigma^2}{\sigma^2},$$

we have

$$\log C \lesssim \log(DM\|a\|_\infty \vee 1) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

Using the standing fixed-power logarithmic simplification therefore gives

$$L \lesssim \log^2(1/\varepsilon) + \log^2(DM\|a\|_\infty \vee 1) + \log^2\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$\|\mathbf{W}\|_\infty \lesssim D + \log^3(1/\varepsilon) + \log^3(DM\|a\|_\infty \vee 1) + \log^3\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

$$S \lesssim D + \log^4(1/\varepsilon) + \log^4(DM\|a\|_\infty \vee 1) + \log^4\left(\frac{1+\sigma^2}{\sigma^2}\right),$$

and

$$\log B \lesssim \log(1/\varepsilon) + \log(DM\|a\|_\infty \vee 1) + \log\left(\frac{1+\sigma^2}{\sigma^2}\right).$$

This proves the stated configuration. □

C Proofs for Generalization

We use the notation P, P_n, and Δ_v from Section 4. The main theorem proof is given first; the risk, truncation, entropy, and concentration lemmas used by that proof are proved in the following subsections.

C.1 Proof of the finite-sample theorem

Proof of Theorem 4.5.

Step 1: Risk representation and fixed-field localization. Lemma 4.1 identifies $P\Delta_v$ with the integrated $L^2(\pi_t)$ velocity error. Lemma 4.2 supplies both the localized second moment and the centered sub-exponential envelope, and hence (13) for every fixed deterministic field. This does not yet control the data-dependent ERM.

Step 2: Finite reduction on the unbounded domain. Apply Lemma 4.4 with confidence $\delta/2$. The bounded-domain ReLU entropy and Gaussian truncation give a deterministic family $v_1, \dots, v_M$ inside $\mathcal{V}(L, W, S, B, \gamma_x)$. On an event of probability at least $1 - \delta/2$, every field has one representative satisfying (15) simultaneously for P and P_n.

Step 3: Uniform oracle inequality. Lemma C.7 applies the fixed-field inequality to the deterministic representatives, takes a union bound, and transfers the result back to the full class. The ERM comparison then gives (17).

Step 4: Approximation–estimation balance. Set $\varepsilon := n^{-1/(2\beta+d)}$ and suppose that this choice satisfies the conditions of Theorem 3.1. Choose L, W, S, B and γ_x as in that theorem and set $\tau := \varepsilon^{2\beta}$. The comparator supplied there satisfies

$$\inf_{v\in\mathcal{V}(L,W,S,B,\gamma_x)} P\Delta_v \lesssim_{d,\beta} \frac{D(1+H^2)}{\sigma}\,\varepsilon^{2\beta}. \tag{103}$$

Its architecture obeys

$$L, \log B \lesssim_{d,\beta} \Gamma_\varepsilon^{c_{\mathrm{arch}}}, \qquad S \lesssim_{d,\beta} D^2\varepsilon^{-d}\Gamma_\varepsilon^{c_{\mathrm{arch}}}.$$

Substituting these bounds into (123) gives

$$\log\frac{8M}{\delta} \lesssim_{d,\beta} D^2\varepsilon^{-d}\big\{\Gamma_\varepsilon + \log(en) + \log(e/\delta)\big\}^{c_{\mathrm{arch}}}. \tag{104}$$

Indeed, $SL \lesssim_{d,\beta} D^2\varepsilon^{-d}\Gamma_\varepsilon^{c_{\mathrm{arch}}}$: here the fixed exponent is chosen after adding the individual, uncoarsened exponents from S, L, and the remaining logarithm in the ReLU covering bound. Combining (103), (104), and (17) yields

$$\begin{aligned} P\Delta_{\widehat{v}} \lesssim_{d,\beta} \frac{D(1+H^2)}{\sigma}\,\varepsilon^{2\beta} \\ + \frac{D^2\varepsilon^{-d}}{n}\left[\frac{1}{\sigma^2} + \left(\frac{1}{\sigma} + \sqrt{D}\right)\log(en)\right]\big\{\Gamma_\varepsilon + \log(en) + \log(e/\delta)\big\}^{c_{\mathrm{arch}}}. \end{aligned}$$

Since $\varepsilon^{2\beta} = \varepsilon^{-d}/n = n^{-2\beta/(2\beta+d)}$ and $\log(en) \lesssim_{d,\beta} \Gamma_\varepsilon$, this is (18). Lemma 4.1 converts $P\Delta_{\widehat{v}}$ to the asserted integrated velocity error. □

C.2 Risk identity and localized excess loss

Proof of Lemma 4.1. Fix $t \in [0,1]$ and write $Y = X_1 - X_0$. Since $v^*(X_t,t) = \mathbb{E}[Y \mid X_t]$,

$$\begin{aligned} \|Y - v(X_t,t)\|^2 = \|Y - v^*(X_t,t)\|^2 + \|v(X_t,t) - v^*(X_t,t)\|^2 \\ - 2\langle Y - v^*(X_t,t), v(X_t,t) - v^*(X_t,t)\rangle. \end{aligned}$$

The last factor in the cross term is measurable with respect to X_t, and

$$\mathbb{E}[Y - v^*(X_t,t) \mid X_t] = 0.$$

Hence the cross term has expectation zero. Taking expectations and integrating over $t \in [0,1]$ gives

$$\mathcal{L}(v) - \mathcal{L}(v^*) = \int_0^1 \mathbb{E}\|v(X_t,t) - v^*(X_t,t)\|^2 dt,$$

which is (8). □

For a fixed deterministic field, conditional label covariance and the bounded structured discrepancy provide the localization needed below.

Lemma C.1 (Conditional covariance of the Flow Matching label)**.** *For every fixed* $t \in [0,1]$,

$$\mathrm{Cov}(X_1 - X_0 \mid X_t) = \frac{\sigma^2}{q_t} I_D + b_t^2\,\mathrm{Cov}(g^*(U) \mid X_t). \tag{105}$$

Consequently,

$$\lambda_{\max}\big(\mathrm{Cov}(X_1 - X_0 \mid X_t)\big) \lesssim \sigma^{-2}, \qquad t \in [0,1]. \tag{106}$$

Proof. Fix $t \in [0,1]$. Conditional on U,

$$X_1 - X_0 = g^*(U) + \sigma\xi - X_0, \qquad X_t = tg^*(U) + (1-t)X_0 + t\sigma\xi.$$

Since X_0 and ξ are independent standard Gaussian vectors,

$$\mathrm{Var}(X_1 - X_0 \mid U) = (1+\sigma^2)I_D, \qquad \mathrm{Var}(X_t \mid U) = q_t I_D,$$

and

$$\mathrm{Cov}(X_1 - X_0, X_t \mid U) = \big(t\sigma^2 - (1-t)\big)I_D.$$

Therefore, Gaussian conditioning gives

$$\begin{aligned}\mathrm{Cov}(X_1 - X_0 \mid X_t, U) &= \left[1 + \sigma^2 - \frac{(t\sigma^2 - (1-t))^2}{q_t}\right] I_D \\ &= \frac{\sigma^2}{q_t} I_D,\end{aligned}$$

where we used

$$(1+\sigma^2)q_t - \big(t\sigma^2 - (1-t)\big)^2 = \sigma^2.$$

The corresponding conditional mean is

$$\begin{aligned}\mathbb{E}[X_1 - X_0 \mid X_t, U] &= g^*(U) + \frac{t\sigma^2 - (1-t)}{q_t}\big(X_t - tg^*(U)\big) \\ &= a_t X_t + (1 - ta_t)\, g^*(U) \\ &= a_t X_t + b_t g^*(U),\end{aligned}$$

where $1 - ta_t = b_t$.

Applying the conditional law of total covariance from (X_t, U) to X_t yields

$$\begin{aligned}\mathrm{Cov}(X_1 - X_0 \mid X_t) &= \mathbb{E}\left[\mathrm{Cov}(X_1 - X_0 \mid X_t, U) \,|\, X_t\right] \\ &\quad + \mathrm{Cov}\left(\mathbb{E}[X_1 - X_0 \mid X_t, U] \,|\, X_t\right) \\ &= \frac{\sigma^2}{q_t} I_D + b_t^2\, \mathrm{Cov}(g^*(U) \mid X_t),\end{aligned}$$

which proves (105).

For the operator-norm bound, $\|g^*(U)\| \le 1$ gives, for every unit vector $u \in \mathbb{R}^D$,

$$\begin{aligned}u^\top \mathrm{Cov}(g^*(U) \mid X_t) u &= \mathrm{Var}\big(u^\top g^*(U) \mid X_t\big) \\ &\le \mathbb{E}\left[(u^\top g^*(U))^2 \,\middle|\, X_t\right] \le 1.\end{aligned}$$

Hence

$$\lambda_{\max}\big(\mathrm{Cov}(g^*(U) \mid X_t)\big) \le 1.$$

By the lower bound on q_t stated in Section 2, and since $0 < \sigma < 1$,

$$\frac{\sigma^2}{q_t} \le 1 + \sigma^2 \le 2.$$

Also,

$$b_t^2 = \frac{(1-t)^2}{q_t^2} \le \frac{1}{q_t} \le \frac{1+\sigma^2}{\sigma^2} \le \frac{2}{\sigma^2}.$$

Combining these estimates with (105) proves (106). □

The covariance bound controls the stochastic label. To convert it into a variance-to-risk relation, we also need a uniform bound on the discrepancy between a structured field and the population velocity.

Lemma C.2 (Uniform discrepancy of structured velocity fields)**.** *For every $v \in \mathcal{V}(L, W, S, B, \gamma_x)$,*

$$\|v(x,t) - v^*(x,t)\|^2 \lesssim \sigma^{-2}, \qquad (x,t) \in \mathbb{R}^D \times [0,1]. \tag{107}$$

Proof. For $v = v_\theta \in \mathcal{V}(L, W, S, B, \gamma_x)$,

$$v(x,t) = a_t x + b_t \operatorname{clip}_2(f_\theta(x,t)),$$

whereas

$$v^*(x,t) = a_t x + b_t m_t(x).$$

Hence the common affine term cancels:

$$v(x,t) - v^*(x,t) = b_t \left[\operatorname{clip}_2(f_\theta(x,t)) - m_t(x)\right].$$

Since

$$\|\operatorname{clip}_2(f_\theta(x,t))\| \le 2$$

and

$$\|m_t(x)\| = \|\mathbb{E}[g^*(U) \mid X_t = x]\| \le \mathbb{E}[\|g^*(U)\| \mid X_t = x] \le 1,$$

we obtain

$$\|v(x,t) - v^*(x,t)\| \le 3|b_t|.$$

Using

$$b_t^2 \le \frac{2}{\sigma^2}$$

from the proof of Lemma C.1 gives

$$\|v(x,t) - v^*(x,t)\|^2 \le \frac{18}{\sigma^2},$$

which proves (107). □

The preceding two bounds yield the localized second-moment estimate for the target-indexed excess loss.

Lemma C.3 (Bernstein condition for the Flow Matching excess loss)**.** *Under the standing assumptions, for every $v \in \mathcal{V}(L, W, S, B, \gamma_x)$,*

$$\mathbb{E}_{X_1}\left[\Delta_v(X_1)^2\right] \lesssim \frac{1}{\sigma^2}\mathbb{E}_{X_1}\Delta_v(X_1). \tag{108}$$

The hidden constant is absolute and, in particular, independent of D, L, W, S, B.

Proof of Lemma C.3. By the definition of Δ_v and Jensen's inequality with respect to the probability measure $dt \otimes P_{X_0}$,

$$\begin{aligned}\mathbb{E}_{X_1}\Delta_v(X_1)^2 \le \int_0^1 \mathbb{E}\Big[&\|X_1 - X_0 - v(X_t,t)\|^2 \\ &- \|X_1 - X_0 - v^*(X_t,t)\|^2\Big]^2 dt.\end{aligned} \tag{109}$$

Fix $t \in [0,1]$. Expanding the squared-loss difference gives

$$\begin{aligned}&\|X_1 - X_0 - v(X_t,t)\|^2 - \|X_1 - X_0 - v^*(X_t,t)\|^2 \\ &\quad = \|v(X_t,t) - v^*(X_t,t)\|^2 - 2\left\langle X_1 - X_0 - v^*(X_t,t), v(X_t,t) - v^*(X_t,t)\right\rangle.\end{aligned}$$

Since

$$v^*(X_t,t) = \mathbb{E}[X_1 - X_0 \mid X_t],$$

the tower property yields

$$\begin{aligned}&\mathbb{E}\Big[\|X_1 - X_0 - v(X_t,t)\|^2 - \|X_1 - X_0 - v^*(X_t,t)\|^2\Big]^2 \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\|X_1 - X_0 - v(X_t,t)\|^2 - \|X_1 - X_0 - v^*(X_t,t)\|^2\right)^2 \middle| X_t\right]\right] \\ &= \mathbb{E}\|v(X_t,t) - v^*(X_t,t)\|^4 \\ &\quad + 4\mathbb{E}\Big[\big(v(X_t,t) - v^*(X_t,t)\big)^\top \operatorname{Cov}(X_1 - X_0 \mid X_t)\big(v(X_t,t) - v^*(X_t,t)\big)\Big].\end{aligned}$$

Lemma C.2 gives

$$\|v(X_t,t)-v^*(X_t,t)\|^4 \lesssim \frac{1}{\sigma^2}\|v(X_t,t)-v^*(X_t,t)\|^2,$$

while Lemma C.1 gives

$$\begin{aligned}&\big(v(X_t,t)-v^*(X_t,t)\big)^\top \mathrm{Cov}(X_1-X_0\mid X_t)\big(v(X_t,t)-v^*(X_t,t)\big)\\ &\lesssim \frac{1}{\sigma^2}\|v(X_t,t)-v^*(X_t,t)\|^2.\end{aligned}$$

Hence, for every fixed $t\in[0,1]$,

$$\begin{aligned}&\mathbb{E}\Big[\|X_1-X_0-v(X_t,t)\|^2-\|X_1-X_0-v^*(X_t,t)\|^2\Big]^2\\ &\lesssim \frac{1}{\sigma^2}\mathbb{E}\|v(X_t,t)-v^*(X_t,t)\|^2.\end{aligned} \tag{110}$$

Combining (109) and (110), and then using the population-risk identity, gives

$$\begin{aligned}\mathbb{E}_{X_1}\Delta_v(X_1)^2 &\lesssim \frac{1}{\sigma^2}\int_0^1 \mathbb{E}\|v(X_t,t)-v^*(X_t,t)\|^2dt\\ &= \frac{1}{\sigma^2}\big[\mathcal{L}(v)-\mathcal{L}(v^*)\big]\\ &= \frac{1}{\sigma^2}\mathbb{E}_{X_1}\Delta_v(X_1).\end{aligned}$$

This proves (108). □

The second moment controls local variance, while fixed-field concentration also requires an envelope for the unbounded target observation.

Lemma C.4 (Uniform tail bound for the Flow Matching excess loss)**.** *For every $v=v_\theta\in\mathcal{V}(L,W,S,B,\gamma_x)$ and every $x_1\in\mathbb{R}^D$,*

$$|\Delta_v(x_1)|\le \frac{3\pi}{2\sigma}\|x_1\|+\frac{9\pi}{4\sigma}+3\sqrt{D}. \tag{111}$$

Consequently,

$$\sup_{v\in\mathcal{V}(L,W,S,B,\gamma_x)}\|\Delta_v(X_1)\|_{\psi_1}\lesssim \frac{1}{\sigma}+\sqrt{D}, \tag{112}$$

and hence

$$\sup_{v\in\mathcal{V}(L,W,S,B,\gamma_x)}\|\Delta_v(X_1)-\mathbb{E}_{X_1}\Delta_v(X_1)\|_{\psi_1}\lesssim \frac{1}{\sigma}+\sqrt{D}. \tag{113}$$

The hidden constants are absolute.

The raw FM square loss is unbounded, but the common affine velocity component cancels in the excess loss and clipping controls the nonlinear discrepancy.

Proof of Lemma C.4. Fix $v=v_\theta\in\mathcal{V}(L,W,S,B,\gamma_x)$ and $x_1\in\mathbb{R}^D$. Within this proof, write

$$\bar f_\theta(x,t):=\mathrm{clip}_2(f_\theta(x,t)).$$

Then $\|\bar f_\theta(x,t)\|\le 2$, while the representation of the population velocity implies $\|m_t(x)\|\le 1$. Hence

$$\|\bar f_\theta(x,t)-m_t(x)\|\le 3. \tag{114}$$

With X_1 fixed at x_1, write $X_t=(1-t)X_0+tx_1$. Since

$$1-ta_t=b_t,\qquad 1+(1-t)a_t=\frac{\sigma^2 t}{q_t},$$

the residuals corresponding to v_θ and v^* satisfy

$$x_1 - X_0 - v_\theta(X_t,t) = b_t\big(x_1 - \bar{f}_\theta(X_t,t)\big) - \frac{\sigma^2 t}{q_t}X_0,$$
$$x_1 - X_0 - v^*(X_t,t) = b_t\big(x_1 - m_t(X_t)\big) - \frac{\sigma^2 t}{q_t}X_0.$$

Subtracting the corresponding squared norms gives

$$\begin{aligned}&\|x_1 - X_0 - v_\theta(X_t,t)\|^2 - \|x_1 - X_0 - v^*(X_t,t)\|^2\\ &= b_t^2\left(\|x_1 - \bar{f}_\theta(X_t,t)\|^2 - \|x_1 - m_t(X_t)\|^2\right)\\ &\quad + 2b_t\frac{\sigma^2 t}{q_t}\left\langle X_0, \bar{f}_\theta(X_t,t) - m_t(X_t)\right\rangle.\end{aligned} \tag{115}$$

Using

$$|\|a\|^2 - \|b\|^2| \le \|a-b\|(\|a\| + \|b\|)$$

together with (114), we obtain

$$\begin{aligned}&\left|\|x_1 - \bar{f}_\theta(X_t,t)\|^2 - \|x_1 - m_t(X_t)\|^2\right|\\ &\quad\le \|\bar{f}_\theta(X_t,t) - m_t(X_t)\|\left(2\|x_1\| + \|\bar{f}_\theta(X_t,t)\| + \|m_t(X_t)\|\right)\\ &\quad\le 6\|x_1\| + 9.\end{aligned}$$

The second term in (115) is bounded by

$$2b_t\frac{\sigma^2 t}{q_t}\|X_0\|\|\bar{f}_\theta(X_t,t) - m_t(X_t)\| \le 6b_t\frac{\sigma^2 t}{q_t}\|X_0\|.$$

Therefore, by the definition of Δ_v,

$$|\Delta_v(x_1)| \le (6\|x_1\| + 9)\int_0^1 b_t^2 dt + 6\mathbb{E}\|X_0\|\int_0^1 b_t\frac{\sigma^2 t}{q_t}dt. \tag{116}$$

Using $q_t = (1-t)^2 + \sigma^2 t^2$ and the change of variables $r = t/(1-t)$,

$$\int_0^1 b_t^2 dt = \int_0^\infty \frac{dr}{(1+\sigma^2 r^2)^2} = \frac{\pi}{4\sigma},$$
$$\int_0^1 b_t\frac{\sigma^2 t}{q_t}dt = \int_0^\infty \frac{\sigma^2 r}{(1+\sigma^2 r^2)^2}dr = \frac{1}{2}.$$

Since $X_0 \sim N(0, I_D)$,

$$\mathbb{E}\|X_0\| \le \sqrt{\mathbb{E}\|X_0\|^2} = \sqrt{D}.$$

Substituting these estimates into (116) yields

$$|\Delta_v(x_1)| \le \frac{3\pi}{2\sigma}\|x_1\| + \frac{9\pi}{4\sigma} + 3\sqrt{D},$$

which proves (111).

For the Orlicz norm under the target distribution, recall that

$$X_1 = g^*(U) + \sigma\xi, \qquad \xi \sim N(0, I_D),$$

with $\|g^*(U)\| \le 1$. Hence

$$\|X_1\| \le 1 + \sigma\|\xi\|.$$

Since the Euclidean norm of a standard Gaussian vector satisfies

$$\big\|\|\xi\|\big\|_{\psi_1} \lesssim \sqrt{D},$$

(111) and the triangle inequality for the ψ_1 norm imply

$$\begin{aligned}\|\Delta_v(X_1)\|_{\psi_1} &\lesssim \frac{1}{\sigma}\left(1 + \big\|\|X_1\|\big\|_{\psi_1}\right) + \sqrt{D} \\ &\lesssim \frac{1}{\sigma}\left(1 + \sigma\sqrt{D}\right) + \frac{1}{\sigma} + \sqrt{D} \\ &\lesssim \frac{1}{\sigma} + \sqrt{D}.\end{aligned}$$

The bound is independent of v, which proves (112). Finally, the standard centering inequality for the sub-exponential Orlicz norm gives

$$\|\Delta_v(X_1) - \mathbb{E}_{X_1}\Delta_v(X_1)\|_{\psi_1} \lesssim \|\Delta_v(X_1)\|_{\psi_1},$$

and (113) follows. □

Combining the localized second moment with this tail envelope gives the fixed-field Bernstein inequality stated in the main text.

Proof of Lemma 4.2. Lemma C.3 and Lemma C.4 give the two bounds in (12). For a fixed deterministic v, Lemma C.4 and the maximal inequality for sub-exponential variables give

$$\left\| \max_{1\le i\le n} |\Delta_v(X_{1,i}) - P\Delta_v| \right\|_{\psi_1} \lesssim (\sigma^{-1} + \sqrt{D})\log(en).$$

Applying the Bernstein inequality for unbounded variables (Lecué & Mendelson, 2012, Proposition 5.2) to both signs, and using (108), proves (13). □

C.3 Truncation, local loss stability, and finite reduction

Lemma C.5 (Local stability of the Flow Matching loss)**.** *Let $v_1, v_2 \in \mathcal{V}(L, W, S, B, \gamma_x)$ be induced by $f_{\theta_1}, f_{\theta_2} \in \mathrm{NN}(L, W, S, B)$ according to the definition in* (4)*. Let $R_0, R_1 \ge 1$ and set $R := R_0 \vee R_1$. Suppose that*

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty(\mathcal{B}(0,R)\times[0,1];\ell_2)} \le \eta.$$

Then, for every $x_1 \in \mathcal{B}(0, R_1)$,

$$\begin{aligned}|\ell(x_1, v_1) - \ell(x_1, v_2)| \le \eta\left[\frac{\pi}{2\sigma}(R_1+2) + \sqrt{D}\right] \\ + \frac{2\pi}{\sigma}(R_1+2)\Pr(\|X_0\| > R_0) + 4\sqrt{D\Pr(\|X_0\| > R_0)}.\end{aligned} \tag{117}$$

Moreover,

$$\begin{aligned}|P(\Delta_{v_1} - \Delta_{v_2})| \lesssim \left(\frac{1}{\sigma} + \sqrt{D}\right)\Big[\eta + \sqrt{\Pr(\|X_0\| > R_0)} \\ + \sqrt{\Pr(\|X_1\| > R_1)}\Big].\end{aligned} \tag{118}$$

The hidden constant is absolute.

Proof. By (4),

$$v_j(x,t) = a_t x + b_t \,\mathrm{clip}_2(f_{\theta_j}(x,t)), \qquad j \in \{1,2\}.$$

Fix $x_1 \in \mathbb{R}^D$ and write

$$X_t = (1-t)X_0 + tx_1.$$

Using

$$1 - ta_t = b_t, \qquad 1 + (1-t)a_t = \frac{\sigma^2 t}{q_t},$$

we obtain

$$x_1 - X_0 - v_j(X_t, t) = b_t\left[x_1 - \mathrm{clip}_2(f_{\theta_j}(X_t, t))\right] - \frac{\sigma^2 t}{q_t}X_0.$$

Therefore,

$$\begin{aligned}&\left|\|x_1 - X_0 - v_1(X_t,t)\|^2 - \|x_1 - X_0 - v_2(X_t,t)\|^2\right| \\ &\le \left[2b_t^2(\|x_1\|+2) + 2b_t\frac{\sigma^2 t}{q_t}\|X_0\|\right] \\ &\qquad \times \|\mathrm{clip}_2(f_{\theta_1}(X_t,t)) - \mathrm{clip}_2(f_{\theta_2}(X_t,t))\|\,.\end{aligned} \tag{119}$$

The clipping map is nonexpansive and takes values in $\mathcal{B}(0,2)$. If $\|x_1\| \le R_1$ and $\|X_0\| \le R_0$, then $\|X_t\| \le R$, so the last factor in (119) is at most η; globally it is bounded by 4. Using

$$\int_0^1 b_t^2\,dt = \frac{\pi}{4\sigma}, \qquad \int_0^1 b_t\frac{\sigma^2 t}{q_t}\,dt = \frac{1}{2},$$

together with

$$\mathbb{E}\|X_0\| \le \sqrt{D}, \qquad \mathbb{E}\big[\|X_0\|\mathbf{1}_{\{\|X_0\|>R_0\}}\big] \le \sqrt{D\Pr(\|X_0\|>R_0)},$$

gives (117).

For the population bound, let

$$A_0 := \{\|X_0\| > R_0\}, \qquad A_1 := \{\|X_1\| > R_1\}.$$

On $A_0^c \cap A_1^c$, $X_t \in \mathcal{B}(0,R)$ for every $t \in [0,1]$, while globally the clipped difference is at most 4. Hence

$$\begin{aligned}&\|\mathrm{clip}_2(f_{\theta_1}(X_t,t)) - \mathrm{clip}_2(f_{\theta_2}(X_t,t))\| \\ &\le \eta + 4\mathbf{1}_{A_0} + 4\mathbf{1}_{A_1}.\end{aligned}$$

Substituting this bound into (119), integrating over t, and taking expectation gives

$$\begin{aligned}|P(\Delta_{v_1} - \Delta_{v_2})| \le \eta&\left[\frac{\pi}{2\sigma}\big(\mathbb{E}\|X_1\|+2\big) + \mathbb{E}\|X_0\|\right] \\ &+ \frac{2\pi}{\sigma}\big(\mathbb{E}\|X_1\|+2\big)\Pr(A_0) + 4\mathbb{E}[\|X_0\|\mathbf{1}_{A_0}] \\ &+ \frac{2\pi}{\sigma}\mathbb{E}[(\|X_1\|+2)\mathbf{1}_{A_1}] + 4\mathbb{E}\|X_0\|\Pr(A_1),\end{aligned}$$

where independence of X_0 and X_1 is used in the cross terms. Since

$$\mathbb{E}\|X_0\| \le \sqrt{D}, \qquad \mathbb{E}\|X_1\|^2 \le 1 + \sigma^2 D,$$

Cauchy–Schwarz yields

$$\mathbb{E}[(\|X_1\|+2)\mathbf{1}_{A_1}] \lesssim (1+\sigma\sqrt{D})\sqrt{\Pr(A_1)}.$$

Together with

$$\mathbb{E}[\|X_0\|\mathbf{1}_{A_0}] \le \sqrt{D\Pr(A_0)},$$

this proves (118). □

The local stability estimate transfers posterior-network approximation to conditional-loss approximation. We next quantify the bounded-domain network cover used in that transfer.

Lemma C.6 (Metric entropy on a bounded state-time domain)**.** *Let $R \ge 1$. For every $\eta \in (0,1)$, there exists a finite internal net*

$$\mathcal{F}_\eta \subset \mathrm{NN}(L,W,S,B)$$

such that

$$\sup_{f_\theta \in \mathrm{NN}(L,W,S,B)} \inf_{f \in \mathcal{F}_\eta} \|f_\theta - f\|_{L^\infty(\mathcal{B}(0,R)\times[0,1];\ell_2)} \le \eta$$

and

$$\log|\mathcal{F}_\eta| \lesssim SL\log\left(\frac{\sqrt{D}\,L(W+1)(B\vee 1)(R+1)}{\eta}\right). \tag{120}$$

Consequently,

$$\begin{aligned}&\log\mathcal{N}\left(\eta, \mathrm{NN}(L,W,S,B), \|\cdot\|_{L^\infty(\mathcal{B}(0,R)\times[0,1];\ell_2)}\right) \\ &\lesssim SL\log\left(\frac{\sqrt{D}\,L(W+1)(B\vee 1)(R+1)}{\eta}\right).\end{aligned} \tag{121}$$

Proof of Lemma C.6. We adapt the parameter-discretization argument of (Suzuki, 2019, Lemma 3) to D-dimensional outputs equipped with the Euclidean norm. Zero-pad to the maximal width-W architecture, fix a support J with $|J| \le S$, and let θ, θ' be supported on J with $\|\theta - \theta'\|_\infty \le h$. For $M := (W+1)(B \vee 1)$, the affine recursion and the 1-Lipschitz property of ReLU give

$$H_\ell \le (R+1)M^\ell, \qquad E_\ell \le ME_{\ell-1} + h(R+1)M^\ell \le \ell(R+1)M^\ell h.$$

Consequently,

$$\|f_\theta - f_{\theta'}\|_{L^\infty(\mathcal{B}(0,R)\times[0,1];\ell_2)} \lesssim \sqrt{D}\, L(R+1)M^L h.$$

Choose $h \asymp \eta/[\sqrt{D}\, L(R+1)M^L]$. Quantizing each active coordinate in $[-B, B]$ at mesh h, while leaving inactive coordinates zero, produces an internal net. The maximal architecture has at most $CL(W+1)^2$ parameter locations, so

$$|\mathcal{F}_\eta| \le \left[CL(W+1)^2\left(1 + \frac{B}{h}\right)\right]^S.$$

Taking logarithms and substituting h yields

$$\log|\mathcal{F}_\eta| \lesssim SL\log\left(\frac{\sqrt{D}\, L(W+1)(B\vee 1)(R+1)}{\eta}\right),$$

which proves (120) and (121). □

The ambient cover is converted into constrained representatives only after the truncation scales have been fixed.

Proof of Lemma 4.4. Fix $\tau, \delta \in (0,1)$ and introduce the proof-local scales

$$\Lambda_{n,\delta,\tau} := \frac{1}{\sigma} + \sqrt{D} + \sqrt{\log\frac{en}{\delta}} + \sqrt{\log\frac{e}{\tau}}, \qquad \eta := \frac{c_0\tau}{\Lambda_{n,\delta,\tau}},$$

$$R_0 := \sqrt{D} + \sqrt{2\log(1/\eta^2)}, \qquad R_1 := 1 + \sigma\sqrt{D} + \sigma\sqrt{2\log\frac{n}{\delta\eta^2}}, \qquad R := R_0 \vee R_1,$$

where $c_0 > 0$ is a sufficiently small absolute constant. Gaussian concentration and $\|g^*\|_{L^\infty} \le 1$ imply

$$\Pr(\|X_0\| > R_0) \le \eta^2, \qquad \Pr(\|X_1\| > R_1) \le \frac{\delta\eta^2}{n},$$

and therefore

$$\Pr\left(\max_{i\le n}\|X_{1,i}\| \le R_1\right) \ge 1 - \delta. \tag{122}$$

Let $\mathcal{F}_\mathcal{V}$ be the posterior-network realizations whose induced velocities belong to $\mathcal{V}(L, W, S, B, \gamma_x)$. Lemma C.6 supplies an ambient η-net on $\mathcal{B}(0,R)\times[0,1]$. Retain the balls that intersect $\mathcal{F}_\mathcal{V}$ and choose an actual constrained realization $\widetilde{f}_j$ from each retained ball. The corresponding fields

$$v_j(x,t) := a_t x + b_t\,\mathrm{clip}_2(\widetilde{f}_j(x,t))$$

are deterministic members of $\mathcal{V}(L, W, S, B, \gamma_x)$, and their number $M_{\tau,\delta}$ is no larger than the ambient covering number. Every constrained realization has a selected representative with internal posterior-network distance at most 2η on the truncated state–time domain.

Since $(R_1+1)/\sigma + \sqrt{D} \lesssim \Lambda_{n,\delta,\tau}$, Lemma C.5, the displayed tail bounds, and a sufficiently small c_0 give

$$\|\Delta_v - \Delta_{v_j}\|_{L^\infty(\mathcal{B}(0,R_1))} \le \frac{\tau}{4}, \qquad |P(\Delta_v - \Delta_{v_j})| \le \frac{3\tau}{4}.$$

On the event in (122), the same selected field v_j therefore satisfies

$$|P(\Delta_v - \Delta_{v_j})| + |P_n(\Delta_v - \Delta_{v_j})| \le \tau,$$

which proves (15).

Finally, set

$$\mathfrak{R}_{n,\delta,\tau} := 2 + \sqrt{D} + \sqrt{\log \frac{en\Lambda_{n,\delta,\tau}^2}{\delta\tau^2}}.$$

Then $R + 1 \lesssim \mathfrak{R}_{n,\delta,\tau}$, and (120) gives

$$\log M_{\tau,\delta} \lesssim SL \log\left(\frac{\sqrt{D}\, L(W+1)(B \vee 1)\Lambda_{n,\delta,\tau}\mathfrak{R}_{n,\delta,\tau}}{\tau}\right). \tag{123}$$

Moreover,

$$\Lambda_{n,\delta,\tau} \lesssim 1 + \frac{1}{\sigma} + \sqrt{D} + \sqrt{\log \frac{en}{\delta}} + \sqrt{\log \frac{e}{\tau}}, \qquad \mathfrak{R}_{n,\delta,\tau} \lesssim 1 + \frac{1}{\sigma} + \sqrt{D} + \sqrt{\log \frac{en}{\delta}} + \sqrt{\log \frac{e}{\tau}}.$$

Substitution into (123) proves the explicit bound (16) stated in the main text. □

C.4 Uniform concentration

Lemma C.7 (Uniform localized concentration of the excess loss)**.** *For every $\tau, \delta \in (0,1)$, let $v_1, \dots, v_M$ be the deterministic family supplied by Lemma 4.4 with confidence parameter $\delta/2$. With probability at least $1 - \delta$,*

$$|(P - P_n)\Delta_v| \le \frac{1}{4}P\Delta_v + C\left\{\left[\frac{1}{\sigma^2} + \left(\frac{1}{\sigma} + \sqrt{D}\right)\log(en)\right]\frac{\log(8M/\delta)}{n} + \tau\right\} \tag{124}$$

simultaneously for all $v \in \mathcal{V}(L, W, S, B, \gamma_x)$, where $C > 0$ is an absolute constant.

Proof. Use the deterministic family supplied by Lemma 4.4 with confidence parameter $\delta/2$. On an event of probability at least $1 - \delta/2$, every v has a representative v_k satisfying (15). Put $h := \log(8M/\delta)$. The fixed-field estimate (13) and a union bound give, on another event of probability at least $1 - \delta/2$,

$$|(P - P_n)\Delta_{v_k}| \le C_{\mathrm{B}}\sqrt{\frac{P\Delta_{v_k}h}{\sigma^2 n}} + C_{\mathrm{B}}\frac{(\sigma^{-1} + \sqrt{D})\log(en)h}{n}$$

for every representative. On the intersection of these events, $P\Delta_{v_k} \le P\Delta_v + \tau$ and transfer through the selected constrained representative gives

$$|(P - P_n)\Delta_v| \le C_{\mathrm{B}}\sqrt{\frac{(P\Delta_v + \tau)h}{\sigma^2 n}} + C_{\mathrm{B}}\frac{(\sigma^{-1} + \sqrt{D})\log(en)h}{n} + \tau.$$

The inequalities $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and Young's inequality prove (124). □

Proof of the oracle inequality (17). Use Lemma 4.4 with confidence $\delta/2$ and write $M = M_{\tau,\delta/2}$. For every deterministic comparator $\bar{v} \in \mathcal{V}(L, W, S, B, \gamma_x)$, the ERM property gives

$$P\Delta_{\widehat{v}} \le P\Delta_{\bar{v}} + (P - P_n)\Delta_{\widehat{v}} + (P_n - P)\Delta_{\bar{v}}.$$

On the simultaneous event from Lemma C.7, apply (124) to both fields and absorb the $\frac{1}{4}P\Delta_{\widehat{v}}$ term. The remaining $\frac{1}{4}P\Delta_{\bar{v}}$ term is absorbed into the universal multiplicative constant. Taking the infimum over $\bar{v}$ proves (17). □

D Auxiliary Results for Distributional Convergence

This appendix collects the auxiliary growth estimate used in the sampling analysis and proves the Euler discretization bound in Proposition 5.4.

D.1 LINEAR-GROWTH BOUND

Lemma D.1 (Linear growth of the velocity fields)**.** *Let*

$$K_\sigma^{\mathrm{sam}} := \max\left\{ \sup_{t\in[0,1]} |a_t|, 2 \sup_{t\in[0,1]} |b_t| \right\}.$$

Then, for every $(x,t) \in \mathbb{R}^D \times [0,1]$*,*

$$\|v^*(x,t)\| \le K_\sigma^{\mathrm{sam}}(1+\|x\|),$$

and, for every $v \in \mathcal{V}(L,W,S,B,\gamma_x)$*,*

$$\|v(x,t)\| \le K_\sigma^{\mathrm{sam}}(1+\|x\|).$$

Proof. By Lemma 2.1,

$$v^*(x,t) = a_t x + b_t m_t(x), \qquad \|m_t(x)\| \le 1,$$

and therefore

$$\|v^*(x,t)\| \le |a_t|\|x\| + |b_t|.$$

For $v \in \mathcal{V}(L,W,S,B,\gamma_x)$, the radius-two clipping in the definition of the velocity class gives

$$\|v(x,t)\| \le |a_t|\|x\| + 2|b_t|.$$

Both bounds are dominated by $K_\sigma^{\mathrm{sam}}(1+\|x\|)$. □

D.2 EULER DISCRETIZATION

Proof of Proposition 5.4. Condition on the target observations. Couple the continuous learned trajectory and the Euler scheme through the same initial condition,

$$\widehat{Z}_0 = \widetilde{Z}_{t_0} \sim \pi_0.$$

By Lemma D.1 and Grönwall's inequality,

$$1 + \|\widehat{Z}_t\| \le e^{K_\sigma^{\mathrm{sam}} t}(1+\|\widehat{Z}_0\|) \le e^{K_\sigma^{\mathrm{sam}}}(1+\|\widehat{Z}_0\|), \qquad t \in [0,1].$$

Consequently, for $t \in [t_k, t_{k+1}]$,

$$\begin{aligned}\|\widehat{Z}_t - \widehat{Z}_{t_k}\| &\le \int_{t_k}^{t} \|\widehat{v}(\widehat{Z}_s, s)\|\, ds \\ &\le K_\sigma^{\mathrm{sam}} e^{K_\sigma^{\mathrm{sam}}}(1+\|\widehat{Z}_0\|)(t-t_k).\end{aligned}$$

Write $h_k := t_{k+1} - t_k$. The continuous learned trajectory satisfies

$$\widehat{Z}_{t_{k+1}} = \widehat{Z}_{t_k} + h_k \widehat{v}(\widehat{Z}_{t_k}, t_k) + r_k,$$

where

$$r_k := \int_{t_k}^{t_{k+1}} \Big[\widehat{v}(\widehat{Z}_t, t) - \widehat{v}(\widehat{Z}_{t_k}, t_k)\Big] dt.$$

Using the spatial class constraint and Assumption 5.3,

$$\begin{aligned}\|\widehat{v}(\widehat{Z}_t, t) - \widehat{v}(\widehat{Z}_{t_k}, t_k)\| \le \bar{\gamma}_x \|\widehat{Z}_t - \widehat{Z}_{t_k}\| \\ + \gamma_t(1+\|\widehat{Z}_{t_k}\|)(t-t_k).\end{aligned}$$

The preceding growth estimates therefore yield

$$\|r_k\| \le \frac{1}{2} e^{K_\sigma^{\mathrm{sam}}} \big(\bar{\gamma}_x K_\sigma^{\mathrm{sam}} + \gamma_t\big)(1+\|\widehat{Z}_0\|) h_k^2.$$

Subtracting the Euler recursion (20) gives

$$\|\widehat{Z}_{t_{k+1}} - \widetilde{Z}_{t_{k+1}}\| \le (1+\bar{\gamma}_x h_k)\|\widehat{Z}_{t_k} - \widetilde{Z}_{t_k}\| + \|r_k\|.$$

Iterating this inequality and using $1 + u \le e^u$ gives

$$\|\widehat{Z}_1 - \widetilde{Z}_{t_M}\| \le \frac{1}{2} e^{K_\sigma^{\mathrm{sam}} + \bar{\gamma}_x} \big(\bar{\gamma}_x K_\sigma^{\mathrm{sam}} + \gamma_t\big)(1 + \|\widehat{Z}_0\|) \sum_{k=0}^{M-1} h_k^2.$$

Since $\sum_{k=0}^{M-1} h_k = 1$ and $h_k \le h$,

$$\sum_{k=0}^{M-1} h_k^2 \le h.$$

Moreover, $\widehat{Z}_0 \sim \mathcal{N}(0, I_D)$, so

$$\|1 + \|\widehat{Z}_0\|\|_{L^2} \le 1 + \sqrt{\mathbb{E}\|\widehat{Z}_0\|^2} = 1 + \sqrt{D}.$$

The coupling of the continuous learned flow and the Euler scheme therefore gives

$$W_2(\widetilde{\pi}_1, \widehat{\pi}_1) \le \frac{1}{2}(1 + \sqrt{D}) e^{K_\sigma^{\mathrm{sam}} + \bar{\gamma}_x} \big(\bar{\gamma}_x K_\sigma^{\mathrm{sam}} + \gamma_t\big) h.$$

Finally, the triangle inequality gives

$$W_2(\widetilde{\pi}_1, \pi_1) \le W_2(\widetilde{\pi}_1, \widehat{\pi}_1) + W_2(\widehat{\pi}_1, \pi_1).$$

Applying the continuous-flow bound (19) to the second term yields

$$W_2(\widehat{\pi}_1, \pi_1) \le \exp\left\{\frac{1}{2} + \|\gamma_x\|_{L^1(0,1)}\right\} \left(\int_0^1 \|\widehat{v}(\cdot, t) - v^*(\cdot, t)\|_{L^2(\pi_t)}^2 \, dt\right)^{1/2}.$$

Combining the last two displays proves (21). □