Intrinsic Wasserstein Rates for Score-Based Generative Models on Smooth Manifolds

Anonymous Authors[1]

Abstract

Score-based generative models are trained in high-dimensional ambient spaces, yet many data distributions are supported on low-dimensional nonlinear structures. We prove that, for compact d-dimensional smooth manifolds $\mathcal{M} \subset [0,1]^D$ with $d > 2$ and β-Hölder densities strictly positive on $\mathcal{M}$, a variance-preserving SGM estimator attains the intrinsic Wasserstein–1 sample exponent $\tilde{\mathcal{O}}(D^{\mathcal{O}_\beta(d)} n^{-(\beta+1)/(d+2\beta)})$, up to logarithmic factors and explicit geometry and density factors. The full nonasymptotic bound explicitly isolates the finite-order geometry envelope, Hölder radius, density lower bound, ambient dependence, and finite-order correction terms. The analysis separates score approximation into a large-noise tangent-cell regime and a small-noise projection-centered, de-Gaussianized Laplace regime. The key technical ingredient is a ReLU implementation of nearest-projection coordinates via finite intrinsic anchors and Gauss–Newton iterations, avoiding black-box high-order D-variate approximation of $\Pi_{\mathcal{M}}$. Consequently, for families with polynomially controlled geometry and density lower bounds, the constructed score-network parameters have polynomial ambient dependence.

1. Introduction

Score-based Generative Models (SGMs), also known as diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2021; Ho et al., 2020; Vahdat et al., 2021), are foundational to high-dimensional generative modeling across domains such as imagery (Dhariwal & Nichol, 2021), video (Ho et al., 2020), audio (Chen et al., 2021; Kong et al., 2021), and biological structures (Watson et al., 2023). However, their statistical behavior remains insufficiently understood when the ambient dimension D is large but the data exhibit an intrinsic dimension $d \ll D$. The manifold hypothesis (Fefferman et al., 2016) postulates that the relevant sample complexity should scale with d, independent of the ambient nonparametric dimension.

This paper provides a rigorous quantitative foundation for the manifold hypothesis within the context of diffusion models. Assuming the data distribution is supported on a compact d-dimensional smooth manifold $\mathcal{M} \subset [0,1]^D$ and admits a β-Hölder density, bounded below with respect to the volume measure on $\mathcal{M}$, we prove a nonasymptotic W_1 guarantee for a variance-preserving SGM with explicit ambient, density, and finite-order geometric dependence. In the fixed-geometry regime with $d > 2$ and the canonical finite-order choice, the bound attains the intrinsic sample exponent $n^{-(\beta+1)/(d+2\beta)}$ up to logarithmic and displayed polynomial ambient factors.

A primary theoretical challenge arises because manifold-supported distributions are singular with respect to the Lebesgue measure in $\mathbb{R}^D$. At time zero, the ambient score is undefined. At any positive time, Gaussian smoothing induces a score characterized by normal localization toward the manifold tube and tangential transport along the manifold surface. These distinct phenomena require different stable approximation strategies. The large-noise regime interacts with tangent patches and can be analyzed using intrinsic cell counts σ_t^{-d}. Conversely, the small-noise regime necessitates a projection-centered ratio expansion because the score's denominator contains a singular normal Gaussian factor. This dichotomy motivates the two-regime analytical framework developed throughout this paper.

While ambient-space SGM theory offers critical insights, it does not directly resolve this question. Convergence analyses have evolved from requiring Log-Sobolev assumptions (Lee et al., 2022; Wibisono & Yang, 2022) to accommodating weaker moment or Fisher-information conditions (Benton et al., 2024; Conforti et al., 2023). Concurrently, score-estimation theories now yield sharp ambient rates for Besov and Hölder densities (Oko et al., 2023; Dou et al., 2024; Stéphanovitch et al., 2025), with recent advances even relaxing the need for density lower bounds (Zhang et al., 2024; Fu & Lee, 2025). Because these

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

Table 1. Comparison of recent theoretical results on diffusion models for manifold data.

Paper	Manifold assumption	Density regularity	Metric	Convergence rate
(Oko et al., 2023)	$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}$ $\mathsf{supp}(\boldsymbol{Z}) = [-1,1]^d$	Besov class density lower bounded density	W_1	$\tilde{\mathcal{O}}(n^{-\frac{\beta+1-\delta}{d+2\beta}})$
(Chen et al., 2023a)	$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}$ sub-Gaussian $P_{\boldsymbol{Z}}$	Lipschitz on-support score	TV	$\tilde{\mathcal{O}}\big(Dn^{-\frac{1-\delta}{d+5}})\big)$
(Tang & Yang, 2024)	compact smooth manifold positive reach	Hölder class density lower bounded density	W_1	$\tilde{\mathcal{O}}(e^D n^{-\frac{\beta+1}{d+2\beta}})$
(Azangulov et al., 2024)	compact smooth manifold positive reach	Hölder class density lower bounded density	W_1	$\tilde{\mathcal{O}}(\sqrt{D} n^{-\frac{\beta+1}{d+2\beta}+\delta})$
(Yakovlev & Puchkin, 2025)	$\boldsymbol{X} = g^*(\boldsymbol{Z}) + \boldsymbol{\xi}$ $\boldsymbol{Z} \sim \mathcal{U}([0,1]^d)$	Hölder class g^*	TV	$\tilde{\mathcal{O}}(D^{6+\binom{d+\beta}{d}} n^{-\frac{\beta}{6\beta+2d}})$
(Zhang et al., 2026)	compact smooth manifold positive reach	Hölder class density lower bounded density	W_1	$\tilde{\mathcal{O}}\Big(D^{\mathcal{O}_\beta(d)} n^{-\frac{\beta+1}{d+2\beta}}\Big)^\dagger$
This paper	compact smooth manifold positive reach	Hölder class density lower bounded density	W_1	$\tilde{\mathcal{O}}\big(D^{\mathcal{O}_\beta(d)} n^{-\frac{\beta+1}{d+2\beta}}\big)^\ddagger$

† The displayed rate in (Zhang et al., 2026) has the intrinsic sample exponent, but the dimension-explicit constants depend on how the projection step is instantiated. ‡ Our formal theorem first yields the bound with the explicit cap term $n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}$ and the two small-noise regime conditions; the table entry records the $d > 2$ simplification after that term is absorbed.

results are tailored to full-dimensional densities in $\mathbb{R}^D$, they do not by themselves explain why the distribution-learning rate should depend on d for singular laws on nonlinear manifolds.

Other complementary perspectives study manifold detection and score singularities (Pidstrigach, 2022; Lu et al., 2023; Li et al., 2025) or broad W_p generalization through Wasserstein dimension (Chakraborty et al., 2026). These address objectives different from smooth-density distribution learning, and therefore do not yield the dimension-explicit intrinsic W_1 rate targeted here.

Comparison with prior manifold analyses. Table 1 highlights prevailing theoretical gaps. Linear-support models recover intrinsic rates but bypass curvature challenges (Oko et al., 2023; Chen et al., 2023a). For nonlinear manifolds, earlier analyses either incur exponential-in-D prefactors (Tang & Yang, 2024) or derive slower rates with high-dimensional approximation factors (Yakovlev & Puchkin, 2025). A recent support-aware analysis obtains a sharper displayed ambient prefactor, roughly $\sqrt{D}$ (Azangulov et al., 2024), but it does so for a more structured pipeline: the estimator first builds sample-dependent local low-rank models of the support, its main score-learning target is weighted by σ_t^2, and the final rate carries an n^δ slack. Our theorem targets a different constructive problem: it certifies an explicit ambient score-network realization of the projection-centered small-noise score, without preprocessing the manifold into local affine subproblems, and attains the exact intrinsic sample exponent up to logarithms. See Sec. A.2 for the theorem-level comparison.

Comparison with the concurrent work (Zhang et al., 2026). The dagger marks their displayed theorem rate in Table 1. They obtain the same intrinsic sample exponent as our result, and the distinction is not the exponent but the dimension-explicit construction behind it. Their proof invokes a generic ambient smooth-map approximation step for the nearest projection on the reach tube. Under a standard D-variate implementation, that step entails a high-order ambient approximation audit for $\Pi_{\mathcal{M}}$. In contrast, our proof adopts a finite-anchor Gauss–Newton network in intrinsic coordinates, yielding a constructive polynomial ambient audit of the score network under the stated polynomial-geometry regime. See Sec. A.1 for the detailed comparison.

1.1. Our contributions

In the displayed bounds below, $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$ denotes the finite-order geometry envelope defined in Theorem B.5. It records reach, volume, and finite-order atlas smoothness factors explicitly; these factors are not hidden by $\tilde{\mathcal{O}}(\cdot)$. The primary contributions are:

(i) *Intrinsic Wasserstein estimation under explicit geometric assumptions:* Under Theorems 2.2, 2.4 and 2.5, the polynomial control $D \vee p_{\min}^{-1} \le n^{a_0}$, and the explicit small-noise reach and denominator-resolution conditions, we prove

$$\mathbb{E}\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{lb}}_{t_0}) \le \tilde{\mathcal{O}}\Big(\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} \times \big[\mathfrak{r}_n + n^{-1/2} + n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}\big]\Big),$$

where $\mathfrak{r}_n = n^{-(\beta+1)/(d+2\beta)}$. For $d > 2$ and the canonical finite-order choice $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$, the finite-order cap term is $\mathcal{O}(\mathfrak{r}_n)$, yielding the intrinsic exponent in the fixed-geometry sense.

(ii) *Explicit ambient dependence and admissible growth regime:* The constructed score networks have leading width and sparsity $n^{d/(d+2\beta)}$ and an explicit ambient factor $(D \vee \log n)^{\mathfrak{p}_{\beta,d}}$, where $\mathfrak{p}_{\beta,d}$ is displayed in Theorem H.1 and satisfies $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$ after the $d > 2$ finite-order choice. For fixed geometry and fixed density lower bound, Theorem H.2 records a sufficient ambient-growth condition $D \vee \log n \le c\, n^{a_{\beta,d}}$ under which the two small-noise checks hold. For growing-ambient families, the polynomial-in-D interpretation additionally requires polynomial control of the finite-order geometry envelope and density factors. Sec. B.7 verifies this geometry regime for affine spheres, flat product tori, and finite-order/reach-controlled closed graph-type embeddings.

(iii) *Constructive two-regime score approximation:* The large-noise branch uses tangent-cell approximation with intrinsic cell counts, while the small-noise branch uses projection-centered de-Gaussianized Laplace expansions, a finite-anchor Gauss–Newton network for projection coordinates, and a polynomial denominator lower bound before reciprocal-network evaluation. Both branches are verified on the forward tube and then passed through clipped score-estimation and slab-wise W_1 perturbation arguments.

Technical overview. The proof separates the Gaussian footprint into two geometric regimes. At large noise, the smoothed score can be approximated by tangent-cell expansions with intrinsic cell counts, so the construction never needs an ambient grid. At small noise, the proof recenters at the nearest point on $\mathcal{M}$, factors out the common normal Gaussian term, and approximates only the tangential Laplace coefficients and the de-Gaussianized ratio. The dimension-explicit step is a finite-anchor Gauss–Newton ReLU network for projection coordinates, which replaces a black-box high-order D-variate approximation of $\Pi_{\mathcal{M}}$. The detailed two-regime proof architecture is given in Sec. 3, with appendix audit maps in Sec. B.2 and B.3.

2. Background

2.1. Notation

Let $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \ge 0\}$ denote the space of non-negative real numbers, $\mathbb{N} := \{0, 1, 2, \dots\}$ the set of natural numbers, and $\mathbb{N}_+ := \mathbb{N} \setminus \{0\}$. We let $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_D)$ represent the Gaussian distribution on $\mathbb{R}^D$ with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 \boldsymbol{I}_D$. For real numbers a and b, we write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. We use standard asymptotic notation: $a = \mathcal{O}(b)$ implies $a \le Cb$ for a universal constant $C > 0$, while $a = \mathcal{O}_\beta(b)$ indicates $a \le C_\beta b$, where the constant C_β may depend on the fixed smoothness parameter β but is independent of $d, D, n, \sigma_t, p_{\min}$, and any numerical geometry constants. The notation $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors, and $\lesssim$ suppresses constants independent of the displayed variables in the surrounding statement. Unadorned constants $C, c > 0$ are universal and may change from line to line; named or subscripted constants may depend only on the quantities explicitly stated in the surrounding lemma, theorem, or convention. For any $\beta > 0$, $\lfloor \beta \rfloor$ denotes the largest integer strictly less than β (equivalently, $\lfloor \beta \rfloor = \lceil \beta \rceil - 1$). For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k) \in \mathbb{N}^k$ and a vector $\boldsymbol{v} = (v_1, \dots, v_k) \in \mathbb{R}^k$, we define $\boldsymbol{v}^{\boldsymbol{\alpha}} := v_1^{\alpha_1} v_2^{\alpha_2} \cdots v_k^{\alpha_k}$ and $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_k$.

The main theorems use a finite geometry order chosen once and then held fixed: $q_{\mathrm{geom}} \in \mathbb{N}$,

$$q_{\mathrm{geom}} \ge q_{\min}(\beta) := \lceil \max\{\beta + 4, 2(\beta \vee 1) + 2\} \rceil . \quad (1)$$

This order controls the finite Taylor expansions used in the tangent-cell proof. For the $d > 2$ optimal W_1 theorem we additionally choose it so that $q_{\mathrm{geom}} - 2 \ge d(\beta+1)/(d-2)$. Since d is an integer and $d > 2$ implies $d/(d-2) \le 3$, this can be enforced uniformly by taking the canonical choice

$$q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta) := \lceil \max\{q_{\min}(\beta),\, 2 + 3(\beta + 1)\} \rceil . \quad (2)$$

Equivalently, any fixed order bounded by a constant depending only on β and satisfying the same lower bounds gives the same ambient exponent. Thus the finite order is $\mathcal{O}_\beta(1)$ and is never allowed to grow with n, D, or d. In the appendix, our bookkeeping collects direct ambient powers into an exponent $\mathfrak{p}_{\beta,d}$ after this finite order has been fixed. Under the preceding $d > 2$ choice, the explicit bookkeeping gives $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$. Within the main text, these rates are summarized as $(D \vee \log n)^{\mathcal{O}_\beta(d)}$ or, after hiding logarithms, as $D^{\mathcal{O}_\beta(d)}$. We also fix the target W_1 sample rate and the large-noise accuracy-cap exponent:

$$\mathfrak{r}_n := n^{-(\beta+1)/(d+2\beta)}, \qquad q_\star := q_{\mathrm{geom}} - 2. \quad (3)$$

The symbol $\mathfrak{r}_n$ is reserved for this final sample rate. Uppercase R-symbols are used only for local radii, cutoffs, or chartwise integral labels. Additional proof-bookkeeping conventions restricted to the appendix are collected in Sec. B.5.

Wasserstein distance. For P, Q two probability distributions on $\mathbb{R}^D$, their Wasserstein-1 distance is

$$\mathsf{W}_1(P, Q) := \inf_{\gamma \in \Gamma(P,Q)} \int \|\boldsymbol{x} - \boldsymbol{y}\|_2 \mathrm{d}\gamma(\boldsymbol{x}, \boldsymbol{y}),$$

and $\Gamma(P, Q)$ is the set of all couplings of P and Q. We use W_1 as it remains meaningful for singular or disjoint manifold-supported measures and respects the ambient geometry of the sampling space.

Neural network class. Following standard neural network approximation theory (Yarotsky, 2017; Schmidt-Hieber,

2020), we model the score function using deep ReLU neural networks with explicit constraints on depth, width, parameter sparsity, and magnitude. Specifically, we define the score network $s(\boldsymbol{x}, t)$ as a function mapping the concatenated input $(\boldsymbol{x}, t) \in \mathbb{R}^D \times \mathbb{R}_+$ to the score vector in $\mathbb{R}^D$.

Definition 2.1 (Sparse ReLU neural network)**.** Let $\rho(u) = \max\{u, 0\}$ act coordinatewise. For a width vector $\boldsymbol{W} = (W_0, W_1, \ldots, W_L)$, the class $\mathcal{NN}(L, \boldsymbol{W}, S, B)$ consists of all maps $f : \mathbb{R}^{W_0} \to \mathbb{R}^{W_L}$ obtained from the recursion $\boldsymbol{z}_0 = \boldsymbol{z}, \boldsymbol{z}_\ell = \rho\big(\boldsymbol{A}^{(\ell)}\boldsymbol{z}_{\ell-1} + \boldsymbol{b}^{(\ell)}\big), 1 \le \ell \le L-1, f(\boldsymbol{z}) = \boldsymbol{A}^{(L)}\boldsymbol{z}_{L-1} + \boldsymbol{b}^{(L)}$, where $\boldsymbol{A}^{(\ell)} \in \mathbb{R}^{W_\ell \times W_{\ell-1}}$ and $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{W_\ell}$ for $\ell = 1, \ldots, L$, and $\sum_{\ell=1}^{L}\big(\|\boldsymbol{A}^{(\ell)}\|_0 + \|\boldsymbol{b}^{(\ell)}\|_0\big) \le S, \max_{1\le\ell\le L}\big(\|\boldsymbol{A}^{(\ell)}\|_\infty \vee \|\boldsymbol{b}^{(\ell)}\|_\infty\big) \le B$. For score networks, we take $W_0 = D + 1$ and $W_L = D$.

2.2. Score-based Generative Models (SGMs)

Forward process. As a forward process $(\boldsymbol{X}_t)_{0\le t\le T}$ in $\mathbb{R}^D$, we consider the Variance Preserving (VP) process:

$$\begin{aligned} &\mathrm{d}\boldsymbol{X}_t = -\alpha(t)\boldsymbol{X}_t\mathrm{d}t + \sqrt{2\alpha(t)}\mathrm{d}\boldsymbol{B}_t, \quad (0 \le t \le T) \\ &\boldsymbol{X}_0 \sim P_0, \end{aligned} \tag{4}$$

where $(\boldsymbol{B}_t)_{0\le t\le T}$ denotes an independent D-dimensional standard Brownian motion and we have that $\boldsymbol{X}_t|\boldsymbol{X}_0 \sim \mathcal{N}(m_t\boldsymbol{X}_0, \sigma_t^2\boldsymbol{I}_D)$, where $m_t = \exp(-\int_0^t \alpha(\tau)\mathrm{d}\tau), \sigma_t^2 = 1 - \exp(-2\int_0^t \alpha(\tau)\mathrm{d}\tau)$.

Reverse process. The VP process (4) has a reverse process $(\boldsymbol{Y}_t)_{0\le t\le T}$, where $\boldsymbol{Y}_t = \boldsymbol{X}_{T-t}$ satisfies

$$\begin{aligned} &\mathrm{d}\boldsymbol{Y}_t = \alpha(T-t)(\boldsymbol{Y}_t + 2\nabla\log p_{T-t}(\boldsymbol{Y}_t))\mathrm{d}t + \sqrt{2\alpha(T-t)}\mathrm{d}\boldsymbol{B}'_t, \\ &\boldsymbol{Y}_0 \sim P_T, \end{aligned}$$

where $(\boldsymbol{B}'_t)_{0\le t\le T}$ is another independent D-dimensional standard Brownian motion and $\nabla\log p_t(\cdot)$ is called the score function. However, the noise distribution P_T and score function $\nabla\log p_t(\cdot)$ are unknown. It has been shown that the VP process converges exponentially to the standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ (Chen et al., 2023b) and for sufficiently large T, we can replace P_T by $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$.

Score matching loss. To estimate the true score function $s^*(\boldsymbol{x}, t) := \nabla\log p_t(\boldsymbol{x})$, we learn a score estimator $s : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ through minimizing the score matching loss:

$$\mathcal{L}_{[t_0,T]}(s) := \int_{t_0}^{T} \mathbb{E}_{\boldsymbol{X}_t}[\|s(\boldsymbol{X}_t, t) - \nabla\log p_t(\boldsymbol{X}_t)\|_2^2]\mathrm{d}t. \tag{5}$$

Because this exact objective is computationally inaccessible, we optimize the equivalent *denoising score matching* loss $\mathbb{E}_{\boldsymbol{X}_0}[\ell(s, \boldsymbol{X}_0)]$ (Hyvärinen, 2005; Vincent, 2011), where:

$$\ell(s, \boldsymbol{X}_0) := \int_{t_0}^{T} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0}[\|s(\boldsymbol{X}_t, t) - \nabla\log p_t(\boldsymbol{X}_t|\boldsymbol{X}_0)\|_2^2]\mathrm{d}t.$$

It is well established that $\mathcal{L}_{[t_0,T]}(s) - \mathcal{L}_{[t_0,T]}(s^*) = \mathbb{E}_{\boldsymbol{X}_0}[\ell(s, \boldsymbol{X}_0) - \ell(s^*, \boldsymbol{X}_0)]$.

Empirical risk minimization (ERM). Given n i.i.d training samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \sim P_0$, we learn the score estimator $\widehat{s}$ over the hypothesis class $\mathcal{NN}$ via the ERM:

$$\widehat{s} \in \arg\min_{s\in\mathcal{NN}} \frac{1}{n}\sum_{i=1}^{n} \ell(s, \boldsymbol{x}^{(i)}). \tag{6}$$

Early stopping. As $t \to 0$, the underlying score function $\nabla\log p_t(\cdot)$ diverges, causing the reverse SDE to become highly ill-conditioned. To ensure numerical stability, we truncate the reverse dynamics, terminating the simulation at a small positive time threshold $t_0 > 0$.

Generating new samples. Using the trained score estimator $\widehat{s}$ across the interval $[t_0, T]$, we construct an approximate reverse process $(\widehat{\boldsymbol{Y}}_t)_{0\le t\le T-t_0}$ and draw new samples $\boldsymbol{X}' := \widehat{\boldsymbol{Y}}_{T-t_0} \sim \widehat{P}_{t_0}$ via: $0 \le t \le T - t_0$,

$$\begin{aligned} &\mathrm{d}\widehat{\boldsymbol{Y}}_t = \alpha(T-t)(\widehat{\boldsymbol{Y}}_t + 2\widehat{s}(\widehat{\boldsymbol{Y}}_t, T-t))\mathrm{d}t + \sqrt{2\alpha(T-t)}\mathrm{d}\boldsymbol{B}_t, \\ &\widehat{\boldsymbol{Y}}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D). \end{aligned}$$

2.3. Manifold Hypothesis

Assumption 2.2 (Compact smooth manifold)**.** The predictor $\boldsymbol{X}_0 \sim P_0$ is supported on $\mathcal{M} \subset [0, 1]^D$, where $\mathcal{M}$ is a compact, C^∞, d-dimensional Riemannian manifold without boundary isometrically embedded in $\mathbb{R}^D$, with $1 \le d \ll D$ and reach $\kappa > 0$ and Riemannian volume $S_\mathcal{M} := \mathrm{vol}_\mathcal{M}(\mathcal{M})$.

Throughout the approximation bounds below, d is treated as the intrinsic dimension and its dependence is kept separate from the ambient-dimension dependence on D.

The reach controls curvature and self-avoidance of the normal tube, while $S_\mathcal{M}$ controls finite chart counts. Their precise contribution to constants is recorded in the finite-order geometry envelope in Theorem B.5. Concrete families satisfying the resulting polynomial-geometry regime, including affine spheres, flat product tori, and closed graph-type embeddings with finite-order derivative and reach control, are given in Sec. B.7.

Definition 2.3 (Hölder functions on a Riemannian manifold)**.** Fix a finite $C^{\lceil\beta\rceil}$ atlas $\{(U_i, \phi_i)\}_{i\in\mathcal{A}}$ on the compact manifold $\mathcal{M}$, with coordinate domains $\phi_i(U_i) \subset \mathbb{R}^d$, and use this atlas in the definition below. For $\beta > 0, B_0 \ge 1$, a function $f : \mathcal{M} \to \mathbb{R}$ belongs to $\mathcal{H}^\beta(\mathcal{M}, B_0)$ if, for every chart (U_i, ϕ_i) in this fixed atlas, with $f^{(i)} := f \circ \phi_i^{-1} : \phi_i(U_i) \to \mathbb{R}$, the following hold:

(i) All $\partial^{\boldsymbol{\alpha}} f^{(i)}$ with $\|\boldsymbol{\alpha}\|_1 \le \lfloor\beta\rfloor$ exist continuously and $\max_{\|\boldsymbol{\alpha}\|_1\le\lfloor\beta\rfloor} \|\partial^{\boldsymbol{\alpha}} f^{(i)}\|_{L^\infty(\phi_i(U_i))} \le B_0$.

(ii) For every $\|\boldsymbol{\alpha}\|_1 = \lfloor \beta \rfloor$,

$$\sup_{\boldsymbol{u}\neq\boldsymbol{v}\in\phi_i(U_i)} \frac{|\partial^{\boldsymbol{\alpha}} f^{(i)}(\boldsymbol{u}) - \partial^{\boldsymbol{\alpha}} f^{(i)}(\boldsymbol{v})|}{\|\boldsymbol{u}-\boldsymbol{v}\|_2^{\beta-\lfloor\beta\rfloor}} \leq B_0.$$

The radius B_0 is therefore a controlled chartwise radius, not a supremum over arbitrary smooth reparametrizations. Passing from this atlas to the reach-scale projection atlas used in the proof only changes the radius by finite transition-map constants; see Theorem D.4.

Assumption 2.4 (Hölder smooth density on manifold)**.** The law P_0 of $\boldsymbol{X}_0$ is absolutely continuous w.r.t the Riemannian volume measure $\mathrm{vol}_{\mathcal{M}}$ on $\mathcal{M}$. We write $\mathrm{d}P_0(\boldsymbol{y}) = p_0(\boldsymbol{y})\mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y})$, where $p_0 : \mathcal{M} \to [0,\infty)$ satisfies $p_0 \in \mathcal{H}^{\beta}(\mathcal{M}, B_0)$ with respect to the fixed controlled finite atlas for a radius $B_0 \geq 1$ and $\int_{\mathcal{M}} p_0(\boldsymbol{y})\mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y}) = 1$.

Assumption 2.5 (Positive density lower bound)**.** There exists a constant $p_{\min} \in (0,1]$ such that $p_0(\boldsymbol{y}) \geq p_{\min}, \forall \boldsymbol{y} \in \mathcal{M}$.

The appendix recalls the standard chart, atlas, and partition-of-unity conventions and constructs the finite projection atlas, reach-volume tube bound, and finite-order geometry envelope $\mathcal{K}_{\mathcal{M},q}$ used to bound approximation constants; see Sec. B.4, B.6 and D.3. These are bookkeeping devices for localization and network-size controls, whereas the fundamental statistical assumptions governing the main theorem are stated explicitly here.

3. Score Approximation on Smooth Manifolds

This section states the deterministic comparison networks later inserted into the clipped ERM oracle and the W_1 perturbation argument. Recall $s^*(\boldsymbol{x},t) = \nabla \log p_t(\boldsymbol{x})$ and $\boldsymbol{X}_t = m_t\boldsymbol{X}_0 + \sigma_t\boldsymbol{Z}$, with $\boldsymbol{X}_0$ supported on $\mathcal{M} \subset [0,1]^D$. The small-noise class has leading intrinsic capacity $n^{d/(d+2\beta)}$, while each large-noise slab has capacity $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$. The cap is kept explicit below and, under the final $d > 2$ choice of $q_\star$, is no larger than the assembled leading capacity up to displayed geometry and logarithmic factors. No statistical estimation is performed in this section.

3.1. Tube Localization

For any ambient coordinate $\boldsymbol{x} \in \mathbb{R}^D$, define the Euclidean distance to the rescaled manifold as:

$$\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) := \inf_{\boldsymbol{y}\in\mathcal{M}} \|\boldsymbol{x} - m_t\boldsymbol{y}\|_2.$$

For any fixed off-tube precision exponent $A > 0$, choose $C_\varrho = C_\varrho(A) > 0$ sufficiently large and define the active tube parameters:

$$\begin{aligned} \varrho_t &:= C_\varrho \sigma_t \sqrt{D \vee \log n}, \\ A_t &:= \{\boldsymbol{x} \in \mathbb{R}^D : \mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \leq \varrho_t\}. \end{aligned} \tag{7}$$

The radius matches the Gaussian footprint up to a logarithmic tail margin, so the core approximation problem is localized to A_t; outside A_t, clipping and Gaussian concentration suffice.

Lemma 3.1 (Informal version of Theorem C.3)**.** *Under Theorem 2.2, fix any $A > 0$ and choose the tube constant $C_\varrho = C_\varrho(A)$ sufficiently large. If an arbitrary comparison score $\widetilde{s}$ satisfies the uniform clipping bound $\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|\widetilde{s}(\boldsymbol{x},t)\|_\infty \leq C\sigma_t^{-1}(D \vee \log n)^{1/2}$, then the off-tube risk is bounded by:*

$$\int_{A_t^c} \|\widetilde{s}(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_2^2 p_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq C_A \sigma_t^{-2}(D \vee \log n)^C n^{-A},$$

where the constant C_A may depend on the prescribed off-tube exponent A.

3.2. Near-Tube Integral Representation

On A_t, a projection atlas and partition of unity reduce the smoothed manifold density to chartwise Gaussian integrals; the full construction is in Sec. D.3. Let $\mathcal{Q}(\boldsymbol{x},t)$ be the Gaussian-weighted manifold density and $\mathcal{P}(\boldsymbol{x},t)$ its first moment. Aggregating them isolates the centered numerator:

$$\mathcal{U}(\boldsymbol{x},t) := \frac{m_t\mathcal{P}(\boldsymbol{x},t) - \boldsymbol{x}\mathcal{Q}(\boldsymbol{x},t)}{\sigma_t}.$$

This directly yields the ambient density and target score representation:

$$p_t(\boldsymbol{x}) = (2\pi\sigma_t^2)^{-D/2}\sigma_t^d\mathcal{Q}(\boldsymbol{x},t), \qquad \sigma_t s^*(\boldsymbol{x},t) = \frac{\mathcal{U}(\boldsymbol{x},t)}{\mathcal{Q}(\boldsymbol{x},t)}.$$

Thus, the stable approximation target is the scaled, centered score $\sigma_t s^*(\boldsymbol{x},t)$, avoiding a separate approximation of $m_t\mathcal{P}$ and $\boldsymbol{x}\mathcal{Q}$ followed by cancellation and division by σ_t.

3.3. Proof Architecture for Score Approximation

The score approximation problem is dictated by the geometry of the Gaussian footprint around the scaled manifold $m_t\mathcal{M}$. On the active tube A_t, the scaled score has the centered-ratio representation $\sigma_t s^*(\boldsymbol{x},t) = \mathcal{U}(\boldsymbol{x},t)/\mathcal{Q}(\boldsymbol{x},t)$ displayed above. The two approximation mechanisms are:

Regime	Geometry mapped by Gaussian	Primary obstruction	Approximation mechanism
Large noise regime $[t_1, T]$	Tangent-scale manifold patch	Avoiding an ambient spatial grid while integrating over the forward tube	Tangent-cell approximation at radius $c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$
Small noise regime $[t_0, t_1]$	Normal fiber intersecting the nearest projection point	A σ_t-mesh is statistically prohibitive, and the raw denominator contains an infinitesimal normal Gaussian factor	Projection-centered, de-Gaussianized Laplace expansion

Thus, large noise is handled by intrinsic tangent cells, while small noise is handled by centering at $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$, factoring out the common normal Gaussian, and expanding only in d tangential coordinates. The former keeps the cell count at $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$; the latter avoids the prohibitive $\sigma_{t_0}^{-d}$ mesh and avoids treating $\Pi_{\mathcal{M}}$ as a generic D-variate smooth function. The key dimension-explicit ingredient in the small-noise branch is the projection network: it computes chart coordinates of $\Pi_{\mathcal{M}}$ by finite intrinsic anchors and Gauss–Newton iterations. In contrast, the large-noise branch uses affine tangent-plane linearization centers for local integration, not nearest projection.

Throughout this section, whenever a chart (U_i, ϕ_i) is used, write $\boldsymbol{z}_i := \phi_i^{-1}$ on its coordinate domain and $\boldsymbol{J}_i(\boldsymbol{u}) := \nabla \boldsymbol{z}_i(\boldsymbol{u}) \in \mathbb{R}^{D\times d}$.

3.4. Large Noise: Tangent-Cell Approximation

On a large-noise slab $I_k = [t_{k-1}, t_k] \subset [t_1, T]$, partition each chart box into coordinate cells of side length $h_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$, with $r_{\star,n}$ defined in Theorem G.2. The resulting complexity is the intrinsic cell count $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$ rather than an ambient grid. For a cell center $\boldsymbol{u}_\nu$, set $\boldsymbol{y}_\nu = \boldsymbol{z}_i(\boldsymbol{u}_\nu)$ and define the affine tangent-plane center

$$\mathsf{T}_\nu(\boldsymbol{x}, t) := m_t \boldsymbol{y}_\nu + \mathsf{P}_\nu^{\tan}(\boldsymbol{x} - m_t \boldsymbol{y}_\nu).$$

For $\boldsymbol{y} = \boldsymbol{z}_i(\boldsymbol{u})$ in the cell,

$$\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2^2 = \|\boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x}, t)\|_2^2 + \|\mathsf{T}_\nu(\boldsymbol{x}, t) - m_t\boldsymbol{y}\|_2^2 + 2\langle \boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x}, t), \mathsf{T}_\nu(\boldsymbol{x}, t) - m_t\boldsymbol{y}\rangle$$

decouples the phase into a tangent Gaussian envelope plus higher-order curvature corrections. The denominator satisfies the tube floor: for $\boldsymbol{x} \in A_t$,

$$\mathcal{Q}(\boldsymbol{x}, t) \geq C^{-1}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C} p_{\min} \exp\{-C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}(D\vee\log n)\}.$$

The network approximates these cellwise envelopes and their stored chart moments, then constructs the reciprocal on the displayed denominator range.

Proposition 3.2 (Informal version of Theorem E.5)**.** *Under Theorems 2.2, 2.4 and 2.5 and the polynomial-growth regime $D \vee p_{\min}^{-1} \leq n^{a_0}$, consider the dyadic large-noise grid with capped cell radius $h_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$. The cap $r_{\star,n}$ is chosen in Theorem G.2 precisely so that the phase-smallness and fixed-order geometry conditions in (65) hold on every large-noise slab. Then, for every I_k, there exists a clipped ReLU network $s_k^{\lg} \in \mathcal{NN}(L_k, \boldsymbol{W}_k, S_k, B_k)$ such that, for all $\boldsymbol{x} \in A_t$ and $t \in I_k$:*

$$\|s_k^{\lg}(\boldsymbol{x}, t) - s^*(\boldsymbol{x}, t)\|_2^2 \leq C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^C p_{\min}^{-C}\sigma_t^{-2} n^{-2(\beta+1)/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.$$

Moreover, $\sup_{\boldsymbol{x}\in\mathbb{R}^D}\|s_k^{\lg}(\boldsymbol{x}, t)\|_\infty \leq C\sigma_t^{-1}(D\vee\log n)^{1/2}$, with capacity bounded by $L_k \vee \log B_k \lesssim \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^C(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$ and $\|\boldsymbol{W}_k\|_\infty \vee S_k \lesssim \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^C p_{\min}^{-C}(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d})(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$.

3.5. Small Noise: Projection-Centered Laplace Approximation

At small noise, extending tangent cells to t_0 would be statistically too costly, so we center at the nearest projection. For $\boldsymbol{x} \in A_t$ and $t \in [t_0, t_1]$, the reach condition makes $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ unique. Let $H := D \vee \log n$ and define

$$\boldsymbol{\pi}(\boldsymbol{x}, t) := \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t), \qquad \boldsymbol{u}_i^{\Pi} := \phi_i(\boldsymbol{\pi}(\boldsymbol{x}, t)),$$
$$\boldsymbol{r} := \boldsymbol{x} - m_t\boldsymbol{\pi}(\boldsymbol{x}, t), \qquad \boldsymbol{\nu} := \boldsymbol{r}/\sigma_t.$$

Then $\nabla \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi})^\top \boldsymbol{r} = 0$, so the normal residual is separated before the tangential Laplace expansion.

By introducing the local parameterization $\boldsymbol{u} = \boldsymbol{u}_i^{\Pi} + \sigma_t\boldsymbol{w}$, the phase expansion rigorously yields:

$$\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi} + \sigma_t\boldsymbol{w})\|_2^2 - \|\boldsymbol{r}\|_2^2}{2\sigma_t^2} = \frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{G}_i(\boldsymbol{u}_i^{\Pi})\boldsymbol{w} + \sum_{\ell=1}^{\lfloor\beta\rfloor}\sigma_t^\ell\Psi_{i,\ell}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{\nu}, \boldsymbol{w}, t) + \mathcal{O}(\sigma_t^\beta).$$

The leading term is a d-dimensional Gaussian in $\boldsymbol{w}$. To remove the unstable normal factor, define

$$\bar{\mathcal{Q}} := e^{\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)}\mathcal{Q}, \qquad \bar{\mathcal{U}} := e^{\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)}\mathcal{U}, \qquad \frac{\bar{\mathcal{U}}}{\bar{\mathcal{Q}}} = \frac{\mathcal{U}}{\mathcal{Q}}.$$

The denominator now has the polynomial floor $\bar{\mathcal{Q}}(\boldsymbol{x}, t) \geq C^{-1}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C} p_{\min}(D\vee\log n)^{-d/4}$. The remaining geometric task is to realize the projection coordinate $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)$ without treating $\Pi_{\mathcal{M}}$ as a black-box D-variate smooth map.

Lemma 3.3 (Informal projection-network primitive; see Theorems F.14 and F.23)**.** *On the active reach tube used in the formal small-noise theorem, for every accuracy $\varepsilon_\Pi \in (0, 1]$ there are ReLU networks approximating $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ and its chart coordinates $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) = \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t))$. The construction uses only intrinsic nonlinear approximation: it fixes finitely many anchors in each d-dimensional chart box, unrolls approximate Gauss–Newton steps*

$$\boldsymbol{u} \mapsto \Pi_i^{(\mathrm{box})}(\boldsymbol{u} + \boldsymbol{A}_i(\boldsymbol{u})(\boldsymbol{x}/m_t - \boldsymbol{z}_i(\boldsymbol{u}))),$$
$$\boldsymbol{A}_i(\boldsymbol{u}) = (\boldsymbol{J}_i(\boldsymbol{u})^\top\boldsymbol{J}_i(\boldsymbol{u}))^{-1}\boldsymbol{J}_i(\boldsymbol{u})^\top,$$

and then combines all chart-anchor candidates by a ReLU objective gate based on $\|\boldsymbol{x}/m_t - \boldsymbol{z}_i(\boldsymbol{u})\|_2^2$. If $A_{\mathcal{M}}$ is the finite intrinsic-anchor count and C_{proj} is the geometry constant from Theorem F.14, the dominant width/sparsity cost is

$$A_{\mathcal{M}}D\big(C_{\mathrm{proj}}D\varepsilon_\Pi^{-2}\big)^{d/(2(\beta\vee 1))}\mathrm{Polylog}(C_{\mathrm{proj}}D\varepsilon_\Pi^{-1}),$$

with depth and log-weight polylogarithmic in $C_{\mathrm{proj}}D\varepsilon_\Pi^{-1}$.

Thus, the projection step has intrinsic exponent and explicit polynomial ambient dependence under polynomial-geometry control, rather than a generic $\varepsilon_\Pi^{-D/\tilde{\beta}}$ ambient smooth-function cost.

Proposition 3.4 (Informal version of Theorem F.29)**.** *Under Theorems 2.2, 2.4 and 2.5, the polynomial-growth condition $D \vee p_{\min}^{-1} \le n^{a_0}$, and the small-noise reach and denominator-resolution checks (232) and (233), there exists a clipped ReLU network $s_{\mathrm{sm}}^{\mathrm{lb}} \in \mathcal{NN}(L_{\mathrm{sm}}, \boldsymbol{W}_{\mathrm{sm}}, S_{\mathrm{sm}}, B_{\mathrm{sm}})$ active over $[t_0, t_1]$ such that: for $\boldsymbol{x} \in A_t, t \in [t_0, t_1]$*

$$\|s_{\mathrm{sm}}^{\mathrm{lb}}(\boldsymbol{x}, t) - s^*(\boldsymbol{x}, t)\|_2^2 \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} \sigma_t^{-2}.$$

Moreover, $\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|s_{\mathrm{sm}}^{\mathrm{lb}}(\boldsymbol{x}, t)\|_\infty \le C\sigma_t^{-1}(D \vee \log n)^{1/2}$, $L_{\mathrm{sm}} \vee \log B_{\mathrm{sm}} \lesssim \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}$ *and* $\|\boldsymbol{W}_{\mathrm{sm}}\|_\infty \vee S_{\mathrm{sm}} \lesssim \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}$.

The two approximation branches are assembled statistically, rather than by a separate deterministic theorem in the main text. In Sec. 4, the small-noise projection–Laplace network is trained once on $[t_0, t_1]$, while the large-noise tangent-cell networks are trained on dyadic slabs over $[t_1, T]$ and combined by ReLU time switches. The formal learned-score assembly is Theorems G.2 and G.3, and the final W_1 telescoping step is Theorem H.1.

4. Score Estimation on Smooth Manifolds

Recall the rate notation $\mathfrak{r}_n$ and $q_\star$ from (3); the large-noise accuracy radius $r_{\star,n}$ is defined in Theorem G.2. This section turns the comparison networks from Sec. 3 into learned-score bounds. The two displays below are the main-text informal versions of Theorems G.2 and G.3. On $[t_0, t_1]$, clipped ERM over the projection–Laplace class gives

$$\mathbb{E}\int_{t_0}^{t_1} \mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}_{\mathrm{sm}}^{\mathrm{lb}}(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t \le \tilde{\mathcal{O}}\Big(\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} D^{\mathcal{O}_\beta(d)}\Big).$$

On each large-noise slab $I_k = [t_{k-1}, t_k]$, clipped ERM over the tangent-cell class gives

$$\mathbb{E}\int_{I_k} \mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}_k^{\mathrm{lg}}(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t \le \tilde{\mathcal{O}}\Big(\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \Big[n^{-2(\beta+1)/(d+2\beta)} + \frac{\sigma_{t_{k-1}}^{-d} + \mathfrak{r}_n^{-d/q_\star}}{n}\Big] D^{\mathcal{O}_\beta(d)}\Big).$$

These are the learned-score inputs used in Sec. 5.

Clipped ERM oracle and covering number bounds. The formal proof is in Sec. G. We minimize the interval-restricted denoising score-matching loss, control the true smoothed risk $\mathcal{L}_I$, and clip learned scores at the tube-scale envelope $\sigma_t^{-1}(D \vee \log n)^{1/2}$.

For a slab $I = [\underline{t}, \overline{t}]$, the notation $\widetilde{\mathcal{NN}}(L, \boldsymbol{W}, S, B)$ denotes the following clipped, compact-input version of the sparse ReLU class. First apply the coordinatewise affine truncation $(\boldsymbol{x}, t) \mapsto (\Pi_{[-B_x,B_x]^D}\boldsymbol{x}, \Pi_{[\underline{t},\overline{t}]}t)$, with $B_x = (D \vee \log n)^C \sigma_{\underline{t}}^{-1}$; then rescale this compact rectangle to $[0, 1]^{D+1}$, apply a network in $\mathcal{NN}(L, \boldsymbol{W}, S, B)$, and finally clip each output coordinate to $C\sigma_t^{-1}(D \vee \log n)^{1/2}$. The off-tube estimates make the discarded input tails negligible, and the compact rescaling is exactly the one used in the covering-number step of the oracle proof.

The clipped ERM oracle inequality (Theorem G.1) balances approximation and complexity:

$$\mathbb{E}\mathcal{L}_I(\widehat{s}) \lesssim \inf_{s\in\mathcal{F}} \mathcal{L}_I(s) + n^{-A} + \frac{(D \vee \log n)^2 \int_I \sigma_t^{-2}\mathrm{d}t \, \log\big(\mathcal{N}(n^{-\mathcal{O}(1)}, \mathcal{F}, \|\cdot\|_\infty) \vee n\big)}{n}.$$

For sparse ReLU networks with structural size M, the concrete entropy evaluation is: whenever $L \vee \log B \le C_{\mathcal{F}}(D \vee \log n)^C$ and $S \vee \|\boldsymbol{W}\|_\infty \le C_{\mathcal{F}} M (D \vee \log n)^C$, Theorem I.14 gives

$$\log\big(\mathcal{N}(n^{-\mathcal{O}(1)}, \widetilde{\mathcal{NN}}(L, \boldsymbol{W}, S, B), \|\cdot\|_\infty) \vee n\big) \le CC_{\mathcal{F}} M (D \vee \log n)^C. \tag{8}$$

On our grids, $\int_I \sigma_t^{-2}\mathrm{d}t$ is logarithmic or bounded, so the stochastic capacity penalty is $\tilde{\mathcal{O}}(M/n)$ up to the displayed geometry factors in $C_{\mathcal{F}}$.

Global noise grid assembly. The small-noise estimator is trained once on $[t_0, t_1]$ with entropy size $M_{\mathrm{sm}} = n^{d/(d+2\beta)}$. The large-noise estimators are trained on dyadic slabs with radius $r_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$ and then merged by ReLU time switches. The $\mathcal{O}(\log n)$ switches add only logarithmic overhead, while the assembled score keeps both $\sigma_{t_{k-1}}^{-d}$ and $r_{\star,n}^{-d}$ explicit.

5. Distribution Estimation in W_1 Distance

This section converts the learned-score bounds from Sec. 4 into the final distributional guarantee. The formal slabwise W_1 perturbation and telescoping argument are in Sec. H.

Set $t_0 \asymp n^{-2(\beta+1)/(d+2\beta)}, t_1 \asymp n^{-2/(d+2\beta)}, T \asymp \log n$, and use the dyadic large-noise grid from (247). Then $\sqrt{t_0} \asymp \mathfrak{r}_n$, $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$, and the terminal Gaussian initialization error is negligible. Let $\widehat{s}^{\mathrm{lb}}$ be the switched clipped estimator obtained by combining the small-noise

and large-noise ERM estimators, and let $\widehat{P}^{\mathrm{lb}}_{t_0}$ be the law at time t_0 of the reverse VP process driven by $\widehat{s}^{\mathrm{lb}}$.

The rate assembly has three terms. Early stopping gives $\mathcal{O}(\sqrt{Dt_0})$, which is absorbed by the displayed ambient polynomial times $\mathfrak{r}_n$, while the terminal error is $\mathcal{O}(\sqrt{D}\, e^{-cT})$ and is made negligible by choosing $T \asymp \log n$ with a sufficiently large constant. On the small-noise interval, the score risk $\tilde{\mathcal{O}}(n^{-2\beta/(d+2\beta)})$ is multiplied by the perturbation width $\sqrt{t_1}$, giving

$$\sqrt{t_1}\, n^{-\beta/(d+2\beta)} \asymp n^{-(\beta+1)/(d+2\beta)} = \mathfrak{r}_n.$$

On the large-noise slabs, the stochastic score term contains $(\sigma_{t_{k-1}}^{-d} + \mathfrak{r}_n^{-d/q_\star})/n$. The dyadic part contributes

$$n^{-1/2} \sum_{k:\, t_k \le 1} t_k^{1/2} \sigma_{t_{k-1}}^{-d/2} \lesssim n^{-1/2}\big(1 + \sigma_{t_1}^{1-d/2}\big) \lesssim n^{-(\beta+1)/(d+2\beta)} + n^{-1/2}.$$

The order-one-time slabs add only another $\tilde{\mathcal{O}}(n^{-1/2})$ term, and the finite-order accuracy cap contributes $n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}$.

Regime for the main guarantee. For the statement below, assume the manifold and density conditions Theorems 2.2, 2.4 and 2.5, the polynomial control $D \vee p_{\min}^{-1} \le n^{a_0}$, the time grid above, the canonical finite-order choice $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$, and the two small-noise sample-size checks (232) and (233). Sec. H.1 records these conditions in checklist form and explains that, for a fixed well-conditioned manifold and fixed $p_{\min} > 0$, the small-noise checks hold for all sufficiently large n. The formal version, including network-size bounds, is Theorem H.1.

Theorem 5.1 (Main W_1 guarantee; informal version of Theorem H.1)**.** *Under the preceding regime, the switched clipped VP-SGM estimator satisfies*

$$\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{lb}}_{t_0}) \le \tilde{\mathcal{O}}\Big(\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} \times \big[\mathfrak{r}_n + n^{-1/2} + n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}\big]\Big),$$

where $q_\star$ and $\mathfrak{r}_n$ are from (3). If $d > 2$, then $q_\star \ge d(\beta+1)/(d-2)$, both bracketed correction terms are $\mathcal{O}(\mathfrak{r}_n)$, and $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$, so the bound simplifies to

$$\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{lb}}_{t_0}) \le \tilde{\mathcal{O}}\big(\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} D^{\mathcal{O}_\beta(d)} n^{-(\beta+1)/(d+2\beta)}\big).$$

The exponent $\mathfrak{p}_{\beta,d}$ and the non-log geometry and density prefactors are made explicit in Theorem H.1. For fixed geometry and fixed density lower bound, the $d > 2$ display matches the intrinsic minimax W_1 sample exponent in the fixed-geometry sense. A sufficient ambient-growth condition for the two small-noise checks is given in Theorem H.2.

6. Conclusion

This paper proves an intrinsic W_1 guarantee for a VP-SGM estimator on compact smooth manifolds by separating the score approximation into large-noise tangent cells and small-noise projection-centered Laplace expansions. The large-noise branch has explicit capacity $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$; the small-noise branch avoids a σ_{t_0}-scale mesh by de-Gaussianizing the ratio and computing projection coordinates through finite-anchor Gauss–Newton networks in local d-dimensional frames.

Under the stated reach, density, denominator, and finite-geometry conditions, the learned reverse process obeys $n^{-(\beta+1)/(d+2\beta)} + n^{-1/2} n^{\frac{d(\beta+1)}{2q_\star(d+2\beta)}}$ up to logarithmic, geometry, density, and explicit ambient factors. For $d > 2$, the canonical finite-order choice makes the cap $\mathcal{O}(n^{-(\beta+1)/(d+2\beta)})$, so the sample exponent matches the intrinsic minimax W_1 exponent in the fixed-geometry sense. The proof remains conditional on exact support, a positive density lower bound, and explicit small-noise admissibility conditions; it also studies an idealized slabwise estimator rather than shared-parameter diffusion training. Open problems include relaxing those support and density assumptions, adding ambient observational noise, and analyzing optimization and discretization error.

References

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

Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Boissonnat, J.-D., Lieutier, A., and Wintraecken, M. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *Journal of applied and computational topology*, 3(1):29–58, 2019.

Brešar, M. and Mijatović, A. Nonasymptotic bounds for forward processes in denoising diffusions: Ornstein–uhlenbeck is hard to beat. *The Annals of Applied Probability*, 35(6):4439–4463, 2025.

Bruno, S., Zhang, Y., Lim, D., Akyildiz, Ö. D., and Sabanis, S. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *CoRR*, abs/2311.13584, 2023.

Chakraborty, S., Berthet, Q., and Bartlett, P. L. Generalization properties of score-matching diffusion models for intrinsically low-dimensional data. *arXiv preprint arXiv:2603.03700*, 2026.

Chen, M., Huang, K., Zhao, T., and Wang, M. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 2023a.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023b.

Conforti, G., Durmus, A., and Silveri, M. G. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

De Santi, R., Vlastelica, M., Hsieh, Y.-P., Shen, Z., He, N., and Krause, A. Provable maximum entropy manifold exploration via diffusion models. *arXiv preprint arXiv:2506.15385*, 2025.

Dhariwal, P. and Nichol, A. Q. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021.

Dou, Z., Kotekal, S., Xu, Z., and Zhou, H. H. From optimal score matching to optimal sampling. *CoRR*, abs/2409.07032, 2024.

Federer, H. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959. doi: 10.1090/S0002-9947-1959-0110078-1.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Fu, G. and Lee, W. S. Approximation and generalization abilities of score-based neural network generative models for sub-gaussian distributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

George, A. J., Veiga, R., and Macris, N. Analysis of diffusion models for manifold data. In *IEEE International Symposium on Information Theory, ISIT 2025, Ann Arbor, MI, USA, June 22-27, 2025*, pp. 1–6. IEEE, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Huang, Z., Wei, Y., and Chen, Y. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *CoRR*, abs/2410.18784, 2024.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Jiao, Y., Zhou, Y., and Li, G. Optimal convergence analysis of DDPM for general distributions. *CoRR*, abs/2510.27562, 2025.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023.

Li, G. and Yan, Y. O(d/t) convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Li, G., Wei, Y., Chen, Y., and Chi, Y. Towards faster non-asymptotic convergence for diffusion-based generative models. *CoRR*, abs/2306.09251, 2023.

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. Accelerating convergence of score-based diffusion models, provably. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a.

Li, X., Dai, Y., and Qu, Q. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b.

Li, X., Shen, Z., Hsieh, Y., and He, N. When scores learn geometry: Rate separations under the manifold hypothesis. *CoRR*, abs/2509.24912, 2025.

Liang, J., Huang, Z., and Chen, Y. Low-dimensional adaptation of diffusion models: Convergence in total variation (extended abstract). In Haghtalab, N. and Moitra, A. (eds.), *The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3723–3729. PMLR, 2025.

Liu, Z., Zhang, W., and Li, T. Improving the euclidean diffusion generation of manifold data by mitigating score function singularity. *arXiv preprint arXiv:2505.09922*, 2025.

Lu, Y., Wang, Z., and Bal, G. Mathematical analysis of singularities in the diffusion model under the submanifold assumption. *arXiv preprint arXiv:2301.07882*, 2023.

Lyu, Y., Nguyen, T. M., Qian, Y., and Tong, X. T. Resolving memorization in empirical diffusion model for manifold data in high-dimensional spaces. *arXiv preprint arXiv:2505.02508*, 2025.

Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582. PMLR, 2023.

Pidstrigach, J. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. Linear convergence of diffusion models under the manifold hypothesis. In Haghtalab, N. and Moitra, A. (eds.), *The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4668–4685. PMLR, 2025.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875, 2020.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2256–2265. JMLR.org, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11895–11907, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Stanczuk, J., Batzolis, G., Deveney, T., and Schönlieb, C. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Stéphanovitch, A., Aamari, E., and Levrard, C. Generalization bounds for score-based generative models: a synthetic proof. *CoRR*, abs/2507.04794, 2025.

Strasman, S., Ocello, A., Boyer, C., Corff, S. L., and Lemaire, V. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, 2025, 2025.

Suzuki, T. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Tang, J. and Yan, Y. Adaptivity and convergence of probability flow odes in diffusion generative models. *CoRR*, abs/2501.18863, 2025.

Tang, R. and Yang, Y. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1648–1656. PMLR, 2024.

Tu, L. W. Manifolds. In *An Introduction to Manifolds*, pp. 47–83. Springer, 2011.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11287–11302, 2021.

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

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. Diffusion models learn low-dimensional distributions via subspace clustering. *CoRR*, abs/2409.02426, 2024.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Wibisono, A. and Yang, K. Y. Convergence in KL divergence of the inexact langevin algorithm with application to score-based generative models. *CoRR*, abs/2211.01512, 2022.

Yakovlev, K. and Puchkin, N. Generalization error bound for denoising score matching under relaxed manifold assumption. In *The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5824–5891. PMLR, 2025.

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

Zhang, K., Yin, H., Liang, F., and Liu, J. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Zhang, M. S., Huan, S., Huang, J., Boffi, N. M., Chen, S., and Chewi, S. Sublinear iterations can suffice even for ddpms. *CoRR*, abs/2511.04844, 2025.

Zhang, Z., Huang, K., Zhao, T., Wang, M., and Chen, M. Diffusion model for manifold data: Score decomposition, curvature, and statistical complexity. *arXiv preprint arXiv:2603.20645*, 2026.

A. More Discussions of Related Work

A.1. Detailed Comparison with Concurrent Work

Although the concurrent study by (Zhang et al., 2026) shares our objective of establishing curvature-driven statistical bounds for diffusion models on nonlinear manifolds, the technical difference lies in how the nearest-point projection $\Pi_{\mathcal{M}}$ is made dimension-explicit inside the proof.

Specifically, (Zhang et al., 2026, Lemma D.18) invokes a generic ambient smooth-function approximation step for $\Pi_{\mathcal{M}}$ on the reach tube. Our contribution is to replace that step by a constructive intrinsic-coordinate implementation. The comparison below therefore concerns the ambient audit of one standard black-box instantiation of the smooth-map step, not an impossibility claim for every realization of the concurrent theorem.

To see the ambient bookkeeping issue, suppose one approximates $\Pi_{\mathcal{M}}$ by a standard D-variate $C^{\tilde{\beta}}$ ReLU approximation theorem at the balancing scale needed to recover intrinsic complexity. In the relevant regime, the effective order is roughly $\tilde{\beta} \gtrsim D\beta/d$, up to logarithmic and auxiliary-accuracy factors. In a local Taylor formulation, approximating order $\tilde{\beta}$ requires a combinatorial number of ambient mixed partial derivatives:

$$\binom{D + \lceil D\beta/d \rceil}{\lceil D\beta/d \rceil} = \exp(D\{(1+\beta/d)\log(1+\beta/d) - (\beta/d)\log(\beta/d)\} + \mathcal{O}(\log D)) \,.$$

At the balancing scale $\tilde{\beta} \asymp D\beta/d$, this monomial count already scales exponentially as $\exp(c_{\beta,d} D)$ for fixed intrinsic parameters d and β.

Furthermore, if the projection map's derivative envelope grows proportionally with the differentiation order (for example, if the $\tilde{\beta}$-th derivative is bounded by $C^{\tilde{\beta}}$), this black-box route can incur an additional exponential-in-D capacity penalty. Our point is therefore constructive: a generic ambient smooth-function approximation of $\Pi_{\mathcal{M}}$ comes with a nontrivial dimension-explicit audit, whereas our proof avoids that route entirely.

By contrast, our architecture avoids high-order ambient Taylor expansions for the projection map. We evaluate the projection coordinates dynamically using finite-anchor Gauss–Newton iterations restricted entirely to d-dimensional local chart coordinates. Coupled with an analytical de-Gaussianization step that enforces a polynomial lower bound prior to the reciprocal network evaluation, our method keeps the ambient dependence of the controlled network parameters explicit and polynomial in D in the polynomial-geometry regime covered by our theorems. These distinctions affect the dimension-explicit prefactors rather than the sample exponent itself, and are summarized in Table 2.

Table 2. Comparison of our work to the concurrent study (Zhang et al., 2026).

Technical Challenge	**(Zhang et al., 2026)**	**This Paper**
Projection Mechanism	Uses a generic ambient smooth-function approximation step for $\Pi_{\mathcal{M}}$ on the reach tube.	Computes $\Pi_{\mathcal{M}}$ dynamically via finite-anchor optimization restricted to d-dimensional intrinsic charts.
Approximation Strategy	The published proof leaves the projection-step ambient audit to an ambient smooth-function approximation; under a standard black-box implementation, the relevant order is roughly $\tilde{\beta} \gtrsim D\beta/d$, up to logarithmic factors.	Resolves coordinates via Gauss–Newton iterations in $\mathbb{R}^d$, naturally achieving the intrinsic $\varepsilon^{-d/\beta}$ covering complexity.
Dimensionality Bottleneck	A standard high-order ambient implementation would involve a combinatorial number of mixed partial derivatives: $\binom{D+\tilde{\beta}}{\tilde{\beta}}$, which is $\exp(c_{\beta,d} D)$ when $\tilde{\beta} \asymp D\beta/d$.	Utilizes strictly finite-order geometry (with q_{geom} fixed before the limit); isolates ambient powers into $(D \vee \log n)^{\mathfrak{p}_{\beta,d}}$, summarized as $D^{\mathcal{O}_\beta(d)}$ in the main $d > 2$ regime.
Denominator Stability	Handles the sharp normal Gaussian singularity through the same projection-centered smooth-map step.	Explicitly de-Gaussianizes the integrals, guaranteeing a stable polynomial lower bound prior to evaluating the reciprocal network.
Complexity Guarantee	Displays the intrinsic sample exponent, while the dimension-explicit prefactors depend on how the projection step is instantiated.	Keeps network parameters polynomial in D under the polynomial-geometry conditions of the formal theorem.

A.2. Comparison with the Support-aware $\sqrt{D}$ Result

The recent paper (Azangulov et al., 2024) proves a $\tilde{\mathcal{O}}(\sqrt{D}n^{-(\alpha+1)/(2\alpha+d)+\gamma})$ Wasserstein bound under the manifold hypothesis. This is a strong ambient factor, but it is obtained for a different estimator architecture from the one studied here.

Their first key step is the score-learning target itself. In Theorem 5 of (Azangulov et al., 2024), the main bound is for the weighted quantity $\int \sigma_t^2 \mathbb{E}\|\widehat{s}_t - s_t\|_2^2\, dt$, rather than the unweighted score loss that enters our slabwise W_1 perturbation argument. This weighting downplays the most singular part of the small-noise regime. The displayed $\sqrt{D}$ factor then appears in their Corollary 6 through the final early-stopping conversion from score error to W_1.

Their second key step is structural. They first estimate the support by a sample-dependent piecewise polynomial manifold surrogate M^*, then use Proposition 21 to show that each local polynomial piece lies in an affine span of dimension $\mathcal{O}(\log n)$. Section 3.1 of (Azangulov et al., 2024) explicitly states that this reduces the ambient score problem to N simpler $\mathcal{O}(\log n)$-dimensional subproblems. The resulting estimator class is therefore support-aware: the local centers and subspaces are extracted from the data before the score network is fit. Indeed, their construction notes that these geometric quantities are found directly from the sample rather than optimized jointly within a generic ambient network.

This distinction matters for the ambient-dimension comparison. Our theorem does not preprocess the manifold into sample-dependent low-rank affine spans. Instead, it analyzes a fully explicit ambient score network that directly computes nearest-projection coordinates and the de-Gaussianized denominator in the small-noise regime. That is a harder constructive problem, so one should not expect the same ambient factor. On the other hand, the result here gives a theorem-level audit of the entire projection-centered score construction and attains the exact intrinsic sample exponent without an extra n^γ slack.

The difference is also visible inside (Azangulov et al., 2024) itself. Their Theorem 7(i) proves the dimension-light rate for the structured support-aware class, whereas Theorem 7(ii) shows that when this estimator is flattened into a more standard ambient ReLU architecture, an explicit factor of D reappears in the score-learning bound. We therefore view the two analyses as complementary: (Azangulov et al., 2024) obtains the sharper displayed ambient prefactor for a support-aware pipeline, while our theorem certifies a less structured ambient-network realization of the manifold score.

A.3. Additional Related Work

While our work focuses on the statistical guarantees of *learning* the score function, a broad body of literature addresses other theoretical aspects of diffusion models.

Convergence analysis on ambient space. Early theoretical works focused on establishing convergence guarantees for data distributions supported on the full ambient space $\mathbb{R}^D$. Initial analyses often relied on stringent structural assumptions, such as Log-Sobolev Inequalities (LSI) or log-concavity, to derive polynomial convergence rates (Lee et al., 2022; Wibisono & Yang, 2022; Bruno et al., 2023). Subsequent research sought to relax these conditions to better model complex, multi-modal data. (Chen et al., 2023b; Lee et al., 2023) demonstrated that L^2-accurate score estimation is sufficient for polynomial convergence without requiring functional inequalities, extending guarantees to distributions with bounded support or decaying tails. More recently, the focus has shifted to minimal assumptions and faster rates. (Conforti et al., 2023) showed that finite Fisher information is sufficient for convergence without early stopping. (Benton et al., 2024) utilized stochastic localization to prove that convergence rates scale linearly with the ambient dimension D (up to logarithmic factors), while (Li & Yan, 2025; Li et al., 2023) provided faster non-asymptotic convergence rates (e.g., $\mathcal{O}(D/T)$) for both ODE and SDE-based samplers. (Jiao et al., 2025) further refined these results for DDPMs, establishing near-optimal rates under relaxed smoothness conditions. Despite these advances, ambient space analyses inherently depend on D, often yielding vacuous bounds when applied to singular measures supported on low-dimensional manifolds ($d \ll D$), as they do not exploit the intrinsic geometry of the data.

Convergence analysis on low-dimensional structures. Recently, (Huang et al., 2024) and (Potaptchik et al., 2025) concurrently proved that the iteration complexity of Denoising Diffusion Probabilistic Models (DDPM) scales linearly with the intrinsic dimension d, rather than the ambient dimension D, establishing the optimal adaptivity of stochastic samplers. Similarly, for deterministic samplers, (Tang & Yan, 2025) and (Liang et al., 2025) showed that Probability Flow ODEs and DDIMs achieve dimension-free convergence rates in TV distance. (George et al., 2025) further analyzed the dynamics of these processes, identifying phase transitions (speciation and collapse) dependent on the dimension ratios. Our work complements these findings by addressing the upstream problem: statistically learning the score function from finite data.

Intrinsic-dimension statistical guarantees. Two very recent works develop complementary intrinsic-dimensional viewpoints. (Chakraborty et al., 2026) studies score-matching diffusion models for general low-dimensional measures through

a (p, q)-Wasserstein dimension, building on empirical-measure Wasserstein rates such as (Weed & Bach, 2019). Their theory yields broad W_p bounds of order roughly $n^{-1/d^\star_{p,q}(\mu)}$ up to logarithmic and algorithmic approximation terms, and it applies beyond smooth compact manifolds under finite-moment conditions. This generality is different from our smooth-density setting: for a d-dimensional compact manifold with β-Hölder density and $d > 2$, the smoothness-adaptive minimax W_1 exponent is $(\beta + 1)/(d + 2\beta)$, which is faster than the empirical-measure exponent $1/d$. (Lyu et al., 2025) also studies finite-sample manifold diffusion from empirical data and uses an inertial correction to avoid memorization, obtaining an ambient-dimension-free W_1 guarantee for C^2 manifolds. Our analysis instead learns a population score over neural-network classes for Hölder densities and audits the concrete projection–Laplace construction, including the de-Gaussianized small-noise denominator and the nearest-projection network.

Diffusion dynamics and schedules. The choice of the forward diffusion process and its noise schedule plays a central role in the performance of score-based generative models. (Strasman et al., 2025) provided error bounds explicitly depending on the noise schedule, highlighting the impact of schedule tuning. (Li et al., 2024a; Zhang et al., 2025) explored algorithmic accelerations, achieving faster convergence rates through higher-order solvers and randomized midpoint methods. Regarding the optimal process choice, (Brešar & Mijatović, 2025) analyzed the forward diffusion and argued that the canonical OU process is hard to beat in terms of convergence to the invariant measure.

Geometric interpretations. Beyond statistical rates, recent works have explored how SGMs encode geometry. (Pidstrigach, 2022) gave conditions under which SGMs detect the underlying data manifold, even when the on-manifold density is not recovered accurately. (Stanczuk et al., 2024) showed that diffusion models approximate the normal bundle of the data manifold, enabling intrinsic dimension estimation. (Wang et al., 2024) framed diffusion training as subspace clustering, showing that optimizing the diffusion loss effectively learns mixtures of low-rank Gaussians. (Lu et al., 2023) and (Liu et al., 2025) analyze the singular structure of Euclidean diffusion scores near embedded manifolds, including the normal–tangential scale separation that motivates modified sampling or training strategies. (Li et al., 2025) gives a particularly sharp small-noise perspective: geometric information in the score appears at order σ^{-2}, whereas density information appears at order one. Their results show that manifold concentration or uniform geometric sampling can tolerate substantially larger score errors than exact distributional recovery. Relatedly, (De Santi et al., 2025) studies maximum-entropy exploration over a diffusion-defined data manifold via fine-tuning. Our theorem addresses the harder distribution-learning target; correspondingly, the small-noise projection–Laplace branch controls the lower-order density-dependent terms rather than only the leading normal projection field. Finally, (Li et al., 2024b) observed that diffusion denoisers exhibit increasing linearity during the generalization phase, approximating optimal linear denoisers.

B. Appendix Roadmap and Conventions

B.1. Roadmap

The appendix is organized so that the main technical results can be read before the supporting algebra. A first pass may focus on the theorem and corollary statements in the right column, then return to the indicated support modules as needed.

Appendix module	Purpose	Main result or use
Sec. C	Off-tube Gaussian localization and Fisher control	Restricts score approximation and estimation to the forward tube
Sec. D	Projection atlas, chart integrals, centered ratio, and noise-regime notation	Defines $\mathcal{Q}, \mathcal{P}, \mathcal{U}$ and the common geometric objects used by both branches
Sec. E	Tangent-cell approximation at capped slab-scale radius $r_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$	Theorem E.5
Sec. F	Small-noise active geometry, finite-anchor projection networks, projection-centered Laplace expansion, and final ratio-network assembly	Theorems F.14 and F.29
Sec. G	Clipped ERM oracle and branchwise learned-score bounds	Theorems G.2 and G.3
Sec. H	Early stopping, terminal initialization, and slabwise W_1 transfer	Theorem H.1
Sec. I	Reusable ReLU approximation, concentration, and covering tools	Black-box support lemmas used throughout the appendix

B.2. Two-regime score-approximation map

The central score-approximation proof splits according to the effective Gaussian footprint. The main text gives the reader-facing version in Sec. 3; the appendix proof uses the following dependency map.

Regime	Geometric center	Technical device	Formal output
Large noise $[t_1, T]$	Cellwise tangent-plane centers at intrinsic radius $c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$	Chart-cell phase expansion, stored chart moments, and the unrefactored denominator tube floor	Theorem E.5, then Theorem G.2
Small noise $[t_0, t_1]$	Nearest projection point $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ and its chart coordinates	Finite-anchor Gauss–Newton projection networks, de-Gaussianized Laplace expansion, and polynomial denominator lower bound	Theorems F.14 and F.29, then Theorem G.3
Assembly	Time switches over the early and dyadic slabs	Clipped learned scores and slabwise W_1 perturbation	Theorem H.1

B.3. Proof audit map

The table below is intended as an audit checklist: each row names the statement where a possible failure would affect the final rate, what should be verified there, and where the statement is used.

Checkpoint	Statement	What to verify	Used in
Tube localization	Theorem C.3	Off-tube score risk is negligible after clipping at $\sigma_t^{-1}(D \vee \log n)^{1/2}$	Approximation and ERM
Centered ratio	(41)	The scaled score is represented as $\mathcal{U}/\mathcal{Q}$ with the cancellation already centered in the numerator	Both approximation branches
Large-noise branch	Theorem E.5	One slab has intrinsic leading size $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$ and scaled score error $n^{-(\beta+1)/(d+2\beta)}$	Large-noise ERM
Projection network	Theorem F.14	Projection centers are computed by intrinsic anchor and Gauss–Newton networks, without high-order ambient approximation of $\Pi_{\mathcal{M}}$	Small-noise branch
Small-noise denominator	Theorem F.27	After removing the normal Gaussian factor, the denominator has a polynomial lower bound before the reciprocal network is applied	Small-noise reciprocal
Small-noise branch	Theorem F.29	The projection–Laplace class has leading size $n^{d/(d+2\beta)}$ and scaled score error $n^{-\beta/(d+2\beta)}$	Small-noise ERM
Score ERM	Theorem G.1	Approximation plus sparse-network entropy gives the stated learned integrated score risks	Score estimation
W_1 transfer	Theorem I.17	Integrated score risk transfers slabwise to sampling error with the displayed time factor	Final theorem

The dependency graph is

$$\text{tube localization} \longrightarrow \begin{cases} \text{large-noise tangent cells,} \\ \text{small-noise projection–Laplace,} \end{cases} \longrightarrow \text{clipped ERM} \longrightarrow \text{slabwise } \mathsf{W}_1 \text{ transfer.}$$

B.4. Standard manifold conventions

We recall the elementary differential-geometric conventions used throughout the appendix, mainly to make the proof self-contained for readers who do not use this notation daily.

Definition B.1 (Chart)**.** Let $\mathcal{M}$ be a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$. A chart on $\mathcal{M}$ is a pair (U, ϕ), where $U \subset \mathcal{M}$ is open and $\phi : U \to \mathbb{R}^d$ is a homeomorphism onto its image.

Two charts (U, ϕ) and (V, ψ) are C^k compatible if the transition maps $\phi \circ \psi^{-1}$ and $\psi \circ \phi^{-1}$ are C^k on their common coordinate domains.

Definition B.2 (C^k atlas)**.** A C^k atlas for $\mathcal{M}$ is a collection of pairwise C^k compatible charts $\{(U_i, \phi_i)\}_{i\in\mathcal{A}}$ such that $\bigcup_{i\in\mathcal{A}} U_i = \mathcal{M}$.

Definition B.3 (Smooth manifold)**.** A smooth manifold is a manifold equipped with a C^∞ atlas.

Definition B.4 (Partition of unity)**.** A C^∞ partition of unity on $\mathcal{M}$ is a collection of nonnegative C^∞ functions $\{\rho_i : \mathcal{M} \to \mathbb{R}_+\}_{i\in\mathcal{A}}$ such that the supports are locally finite and $\sum_{i\in\mathcal{A}} \rho_i = 1$ on $\mathcal{M}$.

Every smooth manifold admits smooth partitions of unity subordinate to open covers (Tu, 2011, Theorem 13.7). In the proof we use a finite partition subordinate to the reach-scale projection atlas constructed in Sec. D.3; multiplying by these cutoffs localizes manifold integrals to single coordinate charts while preserving the finite smoothness needed in the Hölder and Taylor estimates.

B.5. Notation and bookkeeping conventions

The conventions here supplement the global notation in Sec. 2.1 only for proof bookkeeping. Unadorned constants $C, c > 0$ remain universal. Named or subscripted constants, such as C_{proj}, C_Q, or $C_{\mathrm{size,sm}}$, may depend on the fixed objects named in the statement where they are introduced, but not on n, D, σ_t or a tolerance parameter unless that dependence is displayed. The notation $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in displayed sample-size, ambient, time-cutoff, and tolerance scales; it does not hide the density lower bound, reach, volume, or finite-order geometry factors unless this is stated explicitly.

Local approximation and estimation tolerances use descriptive subscripts; no global sample-size tolerance is introduced. We write the smallest large-noise noise scale as $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$ and the centered-residual scale as $\mathfrak{r}_n = n^{-(\beta+1)/(d+2\beta)}$.

The large-noise intrinsic cell radius is not identified with σ_{t_1}: on a slab $I_k = [t_{k-1}, t_k]$ it is $r_k = c_0 \sigma_{t_{k-1}} \wedge r_{\star,n}$, where $r_{\star,n}$ is the accuracy-limited radius defined in Theorem G.2. Thus, when the cap is active, the complexity records both $\sigma_{t_{k-1}}^{-d}$ and $r_{\star,n}^{-d}$ explicitly.

The symbols δ, Δ, and η are reserved for geometric quantities such as active radii, displacements, and chart/projection margins. The tube radius ϱ_t, the generic physical density floor ρ_t, and the partition-of-unity weights ρ_i are distinct. Under the density lower bound, denominator floors used inside reciprocal networks are denoted locally.

In intermediate appendix estimates, $\mathfrak{a}_D, \mathfrak{a}_{\log} \geq 1$ denote fixed finite bookkeeping exponents for direct ambient powers and shell/covering factors, while $\mathfrak{a}_{\mathrm{geom}} \geq 1$ denotes a fixed exponent for the displayed geometry factors $1 \vee S_{\mathcal{M}}$, $1 \vee \kappa^{-1}$, and $\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}$. These exponents never depend on $D, n, \sigma_t, \varepsilon_{\mathrm{rel}}, p_{\min}$ or on the numerical values of the geometry constants. The appendix analysis interval is denoted by $[\underline{t}, \overline{t}]$. When a proof localizes to the high-probability tube, H or H_n denotes $D \vee \log n$ unless another meaning is explicitly declared. On a generic compact time interval, $\sigma_{\underline{t}}$ always means $\inf_{t \in [\underline{t}, \overline{t}]} \sigma_t$ rather than a formal endpoint value; the corresponding schedule lower- and upper-bound convention is stated explicitly in (11).

B.6. Finite-order geometry and tube-volume conventions

Definition B.5 (Finite-order geometry convention)**.** Fix the reach-scale projection atlas used below. For each $q \geq 1$, let $\mathcal{K}_{\mathcal{M},q} \geq e$ denote a finite high-order atlas envelope controlling the C^{q+4} norms of the inverse projection charts, transition maps, partition cutoffs, Jacobian factors, the maps $\boldsymbol{A}_i = \boldsymbol{G}_i^{-1} \boldsymbol{J}_i^{\top}$, and the nearest-projection coordinate and residual maps on the compact reach tubes where they are used. The finite geometry constant $\Gamma_{\mathcal{M},q}$ in the approximation proofs may be chosen so that

$$\Gamma_{\mathcal{M},q} \leq C_{d,q} (1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}(q)} (1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}(q)} \mathcal{K}_{\mathcal{M},q}^{\mathfrak{a}_{\mathrm{geom}}(q)},$$

where $\mathfrak{a}_{\mathrm{geom}}(q)$ depends only on d and q. In final statements, $\mathfrak{a}_{\mathrm{geom}}$ denotes this exponent at $q = q_{\mathrm{geom}}$, enlarged to absorb the finite network-arithmetic powers of $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$. Reach alone does not control $\mathcal{K}_{\mathcal{M},q}$, since the approximation uses finitely many derivatives of charts and projection maps above second order.

Polynomial-geometry regime. The phrase "polynomial ambient dependence" in the main paper is always conditional on the displayed instance-dependent factors being polynomially controlled. Concretely, for the fixed order q_{geom} used in the final theorem, if there is an exponent a_{geom} independent of n and D such that

$$S_{\mathcal{M}} \vee \kappa^{-1} \vee \mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}} \leq D^{a_{\mathrm{geom}}},$$

then $\Gamma_{\mathcal{M},q_{\mathrm{geom}}} \leq D^{C_{\beta,d,a_{\mathrm{geom}}}}$ and every network-size bound displayed below has polynomial ambient dependence. This covers, for example, families admitting reach-scale atlases with polynomially many uniformly well-conditioned charts and polynomially bounded derivatives up to the fixed order $q_{\mathrm{geom}} + 4$. If these finite-order geometry envelopes grow exponentially with D, our statements display that growth rather than hiding it in $\tilde{\mathcal{O}}(\cdot)$ or in universal constants.

B.7. Examples satisfying the polynomial-geometry regime

We record concrete geometric families for which the preceding polynomial regime is verified. The statements below are only about the geometric factors $S_{\mathcal{M}}, \kappa^{-1}$, and $\mathcal{K}_{\mathcal{M},q}$. The density assumptions in the main theorem can then be imposed independently; for example, the uniform density on any of the following manifolds has $p_{\min}^{-1} \leq 1 \vee S_{\mathcal{M}}$ and $B_0 \leq 1 \vee S_{\mathcal{M}}^{-1}$, so the density factors are polynomially controlled in the examples below as well; the radius or pseudoinverse hypotheses give the needed polynomial lower bound on $S_{\mathcal{M}}$.

Proposition B.6 (Basic polynomial-geometry examples)**.** *Fix $d \geq 1$, a finite order $q \geq 1$, and set $q_{\mathrm{emb}} = q + 6$. The following families satisfy the polynomial-geometry regime for the reach-scale projection atlas in Theorem B.5. For each displayed manifold write $\kappa_D = \mathrm{reach}(\mathcal{M}_D)$. In (iii)–(iv), C^k norms of ambient-valued maps use Euclidean target and operator norms in the fixed finite atlas.*

(i) Affine spheres. *Let $D \geq d+1$, let $E_D = \mathbb{R}^{d+1} \times \{0\}^{D-d-1}$, let $\boldsymbol{c}_D = (1/2, \ldots, 1/2)$, and let $0 < r_D \leq 1/4$. Set*

$$\mathcal{M}_D = \{\boldsymbol{c}_D + r_D \boldsymbol{u} : \boldsymbol{u} \in E_D,\ \|\boldsymbol{u}\|_2 = 1\} \subset [0,1]^D.$$

If $r_D^{-1} \leq D^a$, then

$$S_{\mathcal{M}_D} \vee \kappa_D^{-1} \vee \mathcal{K}_{\mathcal{M}_D,q} \leq C_{d,q,a} D^{C_{d,q} a}.$$

(ii) Flat product tori. *Let $D \geq 2d$, let $0 < r_{\min,D} \leq r_{j,D} \leq 1/4$ for $j = 1, \ldots, d$, define*

$$\mathcal{C}_{j,D} := \{(1/2 + r_{j,D}\cos\theta, 1/2 + r_{j,D}\sin\theta) : \theta \in [0, 2\pi)\},$$

and set $\mathcal{M}_D = \mathcal{C}_{1,D} \times \cdots \times \mathcal{C}_{d,D} \times \{1/2\}^{D-2d}$. If $r_{\min,D}^{-1} \leq D^a$, then

$$S_{\mathcal{M}_D} \vee \kappa_D^{-1} \vee \mathcal{K}_{\mathcal{M}_D,q} \leq C_{d,q,a} D^{C_{d,q}a}.$$

(iii) A finite-order embedding criterion. *Let N be a fixed compact smooth d-manifold without boundary, equipped with a fixed finite atlas. Let $\Phi_D : N \to [0,1]^D$ be a $C^{q_{\mathrm{emb}}}$ embedding and set $\mathcal{M}_D = \Phi_D(N)$. Suppose that, in the fixed atlas,*

$$\|\Phi_D\|_{C^{q_{\mathrm{emb}}}(N)} \vee \|(\mathrm{d}\Phi_D)^\dagger\|_{C^{q_{\mathrm{emb}}-1}(N)} \vee \mathrm{reach}(\mathcal{M}_D)^{-1} \leq D^a,$$

where $(\mathrm{d}\Phi_D)^\dagger$ is the Moore–Penrose inverse of the differential in local coordinates. Then

$$S_{\mathcal{M}_D} \vee \kappa_D^{-1} \vee \mathcal{K}_{\mathcal{M}_D,q} \leq C_{N,d,q,a} D^{C_{N,d,q}a}.$$

(iv) Closed graph-type manifolds. *As a special case of (iii), let N be a fixed closed d-manifold and let $\iota : N \to \mathbb{R}^{m_0}$ be a fixed smooth embedding. For $D \geq m_0$, let $F_D : N \to \mathbb{R}^{D-m_0}$ be a $C^{q_{\mathrm{emb}}}$ map. Suppose $\|F_D\|_{C^{q_{\mathrm{emb}}}(N)} \leq D^a$ and, for some $\boldsymbol{\tau}_D \in \mathbb{R}^D$, $\Phi_D = \lambda_D(\iota, F_D) + \boldsymbol{\tau}_D$ maps N into $[0,1]^D$ with $D^{-a} \leq \lambda_D \leq 1$ and*

$$\mathrm{reach}(\Phi_D(N))^{-1} \leq D^a.$$

Then $\mathcal{M}_D = \Phi_D(N)$ satisfies the bound in (iii), with possibly larger constants. Thus closed graph manifolds are covered when finite-order derivatives and reach are polynomially controlled. A mere C^2 curvature bound, without finite-order control and a reach lower bound, is not sufficient for the present theorem.

Proof. For (i), the reach is exactly r_D and $S_{\mathcal{M}_D} = r_D^d \mathrm{vol}(\mathbb{S}^d)$. The reach-scale tangent-projection charts are, up to ambient rotations, the usual graph charts of a round sphere over tangent balls of radius cr_D; differentiating the graph formula shows that their inverse maps, transition maps, and reach-scale cutoffs have C^{q+4} norms bounded by powers of r_D^{-1}. On the $r_D/2$-tube,

$$\Pi_{\mathcal{M}_D}(\boldsymbol{x}) = \boldsymbol{c}_D + r_D \frac{P_{E_D}(\boldsymbol{x} - \boldsymbol{c}_D)}{\|P_{E_D}(\boldsymbol{x} - \boldsymbol{c}_D)\|_2},$$

and the denominator is bounded below by $r_D/2$. Differentiating the displayed formula up to order $q+4$ gives $\mathcal{K}_{\mathcal{M}_D,q} \leq C_{d,q} r_D^{-C_{d,q}}$.

For (ii), the reach is bounded below by $c_d r_{\min,D}$, and the nearest projection in the corresponding product tube is obtained by normalizing each coordinate pair. The volume is $S_{\mathcal{M}_D} = (2\pi)^d \prod_{j=1}^d r_{j,D}$. The reach-scale tangent-projection charts are smooth changes of the product angular charts on angular intervals of length bounded below by a constant multiple of $r_{\min,D}/r_{j,D}$, and quantitative inversion gives finite-order chart, transition, and cutoff bounds by powers of $r_{\min,D}^{-1}$. The coordinatewise projection formula

$$(x_{2j-1}, x_{2j}) \mapsto (1/2, 1/2) + r_{j,D} \frac{(x_{2j-1} - 1/2, x_{2j} - 1/2)}{\|(x_{2j-1} - 1/2, x_{2j} - 1/2)\|_2},$$

has denominators bounded below by a constant multiple of $r_{\min,D}$. Hence $\mathcal{K}_{\mathcal{M}_D,q} \leq C_{d,q} r_{\min,D}^{-C_{d,q}}$.

For (iii), the volume element in a fixed base chart is $\det((\mathrm{d}\Phi_D)^\top \mathrm{d}\Phi_D)^{1/2}$. The C^1 bound on Φ_D gives a polynomial upper bound on every singular value of $\mathrm{d}\Phi_D$, while the pseudoinverse bound gives a polynomial lower bound on the smallest singular value. Since the base atlas on N is fixed, integrating over a fixed positive coordinate volume gives $S_{\mathcal{M}_D} \leq C_N D^{C_d a}$ and $S_{\mathcal{M}_D}^{-1} \leq C_N D^{C_d a}$. The reach assumption gives $\kappa_D^{-1} \leq D^a$.

It remains to bound $\mathcal{K}_{\mathcal{M}_D,q}$. Choose a maximal $c(1 \wedge \kappa_D)$-separated set of centers on $\mathcal{M}_D$. Standard reach-volume estimates bound the number of centers by a fixed power of $(1 \vee S_{\mathcal{M}_D})(1 \vee \kappa_D^{-1})$. At a center $\boldsymbol{y} = \Phi_D(u_0)$, use the tangent-projection map

$$u \mapsto P_{T_{\boldsymbol{y}}\mathcal{M}_D}(\Phi_D(u) - \boldsymbol{y})$$

in a base chart. Its derivative at u_0 has smallest singular value at least $\|(\mathrm{d}\Phi_D)^\dagger\|^{-1}$, and its derivatives up to order q_{emb} are polynomially bounded. The reach chart lemma Theorem D.2 supplies the reach-scale tangent-projection domains. On those domains, the tangent variation bound in (14) and the pseudoinverse hypothesis give a polynomial lower bound on the singular values of the coordinate differential. The quantitative inverse-function theorem and the standard inverse-derivative formulas then give coordinate maps, inverse maps, and transition maps with C^{q+4} norms bounded by $D^{C_{N,d,q}a}$. A scaled bump construction on the same cover gives partition cutoffs with the same type of bound. Jacobian factors, Gram matrices, inverse Gram matrices, and $\boldsymbol{A}_i = \boldsymbol{G}_i^{-1}\boldsymbol{J}_i^\top$ obey the same type of bound by repeated differentiation of determinant and matrix-inverse formulas.

On the corresponding reach tube, the nearest projection is unique. In a tangent graph chart $\boldsymbol{z}_i$, its local coordinate $\boldsymbol{u}^\Pi(\boldsymbol{\xi})$ solves

$$\boldsymbol{J}_i(\boldsymbol{u})^\top(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})) = 0.$$

The derivative of this equation with respect to $\boldsymbol{u}$ equals the negative Gram matrix plus a curvature-normal term. The tangent graph construction and the restriction to a fixed numerical fraction of the reach tube make this derivative uniformly invertible, with inverse norm bounded by $D^{C_{N,d,q}a}$. Applying the quantitative implicit function theorem once more gives C^{q+4} bounds, polynomial in D, for the projection-coordinate and residual maps. Therefore $\mathcal{K}_{\mathcal{M}_D,q} \le C_{N,d,q,a} D^{C_{N,d,q}a}$.

For (iv), the fixed embedding ι gives a fixed lower singular-value bound for $\mathrm{d}\iota$. Since $\|\lambda_D(\mathrm{d}\iota, \mathrm{d}F_D)v\|_2 \ge \lambda_D \|\mathrm{d}\iota\, v\|_2$ and $\lambda_D \ge D^{-a}$, the pseudoinverse norm and its derivatives are polynomially controlled by the $C^{q_{\mathrm{emb}}}$ bound on F_D and matrix-inverse differentiation. Thus the hypotheses of (iii) hold. □

Global tube-volume bound. We use the following standard reach-volume consequence only for localization and geometry bookkeeping. With $\omega_k = \pi^{k/2}/\Gamma(k/2+1)$ and $\omega_0 = 1$, there is a constant $C_{\mathrm{tube}} \le C_d(1 \vee S_{\mathcal{M}})(1 \vee \kappa^{-1})^d$ such that, for all $m \in (0, 1]$ and $r > 0$,

$$\mathrm{vol}((m\mathcal{M})_r) \le C_{\mathrm{tube}}(1+r)^d \omega_{D-d} r^{D-d}. \tag{9}$$

This is the global tube bound used in the off-tube Gaussian tail estimates; the radius is not required to be below the reach.

C. Off-Tube Score Localization

C.1. Off-manifold Gaussian tail and Fisher bounds

Lemma C.1 (Gaussian smoothing off-manifold deviation bound)**.** *Let $\boldsymbol{Y} \sim P$ be supported on a d-dimensional smooth manifold $\mathcal{M} \subseteq [0,1]^D$. Fix $m > 0$ and $\sigma > 0$, and define $\boldsymbol{X} = m\boldsymbol{Y} + \sigma\boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$. For every $A > 0$, there is a constant $C_{\text{off}}(A) > 0$ such that*

$$\Pr[\operatorname{dist}(\boldsymbol{X}, m\mathcal{M}) \geq C_{\text{off}}(A)\sigma\sqrt{D \vee \log n}] \leq n^{-A}.$$

Proof. In the following, we apply the Gaussian concentration inequality for Lipschitz functions to upper bound $\Pr[\operatorname{dist}(\boldsymbol{x}, m\mathcal{M}) \geq C_{\text{off}}(A)\sigma\sqrt{D \vee \log n}]$. Specifically, let the random point be $\boldsymbol{x} = m\boldsymbol{y} + \sigma\boldsymbol{z}$, where $\boldsymbol{y} \in \mathcal{M}$ and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$. We can define a function of the standard Gaussian vector $\boldsymbol{z}$ by:

$$h(\boldsymbol{z}) = \operatorname{dist}(m\boldsymbol{y} + \sigma\boldsymbol{z}, m\mathcal{M}).$$

Notice that the distance function $\operatorname{dist}(\cdot, \mathcal{M})$ is 1-Lipschitz. Then for any $\boldsymbol{z}_1, \boldsymbol{z}_2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$, it holds that

$$\begin{aligned}
|h(\boldsymbol{z}_1) - h(\boldsymbol{z}_2)| &= |\operatorname{dist}(m\boldsymbol{y} + \sigma\boldsymbol{z}_1, m\mathcal{M}) - \operatorname{dist}(m\boldsymbol{y} + \sigma\boldsymbol{z}_2, m\mathcal{M})| \\
&\leq \|(m\boldsymbol{y} + \sigma\boldsymbol{z}_1) - (m\boldsymbol{y} + \sigma\boldsymbol{z}_2)\|_2 \\
&= \|\sigma(\boldsymbol{z}_1 - \boldsymbol{z}_2)\|_2 = \sigma\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2,
\end{aligned}$$

which implies that $h(\boldsymbol{z})$ is σ-Lipschitz. Applying the Gaussian Concentration inequality to the σ-Lipschitz function $h(\boldsymbol{z})$, we get:

$$\Pr[h(\boldsymbol{z}) - \mathbb{E}[h(\boldsymbol{z})] \geq t] \leq \exp\Big(-\frac{t^2}{2\sigma^2}\Big).$$

Since $m\boldsymbol{y} \in m\mathcal{M}$, the distance from $m\boldsymbol{y} + \sigma\boldsymbol{z}$ to the manifold is upper bounded by the distance to $m\boldsymbol{y}$ itself:

$$h(\boldsymbol{z}) = \operatorname{dist}(m\boldsymbol{y} + \sigma\boldsymbol{z}, m\mathcal{M}) \leq \|(m\boldsymbol{y} + \sigma\boldsymbol{z}) - m\boldsymbol{y}\|_2 = \sigma\|\boldsymbol{z}\|_2.$$

Using the standard bound for the expected norm of a Gaussian vector, $\mathbb{E}[\|\boldsymbol{z}\|_2] \leq \sqrt{D}$, we obtain:

$$\mathbb{E}[h(\boldsymbol{z})] \leq \sigma\sqrt{D}.$$

Set $r = C_{\text{off}}(A)\sigma\sqrt{D \vee \log n}$. Taking $C_{\text{off}}(A)$ sufficiently large depending only on A, we have $t := r - \mathbb{E}[h(\boldsymbol{z})] \geq c_{\text{off}}(A)\sigma\sqrt{D \vee \log n}$ for a constant $c_{\text{off}}(A)$ with $c_{\text{off}}(A)^2/2 \geq A$. Substituting this into the concentration bound, we obtain:

$$\Pr[h(\boldsymbol{z}) \geq C_{\text{off}}(A)\sigma\sqrt{D \vee \log n}] \leq \exp\Big(-\frac{c_{\text{off}}(A)^2(D \vee \log n)}{2}\Big) \leq n^{-A}.$$

□

Lemma C.2 (Off-manifold Fisher information bound)**.** *Let $\boldsymbol{Y} \sim P$ be supported on a d-dimensional manifold $\mathcal{M} \subseteq [0,1]^D$. Fix $m > 0$ and $\sigma > 0$, and define $\boldsymbol{X} = m\boldsymbol{Y} + \sigma\boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$. Let $p_\sigma : \mathbb{R}^D \to \mathbb{R}_+$ be the density of $\boldsymbol{X}$. For every desired polynomial tail exponent $A > 0$, there is a threshold constant $C_{\text{off}}(A) > 0$ such that*

$$\int_{\operatorname{dist}(\boldsymbol{x}, m\mathcal{M}) \geq C_{\text{off}}(A)\sigma\sqrt{D \vee \log n}} \Big\|\frac{\nabla p_\sigma(\boldsymbol{x})}{p_\sigma(\boldsymbol{x})}\Big\|_2^2 p_\sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \lesssim \frac{D}{\sigma^2} n^{-A}.$$

Proof. Fix $A > 0$ and let $C_{\text{off}}(A)$ be chosen at the end of the proof. By $\boldsymbol{X} = m\boldsymbol{Y} + \sigma\boldsymbol{Z}$, we can apply Tweedie's formula to express the true score:

$$\frac{\nabla p_\sigma(\boldsymbol{x})}{p_\sigma(\boldsymbol{x})} = \frac{1}{\sigma^2}\mathbb{E}[m\boldsymbol{Y} - \boldsymbol{X} | \boldsymbol{X} = \boldsymbol{x}] = -\frac{1}{\sigma}\mathbb{E}[\boldsymbol{Z} | \boldsymbol{X} = \boldsymbol{x}].$$

Consequently, the squared norm evaluates to:

$$
\begin{aligned}
&\int_{\mathrm{dist}(\boldsymbol{x},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}} \Big\|\frac{\nabla p_\sigma(\boldsymbol{x})}{p_\sigma(\boldsymbol{x})}\Big\|_2^2 p_\sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathrm{dist}(\boldsymbol{x},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}} \frac{1}{\sigma^2}\big\|\mathbb{E}[\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}]\big\|_2^2 p_\sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} && \text{(by Tweedie's formula)} \\
&= \frac{1}{\sigma^2}\mathbb{E}\Big[\|\mathbb{E}[\boldsymbol{Z}|\boldsymbol{X}]\|_2^2\,\mathbb{1}\{\mathrm{dist}(\boldsymbol{X},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}\}\Big] && \text{(by definition of expectation)} \\
&\leq \frac{1}{\sigma^2}\mathbb{E}\Big[\mathbb{E}[\|\boldsymbol{Z}\|_2^2|\boldsymbol{X}]\,\mathbb{1}\{\mathrm{dist}(\boldsymbol{X},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}\}\Big] && \text{(by Jensen's inequality)} \\
&\leq \frac{1}{\sigma^2}\sqrt{\mathbb{E}[\|\boldsymbol{Z}\|_2^4]\,\Pr[\mathrm{dist}(\boldsymbol{X},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}]} && \text{(by Cauchy-Schwarz)} \\
&\leq \frac{1}{\sigma^2}\sqrt{(D^2+2D)\,\Pr[\mathrm{dist}(\boldsymbol{X},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}]} && \text{(by } \mathbb{E}[\|\boldsymbol{Z}\|_2^4]=D^2+2D\text{)} \\
&\lesssim \frac{D}{\sigma^2}\sqrt{\Pr[\mathrm{dist}(\boldsymbol{X},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}]}. && (10)
\end{aligned}
$$

Given the exponent $A>0$ in the statement, choose the threshold $C_{\mathrm{off}}(A)$ large enough so that Theorem C.1 applies with exponent $2A$. Then

$$\Pr[\mathrm{dist}(\boldsymbol{X},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}]\leq n^{-2A}.$$

Combining this with (10) gives

$$\int_{\mathrm{dist}(\boldsymbol{x},m\mathcal{M})\geq C_{\mathrm{off}}(A)\sigma\sqrt{D\vee\log n}} \Big\|\frac{\nabla p_\sigma(\boldsymbol{x})}{p_\sigma(\boldsymbol{x})}\Big\|_2^2 p_\sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \lesssim \frac{D}{\sigma^2}n^{-A}.$$

□

Lemma C.3 (Off-manifold clipped-score error bound)**.** *Suppose Theorem 2.2 holds. Fix $\ell_{\mathrm{off}}\geq 1$. For every $A>0$, let $\varrho_t = C_\varrho(A)\sigma_t\sqrt{D\vee\log n}$, with $C_\varrho(A)$ chosen large enough for polynomial tail exponent A. For any measurable function $\widetilde{s}:\mathbb{R}^D\times[\underline{t},\overline{t}]\to\mathbb{R}^D$ satisfying*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D}\|\widetilde{s}(\boldsymbol{x},t)\|_\infty \lesssim \sigma_t^{-1}\sqrt{D\ell_{\mathrm{off}}},$$

it holds that

$$\int_{\mathrm{dist}(\boldsymbol{x}_t,m_t\mathcal{M})\geq\varrho_t}\|\widetilde{s}(\boldsymbol{x}_t,t)-\nabla\log p_t(\boldsymbol{x}_t)\|_2^2 p_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \lesssim \big(D+D^2\ell_{\mathrm{off}}\big)\sigma_t^{-2}n^{-A}.$$

Proof. Let $\ell_{\mathrm{off}}\geq 1$ be the logarithmic factor in the coordinatewise bound of the lemma. Fix $\varrho_t = C_\varrho(A)\sigma_t\sqrt{D\vee\log n}$ with $C_\varrho(A)$ chosen large enough for the polynomial tail exponent required below. By Theorem C.2 it holds that

$$\int_{\mathrm{dist}(\boldsymbol{x}_t,m_t\mathcal{M})\geq\varrho_t}\Big\|\frac{\nabla p_t(\boldsymbol{x}_t)}{p_t(\boldsymbol{x}_t)}\Big\|_2^2 p_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \lesssim \frac{D}{\sigma_t^2}n^{-A}.$$

By Theorem C.1 and the coordinatewise uniform assumption $\sup_{\boldsymbol{x}\in\mathbb{R}^D}\|\widetilde{s}(\boldsymbol{x},t)\|_\infty\lesssim\sigma_t^{-1}\sqrt{D\ell_{\mathrm{off}}}$, we have

$$
\begin{aligned}
\int_{\mathrm{dist}(\boldsymbol{x}_t,m_t\mathcal{M})\geq\varrho_t}\|\widetilde{s}(\boldsymbol{x}_t,t)\|_2^2 p_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t &\leq D\Big(\sup_{\boldsymbol{x}\in\mathbb{R}^D}\|\widetilde{s}(\boldsymbol{x},t)\|_\infty\Big)^2\cdot\Pr\big[\mathrm{dist}(\boldsymbol{X}_t,m_t\mathcal{M})\geq\varrho_t\big] \\
&\lesssim \sigma_t^{-2}D^2\ell_{\mathrm{off}}n^{-A}.
\end{aligned}
$$

Using the triangle inequality $\|a-b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$, we combine these bounds:

$$\begin{aligned}
&\int_{\mathrm{dist}(\boldsymbol{x}_t, m_t\mathcal{M}) \ge \varrho_t} \|\nabla \log p_t(\boldsymbol{x}_t) - \widetilde{s}(\boldsymbol{x}_t, t)\|_2^2 p_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \\
&\lesssim \int_{\mathrm{dist}(\boldsymbol{x}_t, m_t\mathcal{M}) \ge \varrho_t} \left\| \frac{\nabla p_t(\boldsymbol{x}_t)}{p_t(\boldsymbol{x}_t)} \right\|_2^2 p_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t + \int_{\mathrm{dist}(\boldsymbol{x}_t, m_t\mathcal{M}) \ge \varrho_t} \|\widetilde{s}(\boldsymbol{x}_t, t)\|_2^2 p_t(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t \\
&\lesssim \frac{D}{\sigma_t^2} n^{-A} + \frac{D^2 \ell_{\mathrm{off}}}{\sigma_t^2} n^{-A} \lesssim \frac{D + D^2 \ell_{\mathrm{off}}}{\sigma_t^2} n^{-A}.
\end{aligned}$$

□

D. Setup for Branchwise Score Approximation

To approximate the score near the forward-smeared manifold under a density lower bound, we resolve the geometry of the underlying data manifold $\mathcal{M}$ and approximate the centered numerator ratio. The setup section fixes the noise regimes, geometric charts, amplitudes, and chart-integral notation. The large-noise component uses the tangent-cell construction in Sec. E. The small-noise component uses the unified projection–Laplace pipeline in Sec. F; its internal subsections isolate active geometry, finite-anchor projection networks, the Laplace expansion, and final ratio-network assembly. The learned-score section combines these two deterministic branches directly. For readability, this appendix is best read in three passes. First read the definitions of $\mathcal{Q}, \mathcal{P}, \mathcal{U}$ in Sec. D.3.3 and the final branch statements Theorems E.5 and F.29. Second read the proof-architecture paragraphs at the starts of Sec. E and F. Finally consult the supporting lemmas for the particular denominator, projection, or network-size estimate used in the main text. For the VP forward process considered here, $\boldsymbol{X}_t = m_t \boldsymbol{X}_0 + \sigma_t \boldsymbol{Z}$, the deterministic schedules satisfy $m_t \le 1$ and $\sigma_t \le 1$. On the truncated time interval $[\underline{t}, \overline{t}]$ we assume a positive lower bound for m_t, and we use the exact infimum of the noise schedule as the noise lower bound:

$$0 < \underline{m} \le m_t \le \overline{m} \le 1, \qquad \sigma_{\underline{t}} := \inf_{t \in [\underline{t}, \overline{t}]} \sigma_t, \qquad 0 < \sigma_{\underline{t}} \le \sigma_t \le \overline{\sigma} \le 1. \tag{11}$$

If the noise schedule is monotone increasing, this notation agrees with the endpoint value $\sigma(\underline{t})$. All schedule-approximation size bounds below are stated for the fixed compact time interval $[\underline{t}, \overline{t}]$; logarithmic factors in $\underline{t}^{-1}$ are absorbed into the constants unless explicitly displayed.

D.1. Noise regimes and branch-scale validity

Definition D.1 (Branch-scale small- and large-noise validity regions)**.** Fix constants $C_{\mathrm{sm}} \in (0, 1]$ and $c_{\mathrm{lg}} > 0$ satisfying

$$4c_{\mathrm{lg}} \le C_{\mathrm{sm}}.$$

Their final numerical values are chosen after the uniform Laplace remainder constant is fixed; see (197). For every $0 < \varepsilon < 1$, define the small-noise validity region by

$$\mathcal{I}_{\mathrm{sm}}(\varepsilon) := \{t \in [\underline{t}, \overline{t}] : \sigma_t \le C_{\mathrm{sm}} \varepsilon^{1/\beta}\}. \tag{12}$$

and the large-noise validity region by

$$\mathcal{I}_{\mathrm{lg}}(\varepsilon) := \{t \in [\underline{t}, \overline{t}] : \sigma_t \ge c_{\mathrm{lg}} \varepsilon^{1/\beta}\}. \tag{13}$$

These auxiliary validity regions need not be disjoint. The parameter ε is a branch-scale parameter; target tolerances are named by role, such as coordinate accuracy ε_u, block accuracy $\varepsilon_{\mathrm{blk}}$, and scaled score accuracy $\varepsilon_{\mathrm{sc}}$. In the final large-noise application, the branch scale is $n^{-\beta/(d+2\beta)}$, so the validity condition is $\sigma_t \gtrsim n^{-1/(d+2\beta)}$.

D.2. Branchwise approximation roadmap

There are two approximation mechanisms under Theorem 2.5. On large-noise slabs, the denominator $\mathcal{Q}$ has an explicit lower bound on the near-manifold tube, so the tangent-cell moment approximation can be used directly and the network size is driven by the intrinsic cell count. On the small-noise interval, the denominator is controlled after de-Gaussianization, and the proof uses a projection-centered Laplace expansion together with neural approximation of nearest-point projection maps.

For navigation, the formal dependency structure is:

Component	Formal role
Sec. D.3	Build the projection atlas, partition of unity, chart integrals $\mathcal{Q}, \mathcal{P}, \mathcal{U}$, and the centered ratio $s^* = \sigma_t^{-1}\mathcal{U}/\mathcal{Q}$.
Sec. E	Approximate $\mathcal{Q}$ and $\mathcal{U}$ on tangent cells at radius $r_k = c_0 \sigma_{t_{k-1}} \wedge r_{\star,n}$ under a direct tube lower bound for $\mathcal{Q}$, with $r_{\star,n}$ defined in Theorem G.2.
Theorem E.5	Convert the large-noise chart-cell approximation and denominator floor into a global clipped score approximation on each large-noise slab.
Sec. F	Develop the active projection geometry, Gauss–Newton projection networks, Laplace expansion, and final ratio-network assembly needed at small noise.
Theorems F.27 and F.28	Remove the common normal Gaussian factor, prove a polynomial denominator floor, and approximate the de-Gaussianized chart integrals.
Theorem F.29	Assemble the small-noise projection–Laplace networks into the final clipped score approximation on $[t_0, t_1]$.
Sec. G and H	Insert the deterministic comparison networks into the clipped ERM oracle and then into the slabwise W_1 perturbation argument.

D.3. Projection atlas and chart-integral decomposition

D.3.1. Positive-reach projection charts

Let $\mathcal{B}(\boldsymbol{y}, r)$ denote the open Euclidean ball with radius $r > 0$ and center $\boldsymbol{y} \in \mathbb{R}^D$. We will use projection charts whose validity follows from positive reach.

Lemma D.2 (Projection charts from positive reach)**.** *Let Theorem 2.2 hold. There is a numerical constant $c_{\mathrm{ch}} \in (0, 1/8]$, depending only on d, such that the chart radius may be chosen as*

$$r_{\mathrm{ch}} := c_{\mathrm{ch}}(1 \wedge \kappa).$$

This radius has the following property. For every $\boldsymbol{y} \in \mathcal{M}$, if $\boldsymbol{V}_{\boldsymbol{y}} \in \mathbb{R}^{D\times d}$ has orthonormal columns spanning $T_{\boldsymbol{y}}\mathcal{M}$, then there is an open neighborhood $U_{\boldsymbol{y}} \subset \mathcal{M}$ satisfying

$$\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r_{\mathrm{ch}}/2) \subset U_{\boldsymbol{y}} \subset \mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r_{\mathrm{ch}}),$$

such that the tangent projection

$$\pi_{\boldsymbol{y}} : U_{\boldsymbol{y}} \to \boldsymbol{V}_{\boldsymbol{y}}^\top (U_{\boldsymbol{y}} - \boldsymbol{y}) \subset \mathbb{R}^d, \qquad \pi_{\boldsymbol{y}}(\boldsymbol{z}) := \boldsymbol{V}_{\boldsymbol{y}}^\top (\boldsymbol{z} - \boldsymbol{y}),$$

is a C^∞ diffeomorphism onto its image. Moreover, for every $\boldsymbol{z} \in U_{\boldsymbol{y}}$ and every $\boldsymbol{v} \in T_{\boldsymbol{z}}\mathcal{M}$,

$$(1 - 2r_{\mathrm{ch}}/\kappa)\|\boldsymbol{v}\|_2^2 \le \|\boldsymbol{V}_{\boldsymbol{y}}^\top \boldsymbol{v}\|_2^2 \le \|\boldsymbol{v}\|_2^2. \tag{14}$$

Proof. We use the standard local-graph theorem for manifolds of positive reach. For a closed embedded C^2 manifold of reach at least κ, Federer (Federer, 1959, Theorem 4.8(12)) gives the tubular-neighborhood property for normal segments of length below κ, hence the usual tangent-ball exclusion, while (Federer, 1959, Theorem 4.8(7)) gives the quadratic distance-to-tangent-plane bound. The tangent variation estimate recorded for instance in (Boissonnat et al., 2019, Corollary 3), together with the local inverse theorem, then gives a numerical $c_{\mathrm{ch}} \le 1/8$, depending only on d, such that $\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, c_{\mathrm{ch}}\kappa)$ is a C^∞ graph over $T_{\boldsymbol{y}}\mathcal{M}$. If $\kappa \ge 1$, the smaller radius c_{ch} is also admissible. Thus $r_{\mathrm{ch}} = c_{\mathrm{ch}}(1 \wedge \kappa)$ is admissible uniformly in $\boldsymbol{y}$. Equivalently, the orthogonal projection onto $T_{\boldsymbol{y}}\mathcal{M}$ is one-to-one on this patch and is a C^∞ diffeomorphism onto its image. Thus there is an open set $U_{\boldsymbol{y}} \subset \mathcal{M}$ with

$$\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r_{\mathrm{ch}}/2) \subset U_{\boldsymbol{y}} \subset \mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r_{\mathrm{ch}})$$

such that $\pi_{\boldsymbol{y}}$ is a C^∞ diffeomorphism onto its image.

It remains to record the quantitative derivative bound used below. A standard consequence of reach is the tangent-space variation estimate (Boissonnat et al., 2019, Corollary 3)

$$\|P_{T_{\boldsymbol{z}}\mathcal{M}} - P_{T_{\boldsymbol{y}}\mathcal{M}}\|_{\mathrm{op}} \le \frac{2\|\boldsymbol{z} - \boldsymbol{y}\|_2}{\kappa}, \qquad \|\boldsymbol{z} - \boldsymbol{y}\|_2 < \kappa/2,$$

where P_T denotes the orthogonal projection onto a subspace T. If $\boldsymbol{z} \in U_{\boldsymbol{y}}$ and $\boldsymbol{v} \in T_{\boldsymbol{z}}\mathcal{M}$ with $\|\boldsymbol{v}\|_2 = 1$, then

$$\|P_{(T_{\boldsymbol{y}}\mathcal{M})^\perp}\boldsymbol{v}\|_2 = \|(P_{T_{\boldsymbol{z}}\mathcal{M}} - P_{T_{\boldsymbol{y}}\mathcal{M}})\boldsymbol{v}\|_2 \le \|P_{T_{\boldsymbol{z}}\mathcal{M}} - P_{T_{\boldsymbol{y}}\mathcal{M}}\|_{\mathrm{op}}.$$

Because $r_{\mathrm{ch}} \le \kappa/8$, this last quantity is at most 1. Hence

$$\begin{aligned}\|\boldsymbol{V}_{\boldsymbol{y}}^\top \boldsymbol{v}\|_2^2 &= \|P_{T_{\boldsymbol{y}}\mathcal{M}}\boldsymbol{v}\|_2^2 = 1 - \|P_{(T_{\boldsymbol{y}}\mathcal{M})^\perp}\boldsymbol{v}\|_2^2 \\ &\ge 1 - \|P_{T_{\boldsymbol{z}}\mathcal{M}} - P_{T_{\boldsymbol{y}}\mathcal{M}}\|_{\mathrm{op}} \ge 1 - \frac{2r_{\mathrm{ch}}}{\kappa}.\end{aligned}$$

The upper bound in (14) is immediate because $\boldsymbol{V}_{\boldsymbol{y}}^\top$ is an orthogonal projection. Homogeneity gives the bound for general $\boldsymbol{v}$. □

By compactness, choose points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{C_{\mathcal{M}}} \in \mathcal{M}$ such that

$$\mathcal{M} \subset \bigcup_{i=1}^{C_{\mathcal{M}}} \big(\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}_i, r_{\mathrm{ch}}/2)\big).$$

Let $U_i := U_{\boldsymbol{y}_i}$ and let $\boldsymbol{V}_i := \boldsymbol{V}_{\boldsymbol{y}_i}$. The covering number may be chosen to satisfy the standard reach-volume bound

$$C_{\mathcal{M}} \le C_d\, S_{\mathcal{M}}\, r_{\mathrm{ch}}^{-d} \le C_d\, S_{\mathcal{M}}\, (1 \vee \kappa^{-1})^d, \tag{15}$$

where C_d depends only on d. Indeed, take a maximal $r_{\mathrm{ch}}/4$-separated subset of $\mathcal{M}$; the balls $\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}_i, r_{\mathrm{ch}}/2)$ cover $\mathcal{M}$, while the intrinsic volumes of the disjoint smaller balls are bounded below by $c_d r_{\mathrm{ch}}^d$ because the second fundamental form is bounded by κ^{-1} on the reach scale.

For any $\boldsymbol{y} \in U_i$, define the chart by applying the tangent projection from Theorem D.2 and then rescaling and translating the resulting coordinate image:

$$\phi_i(\boldsymbol{y}) := \eta_i\big(\boldsymbol{V}_i^\top(\boldsymbol{y} - \boldsymbol{y}_i) + \boldsymbol{c}_i\big), \tag{16}$$

where $\eta_i \in (0, 1]$ and $\boldsymbol{c}_i \in \mathbb{R}^d$ are chosen so that $\phi_i(U_i) \subseteq [0, 1]^d$. Indeed,

$$\phi_i(\boldsymbol{y}) = \eta_i\big(\pi_{\boldsymbol{y}_i}(\boldsymbol{y}) + \boldsymbol{c}_i\big), \qquad \pi_{\boldsymbol{y}_i}(\boldsymbol{y}) = \boldsymbol{V}_i^\top(\boldsymbol{y} - \boldsymbol{y}_i).$$

Since $\pi_{\boldsymbol{y}_i}$ is a diffeomorphism on U_i by Theorem D.2, and the map $\boldsymbol{u} \mapsto \eta_i(\boldsymbol{u} + \boldsymbol{c}_i)$ is an invertible affine map of $\mathbb{R}^d$, so is ϕ_i. Note that each ϕ_i is the restriction of an affine map on $\mathbb{R}^D$, and therefore can be exactly represented by a one-hidden-layer ReLU network.

Let

$$\phi_i^{-1} : \phi_i(U_i) \to U_i \subset \mathcal{M}.$$

The following proposition establishes a uniform bound on the volume distortion within each local chart:

Proposition D.3 (Uniform Jacobian bounds for projection charts)**.** *Under Theorem 2.2, the Jacobian determinant of the inverse chart ϕ_i^{-1} satisfies, for every $\boldsymbol{u} \in \phi_i(U_i)$,*

$$\sqrt{\det(\nabla\phi_i^{-1}(\boldsymbol{u})^\top \nabla\phi_i^{-1}(\boldsymbol{u}))} \le \frac{1}{\eta_i^d}\Big(1 - \frac{2r_{\mathrm{ch}}}{\kappa}\Big)^{-d/2}. \tag{17}$$

Proof. Fix i and write $\boldsymbol{z} = \phi_i^{-1}(\boldsymbol{u})$. The derivative of ϕ_i at $\boldsymbol{z}$, restricted to $T_{\boldsymbol{z}}\mathcal{M}$, is

$$\mathrm{D}\phi_i(\boldsymbol{z})|_{T_{\boldsymbol{z}}\mathcal{M}} = \eta_i\, \boldsymbol{V}_i^\top|_{T_{\boldsymbol{z}}\mathcal{M}}.$$

By (14), every singular value of $\boldsymbol{V}_i^\top|_{T_{\boldsymbol{z}}\mathcal{M}}$ is at least $(1 - 2r_{\mathrm{ch}}/\kappa)^{1/2}$ and at most 1. Hence every singular value of $\mathrm{D}\phi_i(\boldsymbol{z})|_{T_{\boldsymbol{z}}\mathcal{M}}$ is at least $\eta_i(1 - 2r_{\mathrm{ch}}/\kappa)^{1/2}$. Since $\nabla\phi_i^{-1}(\boldsymbol{u})$ is the inverse linear map from $\mathbb{R}^d$ to $T_{\boldsymbol{z}}\mathcal{M}$, its singular values are bounded above by $\eta_i^{-1}(1 - 2r_{\mathrm{ch}}/\kappa)^{-1/2}$. Multiplying the d singular values gives (17). □

We define

$$B^{\mathrm{Jac}} := \sup_i \eta_i^{-d}(1 - 2r_{\mathrm{ch}}/\kappa)^{-d/2}. \tag{18}$$

D.3.2. Partition-of-unity amplitudes and density regularity

Let $\{\rho_i\}_{i=1}^{C_{\mathcal{M}}}$ be a C^∞ partition of unity subordinate to the finite projection atlas $\{U_i\}_{i=1}^{C_{\mathcal{M}}}$. After a finite refinement of the projection cover, we choose it with compactly supported chart pieces

$$\mathsf{supp}_{\mathcal{M}}(\rho_i) \Subset U_i, \qquad i = 1, \dots, C_{\mathcal{M}},$$

which is possible because the smaller sets $\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}_i, r_{\mathrm{ch}}/2)$ still cover the compact manifold and each is compactly contained in U_i. Let

$$\mathcal{S}_i := \mathsf{supp}_{\mathcal{M}}(\rho_i) \Subset U_i, \qquad i = 1, \dots, C_{\mathcal{M}}, \tag{19}$$

be the compact support of the partition of unity on chart i. The refinement is chosen so that, for every resulting index, the coordinate support $\phi_i(\mathcal{S}_i)$ is contained in nested axis-parallel coordinate boxes compactly contained in $\phi_i(U_i)$. We fix real numbers

$$a_{i,r}^{\square} < a_{i,r}^{\circ} < b_{i,r}^{\circ} < b_{i,r}^{\square}, \qquad r = 1, \dots, d,$$

so that, with

$$K_i^\circ := \prod_{r=1}^{d} [a_{i,r}^\circ, b_{i,r}^\circ], \qquad \square_i := \prod_{r=1}^{d} [a_{i,r}^\square, b_{i,r}^\square], \tag{20}$$

we have

$$\phi_i(\mathcal{S}_i) \subset \mathrm{int}(K_i^\circ) \subset K_i^\circ \Subset \mathrm{int}(\square_i) \subset \square_i \Subset \phi_i(U_i). \tag{21}$$

These fixed boxes are used both by the tangent-cell construction and, under the density lower bound, by the projection-centered Laplace expansion. This only changes the finite atlas cardinality by a geometry-dependent factor. Since $\mathsf{supp}(\rho_i) \subset U_i$, we may insert $1 = \sum_{i=1}^{C_\mathcal{M}} \rho_i$ into the manifold integral and then change variables chart by chart. For each chart i, define the pulled-back partition weight

$$\bar{\rho}_i(\boldsymbol{u}) := (\rho_i \circ \phi_i^{-1})(\boldsymbol{u}), \qquad \boldsymbol{u} \in \phi_i(U_i). \tag{22}$$

For each chart $i \in \{1, \dots, C_\mathcal{M}\}$, define $\boldsymbol{z}_i : \phi_i(U_i) \to U_i, \boldsymbol{J}_i : \phi_i(U_i) \to \mathbb{R}^{D\times d}, \boldsymbol{G}_i : \phi_i(U_i) \to \mathbb{R}^{d\times d}$:

$$\boldsymbol{z}_i(\boldsymbol{u}) := \phi_i^{-1}(\boldsymbol{u}) \in U_i \subset \mathcal{M}, \tag{23}$$

$$\boldsymbol{J}_i(\boldsymbol{u}) := \nabla \boldsymbol{z}_i(\boldsymbol{u}) \in \mathbb{R}^{D\times d}, \tag{24}$$

$$\boldsymbol{G}_i(\boldsymbol{u}) := \boldsymbol{J}_i(\boldsymbol{u})^\top \boldsymbol{J}_i(\boldsymbol{u}) \in \mathbb{R}^{d\times d}. \tag{25}$$

Then, the density of $\boldsymbol{X}_t = m_t \boldsymbol{X}_0 + \sigma_t \boldsymbol{Z}, \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ is

$$\begin{aligned}
p_t(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \int_\mathcal{M} \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2}{2\sigma_t^2}\Big) p_0(\boldsymbol{y}) \mathrm{dvol}_\mathcal{M}(\boldsymbol{y}) \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_\mathcal{M}} \int_\mathcal{M} \rho_i(\boldsymbol{y}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2}{2\sigma_t^2}\Big) p_0(\boldsymbol{y}) \mathrm{dvol}_\mathcal{M}(\boldsymbol{y}) \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \bar{\rho}_i(\boldsymbol{u}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \phi_i^{-1}(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big) p_0(\phi_i^{-1}(\boldsymbol{u})) \sqrt{(\det \nabla \phi_i^{-1}(\boldsymbol{u})^\top \nabla \phi_i^{-1}(\boldsymbol{u}))} \mathrm{d}\boldsymbol{u} \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \bar{\rho}_i(\boldsymbol{u}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big) p_0(\boldsymbol{z}_i(\boldsymbol{u})) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} \mathrm{d}\boldsymbol{u}
\end{aligned} \tag{26}$$

Lemma D.4 (Hölder regularity in projection coordinates)**.** *Assume $p_0 \in \mathcal{H}^\beta(\mathcal{M}, B_0)$ in the sense of Theorem 2.3. Then each projection-coordinate representative $p_0 \circ \boldsymbol{z}_i$ belongs to $\mathcal{H}^\beta(\phi_i(U_i))$, and*

$$\max_{1\le i\le C_\mathcal{M}} \|p_0 \circ \boldsymbol{z}_i\|_{\mathcal{H}^\beta(\phi_i(U_i))} < \infty.$$

The resulting bound depends on B_0, β, and finitely many $C^{\lceil\beta\rceil}$-norms of transition maps between the original smooth atlas and the projection atlas.

Proof. Fix a projection chart (U_i, ϕ_i). Since $\overline{U_i}$ is compact in $\mathcal{M}$, it is covered by finitely many relatively compact chart neighborhoods $V'_\alpha \Subset V_\alpha$, where (V_α, ψ_α) are charts from the atlas used to define $\mathcal{H}^\beta(\mathcal{M}, B_0)$. Let $W_{i,\alpha} := U_i \cap V'_\alpha$. On $\phi_i(W_{i,\alpha})$,

$$p_0 \circ \boldsymbol{z}_i = (p_0 \circ \psi_\alpha^{-1}) \circ (\psi_\alpha \circ \boldsymbol{z}_i).$$

The first factor has $\mathcal{H}^\beta$-norm at most B_0 by assumption. The transition map $\psi_\alpha \circ \boldsymbol{z}_i$ is C^∞, and all of its derivatives up to order $\lceil\beta\rceil$ are bounded on the compact closure of $\phi_i(W_{i,\alpha})$. The standard composition theorem for Hölder classes, proved by the chain rule for derivatives up to order $\lfloor\beta\rfloor$ and the Hölder bound for the top derivative, gives

$$\|p_0 \circ \boldsymbol{z}_i\|_{\mathcal{H}^\beta(\phi_i(W_{i,\alpha}))} \le C_{i,\alpha} B_0,$$

where $C_{i,\alpha}$ depends only on the stated transition-map norms.

To pass from these local estimates to the full coordinate domain $\phi_i(U_i)$, choose a smooth partition of unity $\{\chi_{i,\alpha}\}_\alpha$ on U_i subordinate to the finite cover $\{W_{i,\alpha}\}_\alpha$, and write

$$p_0 \circ \boldsymbol{z}_i = \sum_\alpha \big((\chi_{i,\alpha} p_0) \circ \boldsymbol{z}_i\big).$$

Each summand has compact support in $\phi_i(W_{i,\alpha})$. Extending it by zero to $\phi_i(U_i)$ preserves its $\mathcal{H}^\beta$-norm up to a constant depending only on the cutoff derivatives and the positive buffer between $\mathsf{supp}\chi_{i,\alpha}$ and $U_i \setminus W_{i,\alpha}$. The algebra property of Hölder spaces controls multiplication by $\chi_{i,\alpha} \circ \boldsymbol{z}_i$. Since the number of subpatches is finite, summing the extended pieces gives a finite $\mathcal{H}^\beta(\phi_i(U_i))$ bound. Taking the maximum over the finite projection atlas gives the claim. □

By Theorem D.4, the pulled-back density has a finite chartwise Hölder norm. We write

$$B_p^{\mathrm{proj}} := \max_{1 \le i \le C_{\mathcal{M}}} \|p_0 \circ \boldsymbol{z}_i\|_{\mathcal{H}^\beta(\phi_i(U_i))} < \infty. \tag{27}$$

The constant B_p^{proj} depends on B_0, β, and the finite atlas transition maps; it is generally not equal to B_0.

Similarly to $p_t(\boldsymbol{x})$, the gradient of the density function is given by

$$\begin{aligned}\nabla p_t(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \int_{\mathcal{M}} \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2}{2\sigma_t^2}\Big)\Big(-\frac{\boldsymbol{x} - m_t \boldsymbol{y}}{\sigma_t^2}\Big) p_0(\boldsymbol{y}) \mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y}) \\ &= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_{\mathcal{M}}} \int_{\phi_i(U_i)} \bar{\rho}_i(\boldsymbol{u}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big)\Big(-\frac{\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})}{\sigma_t^2}\Big) p_0(\boldsymbol{z}_i(\boldsymbol{u})) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} \mathrm{d}\boldsymbol{u}.\end{aligned} \tag{28}$$

Define $a_i^{(\mathcal{Q})} : \phi_i(U_i) \to \mathbb{R}$ and $\boldsymbol{a}_i^{(\mathcal{P})} : \phi_i(U_i) \to \mathbb{R}^D$:

$$a_i^{(\mathcal{Q})}(\boldsymbol{u}) := \bar{\rho}_i(\boldsymbol{u}) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} p_0(\boldsymbol{z}_i(\boldsymbol{u})), \qquad \boldsymbol{a}_i^{(\mathcal{P})}(\boldsymbol{u}) := \boldsymbol{z}_i(\boldsymbol{u}) a_i^{(\mathcal{Q})}(\boldsymbol{u}). \tag{29}$$

For $k = 1, \ldots, D$, write $a_i^{(\mathcal{P}_k)} := a_{i,k}^{(\mathcal{P})}$, and define the integral-type set

$$\mathcal{R} := \{\mathcal{Q}, \mathcal{P}_1, \ldots, \mathcal{P}_D\}.$$

Thus $R \in \mathcal{R}$ denotes a type, not the already summed global integral; $a_i^{(R)}$ denotes $a_i^{(\mathcal{Q})}$ when $R = \mathcal{Q}$, and $a_i^{(\mathcal{P}_k)}$ when $R = \mathcal{P}_k$.

The following lemma shows that $a_i^{(\mathcal{Q})}$ and $\boldsymbol{a}_i^{(\mathcal{P})}$ are β-Hölder smooth on $\phi_i(U_i)$:

Lemma D.5 (Chart-amplitude Hölder regularity)**.** *Assume Theorems 2.2 and 2.4 hold. There is a finite constant B_a such that*

$$a_i^{(\mathcal{Q})}, a_{i,k}^{(\mathcal{P})} \in \mathcal{H}^\beta(\phi_i(U_i), B_a), \qquad k = 1, \ldots, D. \tag{30}$$

One may take

$$B_a = c_\beta B_p^{\mathrm{proj}} (1 \vee B^{\mathrm{Jac}}) \left(1 \vee \max_{i,k} \max_{\|\boldsymbol{\alpha}\|_1 \le \lceil \beta \rceil} \Big\{ \|\partial^{\boldsymbol{\alpha}} \bar{\rho}_i\|_\infty \vee \|\partial^{\boldsymbol{\alpha}} \sqrt{\det \boldsymbol{G}_i}\|_\infty \vee \|\partial^{\boldsymbol{\alpha}} z_{i,k}\|_\infty \Big\}\right).$$

Moreover, because the manifold is embedded in $[0,1]^D$, we have the coordinate-wise bounds:

$$|a_{i,k}^{(\mathcal{P})}(\boldsymbol{u})| \le a_i^{(\mathcal{Q})}(\boldsymbol{u}) \le B_p^{\mathrm{proj}} B^{\mathrm{Jac}}. \tag{31}$$

Proof of Theorem D.5. **Step 1: Uniform L^∞ bounds.** Because $p_0 \circ \boldsymbol{z}_i$ is β-Hölder smooth on the parameter space $\phi_i(U_i) \subset [0,1]^d$. Specifically, its Hölder norm (and consequently its supremum norm) is bounded by:

$$p_0 \circ \boldsymbol{z}_i \in \mathcal{H}^\beta(\phi_i(U_i), B_p^{\mathrm{proj}}), \quad \text{for each } i = 1, 2, \ldots, C_{\mathcal{M}}. \tag{32}$$

By definition of the partition of unity, $0 \le \bar{\rho}_i(\boldsymbol{u}) \le 1$ and by Theorem D.3, the Jacobian determinant is uniformly bounded by B^{Jac}. Multiplying these supremum bounds yields:

$$|a_i^{(\mathcal{Q})}(\boldsymbol{u})| = |\bar{\rho}_i(\boldsymbol{u})\sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})}p_0(\boldsymbol{z}_i(\boldsymbol{u}))| \le 1 \cdot B^{\mathrm{Jac}} \cdot B_p^{\mathrm{proj}}.$$

Since the manifold is contained in $[0,1]^D$, we have $|z_{i,k}(\boldsymbol{u})| \le 1$, which immediately implies $|a_{i,k}^{(\mathcal{P})}(\boldsymbol{u})| \le a_i^{(\mathcal{Q})}(\boldsymbol{u})$, verifying (31).

Step 2: Hölder regularity and the explicit bound B_a. Because the manifold and the partition of unity are C^∞, the geometric multipliers $\bar{\rho}_i$, $\boldsymbol{z}_i$, and $\sqrt{\det \boldsymbol{G}_i}$ (which is smooth since $\boldsymbol{G}_i$ is strictly positive definite) possess uniformly bounded derivatives of all orders on the compact chart supports.

Let C_{geom} be one plus the maximum $C^{\lceil\beta\rceil}$ norm of these geometric factors across all finitely many charts $i \in \{1,\dots,C_{\mathcal{M}}\}$ and coordinates $k \in \{1,\dots,D\}$:

$$C_{\mathrm{geom}} := 1 \vee \max_{i,k} \max_{|\boldsymbol{\alpha}|\le\lceil\beta\rceil} \Big\{ \|\partial^{\boldsymbol{\alpha}}\bar{\rho}_i\|_\infty \vee \|\partial^{\boldsymbol{\alpha}}\sqrt{\det \boldsymbol{G}_i}\|_\infty \vee \|\partial^{\boldsymbol{\alpha}} z_{i,k}\|_\infty \Big\}.$$

Consequently, the $\mathcal{H}^\beta$ norms of $\bar{\rho}_i$, $\sqrt{\det \boldsymbol{G}_i}$, and $z_{i,k}$ are all strictly bounded by C_{geom}.

We invoke the algebra property of Hölder spaces: for any $f, g \in \mathcal{H}^\beta$, there exists a constant c_β such that $\|fg\|_{\mathcal{H}^\beta} \le c_\beta(\|f\|_{\mathcal{H}^\beta}\|g\|_\infty + \|f\|_\infty\|g\|_{\mathcal{H}^\beta})$. Applying this iteratively to the product $a_i^{(\mathcal{Q})} = (\bar{\rho}_i\sqrt{\det \boldsymbol{G}_i}) \cdot (p_0 \circ \boldsymbol{z}_i)$, we obtain:

$$\begin{aligned}
\|a_i^{(\mathcal{Q})}\|_{\mathcal{H}^\beta} &\le c_\beta^2 \Big(\|\bar{\rho}_i\|_{\mathcal{H}^\beta}\|\sqrt{\det \boldsymbol{G}_i}\|_\infty + \|\bar{\rho}_i\|_\infty\|\sqrt{\det \boldsymbol{G}_i}\|_{\mathcal{H}^\beta} \Big) \|p_0 \circ \boldsymbol{z}_i\|_\infty + c_\beta\|\bar{\rho}_i\sqrt{\det \boldsymbol{G}_i}\|_\infty\|p_0 \circ \boldsymbol{z}_i\|_{\mathcal{H}^\beta} \\
&\le c_\beta^2 \big(C_{\mathrm{geom}} B^{\mathrm{Jac}} + 1 \cdot C_{\mathrm{geom}} \big) B_p^{\mathrm{proj}} + c_\beta(1 \cdot B^{\mathrm{Jac}}) B_p^{\mathrm{proj}} \\
&\le c_\beta^2 B_p^{\mathrm{proj}} \Big[C_{\mathrm{geom}}(B^{\mathrm{Jac}} + 1) + B^{\mathrm{Jac}} \Big].
\end{aligned}$$

Similarly, for $a_{i,j}^{(\mathcal{P})} = z_{i,j} a_i^{(\mathcal{Q})}$, applying the product rule again yields:

$$\begin{aligned}
\|a_{i,j}^{(\mathcal{P})}\|_{\mathcal{H}^\beta} &\le c_\beta \Big(\|z_{i,j}\|_{\mathcal{H}^\beta}\|a_i^{(\mathcal{Q})}\|_\infty + \|z_{i,j}\|_\infty\|a_i^{(\mathcal{Q})}\|_{\mathcal{H}^\beta} \Big) \\
&\le c_\beta \Big(C_{\mathrm{geom}} B^{\mathrm{Jac}} B_p^{\mathrm{proj}} + 1 \cdot \|a_i^{(\mathcal{Q})}\|_{\mathcal{H}^\beta} \Big).
\end{aligned}$$

Taking the maximum of these explicit bounds over all charts and coordinates, we may choose the universal Hölder constant as:

$$B_a := c'_\beta B_p^{\mathrm{proj}} C_{\mathrm{geom}}(1 \vee B^{\mathrm{Jac}}).$$

□

D.3.3. Chart-integral representation of density and score

Define $F_{i,\boldsymbol{x},t} : \phi_i(U_i) \to \mathbb{R}$:

$$F_{i,\boldsymbol{x},t}(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2. \tag{33}$$

Fix $n \ge 2$. In statements using polynomial regimes $\varepsilon \ge n^{-a_\varepsilon}$ and $D \vee \sigma_{\underline{t}}^{-1} \le n^{a_{\mathrm{par}}}$, regard $a_\varepsilon, a_{\mathrm{par}}$ as fixed. Choose once and for all a threshold $C_\varrho = C_\varrho(a_\varepsilon, a_{\mathrm{par}}) \ge 1$, independent of D, n, $\sigma_{\underline{t}}$, and all numerical ε-values after those exponents are fixed, large enough that the off-manifold deviation and Fisher bounds in Theorems C.1 and C.2 hold with threshold $C_\varrho \sigma_t \sqrt{D \vee \log n}$, with the polynomial tail exponent required by the branchwise approximation and estimation arguments. This same constant determines the near-manifold domain used throughout the branchwise near-manifold approximation results. Enlarge the fixed tube-volume constant C_{tube} from (9), if necessary, so that $C_{\mathrm{tube}} \ge C_\varrho$. Dependence on the tube radius is tracked through the displayed quantity M_x^ϱ below; it is not hidden in constants that are declared independent of D and n.

$$\varrho_t := C_\varrho \sigma_t \sqrt{D \vee \log n}, \tag{34}$$

For each $t \in [\underline{t}, \overline{t}]$, define the time-indexed near-manifold tube

$$A_t := \{\boldsymbol{x} \in \mathbb{R}^D : \mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le \varrho_t\}. \tag{35}$$

Thus a condition such as $\boldsymbol{x} \in A_t$ is always interpreted at the same time t. Since $\mathcal{M} \subset [0,1]^D$, $m_t \le 1$, and $\sigma_t \le 1$, every $t \in [\underline{t}, \overline{t}]$ and $\boldsymbol{x} \in A_t$ satisfy

$$\|\boldsymbol{x}\|_\infty \le 1 + C_\varrho \sqrt{D \vee \log n}.$$

We therefore write

$$M_x^\varrho := 1 + C_\varrho \sqrt{D \vee \log n}. \tag{36}$$

For $i \in \{1, \dots, C_\mathcal{M}\}$, define the chart integrals for all pairs $(\boldsymbol{x}, t)$ with $t \in [\underline{t}, \overline{t}]$ and $\boldsymbol{x} \in A_t$:

$$\mathcal{P}_i(\boldsymbol{x}, t) := \sigma_t^{-d} \int_{\phi_i(U_i)} \exp\Big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\Big) \boldsymbol{a}_i^{(\mathcal{P})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \tag{37}$$

$$\mathcal{Q}_i(\boldsymbol{x}, t) := \sigma_t^{-d} \int_{\phi_i(U_i)} \exp\Big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\Big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \tag{38}$$

Let

$$\mathcal{P}(\boldsymbol{x}, t) := \sum_{i=1}^{C_\mathcal{M}} \mathcal{P}_i(\boldsymbol{x}, t) \in \mathbb{R}^D, \qquad \mathcal{Q}(\boldsymbol{x}, t) := \sum_{i=1}^{C_\mathcal{M}} \mathcal{Q}_i(\boldsymbol{x}, t) \in \mathbb{R}.$$

Define the centered local and global numerator integrals

$$\begin{aligned} \mathcal{U}_i(\boldsymbol{x}, t) &:= \frac{m_t \mathcal{P}_i(\boldsymbol{x}, t) - \boldsymbol{x}\mathcal{Q}_i(\boldsymbol{x}, t)}{\sigma_t} \\ &= \sigma_t^{-d} \int_{\phi_i(U_i)} \frac{m_t \boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{x}}{\sigma_t} \exp\Big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\Big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \end{aligned} \tag{39}$$

$$\mathcal{U}(\boldsymbol{x}, t) := \sum_{i=1}^{C_\mathcal{M}} \mathcal{U}_i(\boldsymbol{x}, t) = \frac{m_t \mathcal{P}(\boldsymbol{x}, t) - \boldsymbol{x}\mathcal{Q}(\boldsymbol{x}, t)}{\sigma_t}. \tag{40}$$

Then

$$s^*(\boldsymbol{x}, t) = \nabla \log p_t(\boldsymbol{x}) = \frac{1}{\sigma_t} \frac{\mathcal{U}(\boldsymbol{x}, t)}{\mathcal{Q}(\boldsymbol{x}, t)} \qquad (p_t(\boldsymbol{x}) > 0). \tag{41}$$

For a chart-integral floor $C_t > 0$, the centered floor-regularized score is

$$s_{\mathrm{reg},C_t}(\boldsymbol{x}, t) := \frac{1}{\sigma_t} \frac{\mathcal{U}(\boldsymbol{x}, t)}{\mathcal{Q}(\boldsymbol{x}, t) \vee C_t}. \tag{42}$$

For reference, by (26) and (28), the same regularized score can also be written in the uncentered form

$$\begin{aligned} \frac{\nabla p_t(\boldsymbol{x})}{p_t(\boldsymbol{x}) \vee \rho_t} &= \frac{\frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \exp\big(-\frac{\|\boldsymbol{x} - m_t \phi_i^{-1}(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\big) \big(-\frac{\boldsymbol{x} - m_t \phi_i^{-1}(\boldsymbol{u})}{\sigma_t^2}\big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}}{\frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \exp\big(-\frac{\|\boldsymbol{x} - m_t \phi_i^{-1}(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \vee \rho_t} \\ &= -\frac{\boldsymbol{x}}{\sigma_t^2} \frac{\sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \exp\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}}{\sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \exp\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \vee (2\pi\sigma_t^2)^{D/2} \rho_t} \\ &\quad + \frac{m_t}{\sigma_t^2} \frac{\sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \exp\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big) \boldsymbol{a}_i^{(\mathcal{P})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}}{\sum_{i=1}^{C_\mathcal{M}} \int_{\phi_i(U_i)} \exp\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u} \vee (2\pi\sigma_t^2)^{D/2} \rho_t} \\ &= -\frac{\boldsymbol{x}}{\sigma_t^2} \frac{\mathcal{Q}(\boldsymbol{x}, t)}{\mathcal{Q}(\boldsymbol{x}, t) \vee C_t} + \frac{m_t}{\sigma_t^2} \frac{\mathcal{P}(\boldsymbol{x}, t)}{\mathcal{Q}(\boldsymbol{x}, t) \vee C_t}, \end{aligned} \tag{43}$$

where

$$C_t := \sigma_t^{-d}(2\pi\sigma_t^2)^{D/2}\rho_t. \tag{44}$$

Conversely, choosing the chart-integral floor C_t fixes the density-level floor as

$$\rho_t = (2\pi\sigma_t^2)^{-D/2}\sigma_t^d C_t. \tag{45}$$

Lemma D.6 (Uniform bounds for chart moment and density integrals)**.** *Let*

$$V_{\text{chart}} := 1 \vee \sum_{i=1}^{C_{\mathcal{M}}} \text{vol}(\phi_i(U_i)) < \infty.$$

There is a finite $M_0 := B_p^{\text{proj}} B^{\text{Jac}} \sigma_{\underline{t}}^{-d} V_{\text{chart}}$ *such that*

$$0 \le \mathcal{Q}(\boldsymbol{x}, t) \le M_0, \qquad |\mathcal{P}_k(\boldsymbol{x}, t)| \le \mathcal{Q}(\boldsymbol{x}, t) \le M_0, \quad k = 1, \dots, D. \tag{46}$$

Proof of Theorem D.6. We first show $|\mathcal{P}_k(\boldsymbol{x}, t)| \le \mathcal{Q}(\boldsymbol{x}, t)$. Since $\mathcal{M} \subset [0,1]^D$, $0 \le z_{i,k}(\boldsymbol{u}) \le 1$, so $|a_{i,k}^{(\mathcal{P})}(\boldsymbol{u})| \le a_i^{(\mathcal{Q})}(\boldsymbol{u})$ and $|\mathcal{P}_{i,k}| \le \mathcal{Q}_i$. Summing over i gives $|\mathcal{P}_k| \le \mathcal{Q}$.

We next prove the uniform upper bound for $\mathcal{Q}$. Note that $\exp\big(-F_{i,\boldsymbol{x},t}(\boldsymbol{u})/\sigma_t^2\big) \le 1$, $\sigma_t^{-d} \le \sigma_{\underline{t}}^{-d}$, and by (31) we have

$$\mathcal{Q}_i(\boldsymbol{x}, t) \le B_p^{\text{proj}} B^{\text{Jac}} \sigma_{\underline{t}}^{-d} \text{vol}(\phi_i(U_i)).$$

Summing over charts gives the displayed M_0. □

D.4. Centered ratio target and stability

We use the integral-type set $\mathcal{R} = \{\mathcal{Q}, \mathcal{P}_1, \dots, \mathcal{P}_D\}$ fixed after (29). In this subsection R_i denotes the corresponding local chart component: $R_i = \mathcal{Q}_i$ when $R = \mathcal{Q}$, and $R_i = \mathcal{P}_{i,k}$ when $R = \mathcal{P}_k$. The common cell notation below is used by the tangent-cell construction and by the density-lower-bound projection–Laplace construction. The method-specific approximation results are stated in the following sections.

D.4.1. Intrinsic cell notation and density floors

For each chart, fix the common compact support envelope

$$K_i^{\text{cell}} := \phi_i(\mathcal{S}_i). \tag{47}$$

By (21), $K_i^{\text{cell}} \Subset \square_i$. Moreover, for every $R \in \mathcal{R}$, the amplitude $a_i^{(R)}$ vanishes outside K_i^{cell}, because each amplitude contains the factor $\bar{\rho}_i$.

For $R \in \mathcal{R}$, define the chart integral

$$R_i(\boldsymbol{x}, t) := \sigma_t^{-d} \int_{K_i^{\text{cell}}} \exp\left(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\right) a_i^{(R)}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \tag{48}$$

We use the same amplitude as a function on $\square_i$; by Theorem D.5 on $\phi_i(U_i)$ and the restriction $\square_i \Subset \phi_i(U_i)$, its restriction satisfies $a_i^{(R)}|_{\square_i} \in \mathcal{H}^\beta(\square_i, B_a)$. Since $K_i^{\text{cell}} \Subset \square_i$, set

$$\delta_i^{\text{cell}} := \text{dist}(K_i^{\text{cell}}, \partial\square_i) > 0.$$

For every $0 < h \le \delta_i^{\text{cell}}/(2\sqrt{d})$, let $\mathcal{Q}_{i,h}$ be the finite family of half-open cubes

$$Q_\nu = \boldsymbol{u}_\nu + h[0,1)^d$$

from the standard h-lattice whose closures intersect K_i^{cell}. Then

$$K_i^{\text{cell}} \subset \bigcup_{\nu \in \mathcal{Q}_{i,h}} \overline{Q}_\nu \subset \square_i, \qquad |\mathcal{Q}_{i,h}| \le C_i h^{-d}.$$

Indeed, every point in such a cube is within distance $\sqrt{d}\,h$ of K_i^{cell}, and the cardinality bound follows by comparing the volume of the $\sqrt{d}\,h$-neighborhood of the fixed compact set K_i^{cell} with the cube volume h^d. Since $a_i^{(R)}$ vanishes outside K_i^{cell}, integration over K_i^{cell} is equivalent to the sum of the integrals over these full cubes; cube boundaries have Lebesgue measure zero.

E. Large-Noise Tangent-Cell Score Approximation

With a density lower bound, the unrefactored denominator $\mathcal{Q}$ has an explicit tube lower bound in the large-noise regime. Thus the separated tangent-cell moment approximation can be used directly for $\mathcal{Q}$ and $\mathcal{U}$, followed by an ordinary reciprocal network.

Large-noise proof architecture. The target result is Theorem E.5. The proof has four modules: derive tangent-cell phase bounds; approximate each chart integral with cells of radius $r_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$ in the statistical application (Theorem G.2); use the density lower bound to construct the reciprocal of $\mathcal{Q}$; and assemble the centered ratio $\mathcal{U}/\mathcal{Q}$ with clipping. The important size fact is that the cell count is explicitly recorded as $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$ rather than hidden in asymptotic notation.

E.1. Tangent-cell phase decomposition and gated chart approximation

The tangent-cell construction follows the tangent-space moment idea. The network is fixed in advance: every cell is represented by a gated subnetwork, and the gate is an explicit ReLU ramp in a computable phase proxy. Thus no input-dependent choice of active cells is made after the network has been constructed.

Lemma E.1 (Tangent-cell Gaussian phase bounds)**.** *Let ν be a cell index in a finite coordinate-cube partition of $\square_i$. Let $Q_\nu \subset \square_i$ be the corresponding cube of side length h, let $\boldsymbol{u}_\nu$ be its center, and set $\boldsymbol{y}_\nu = \boldsymbol{z}_i(\boldsymbol{u}_\nu)$. Let $\mathsf{P}_\nu^{\tan}$ be the orthogonal linear projection in $\mathbb{R}^D$ onto the tangent space $T_{\boldsymbol{y}_\nu}\mathcal{M}$. Define*

$$\mathsf{T}_\nu(\boldsymbol{x},t) := m_t\boldsymbol{y}_\nu + \mathsf{P}_\nu^{\tan}(\boldsymbol{x} - m_t\boldsymbol{y}_\nu), \qquad \Phi_\nu^0(\boldsymbol{x},t) := \frac{\|\boldsymbol{x} - m_t\boldsymbol{y}_\nu\|_2^2}{2\sigma_t^2}.$$

The symbol T_ν denotes an affine tangent-plane linearization center; it is not the nearest-point projection $\Pi_{\mathcal{M}}$ used in the small-noise branch. There are constants $c_{\mathrm{ph}}, C_{\mathrm{ph}} > 0$, depending only on the fixed atlas and on $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$, such that if $h \le c_{\mathrm{ph}}\sigma_t$, then the following hold for all $\boldsymbol{u} \in Q_\nu$, $\boldsymbol{y} = \boldsymbol{z}_i(\boldsymbol{u})$. First, if $\Phi_\nu^0(\boldsymbol{x},t) \le H$ with $H \ge 1$, then

$$\frac{\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2}{\sigma_t} \le C_{\mathrm{ph}} H^{1/2}, \tag{49}$$

$$\frac{|T_\nu(\boldsymbol{x},\boldsymbol{y},t)|}{\sigma_t^2} \le C_{\mathrm{ph}}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}\, h\, H^{1/2}, \tag{50}$$

$$\frac{m_t^2 D_\nu(\boldsymbol{x},\boldsymbol{y},t)}{2\sigma_t^2} \le C_{\mathrm{ph}} H. \tag{51}$$

Second, if $\Phi_\nu^0(\boldsymbol{x},t) \ge AH$ for a sufficiently large numerical constant $A \ge A_{\mathrm{ph}}$, then

$$\frac{\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2^2}{2\sigma_t^2} \ge c_{\mathrm{ph}} AH. \tag{52}$$

Proof. The chart map is C^2 with norm bounded by $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$. Hence, for $\boldsymbol{u} \in Q_\nu$,

$$\boldsymbol{z}_i(\boldsymbol{u}) = \boldsymbol{y}_\nu + \boldsymbol{J}_i(\boldsymbol{u}_\nu)(\boldsymbol{u} - \boldsymbol{u}_\nu) + \boldsymbol{e}_\nu(\boldsymbol{u}), \qquad \|\boldsymbol{e}_\nu(\boldsymbol{u})\|_2 \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}} h^2.$$

The vector $\boldsymbol{J}_i(\boldsymbol{u}_\nu)(\boldsymbol{u} - \boldsymbol{u}_\nu)$ lies in $T_{\boldsymbol{y}_\nu}\mathcal{M}$, while $\boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t)$ is orthogonal to this tangent space. Therefore

$$T_\nu(\boldsymbol{x},\boldsymbol{y},t) = \langle \boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t), \mathsf{T}_\nu(\boldsymbol{x},t) - m_t\boldsymbol{y}\rangle$$

only sees the quadratic normal error $\boldsymbol{e}_\nu$ in the cell, and

$$|T_\nu(\boldsymbol{x},\boldsymbol{y},t)| \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}\, m_t h^2 \|\boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t)\|_2.$$

If $\Phi_\nu^0 \le H$, then

$$\|\boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t)\|_2 \le \|\boldsymbol{x} - m_t\boldsymbol{y}_\nu\|_2 \le \sqrt{2H}\,\sigma_t.$$

Since $h \le c_{\mathrm{ph}}\sigma_t$ and $m_t \le 1$, this proves (50). The same triangle inequality gives

$$\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2 \le \|\boldsymbol{x} - m_t\boldsymbol{y}_\nu\|_2 + m_t\|\boldsymbol{y} - \boldsymbol{y}_\nu\|_2 \le C\sigma_t H^{1/2},$$

after decreasing c_{ph}, which proves (49). Moreover, since $m_t^2 D_\nu = \|\mathsf{T}_\nu(\boldsymbol{x},t) - m_t\boldsymbol{y}\|_2^2$, we avoid any division by the possibly small schedule value m_t. Indeed,

$$\mathsf{T}_\nu(\boldsymbol{x},t) - m_t\boldsymbol{y} = \mathsf{P}_\nu^{\tan}(\boldsymbol{x} - m_t\boldsymbol{y}_\nu) - m_t\boldsymbol{J}_i(\boldsymbol{u}_\nu)(\boldsymbol{u} - \boldsymbol{u}_\nu) - m_t\boldsymbol{e}_\nu(\boldsymbol{u}),$$

so the right-hand side is bounded by $C(\|\boldsymbol{x} - m_t\boldsymbol{y}_\nu\|_2 + h)$. This gives (51) because $h \le c_{\mathrm{ph}}\sigma_t$ and $\Phi_\nu^0 \le H$.

For the tail statement, if $\Phi_\nu^0 \ge AH$, then

$$\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2 \ge \|\boldsymbol{x} - m_t\boldsymbol{y}_\nu\|_2 - m_t\|\boldsymbol{y} - \boldsymbol{y}_\nu\|_2 \ge \sigma_t\sqrt{2AH} - Ch.$$

Choosing A_{ph} large and c_{ph} small gives $\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2^2/(2\sigma_t^2) \ge c_{\mathrm{ph}}AH$. □

Lemma E.2 (Scalar exponential Taylor bounds)**.** *There are universal constants $A_{\exp}, c_{\exp} > 0$ such that the following hold. If $0 \le z \le H$ and $J \ge A_{\exp}H$, then*

$$\left| e^{-z} - \sum_{\ell=0}^{J} \frac{(-z)^\ell}{\ell!} \right| \le e^{-c_{\exp}J}. \tag{53}$$

If $|z| \le a \le 1$, then for every integer $K \ge 0$,

$$\left| e^{-z} - \sum_{\ell=0}^{K} \frac{(-z)^\ell}{\ell!} \right| \le C_K a^{K+1} e^a. \tag{54}$$

Proof. For (54), Taylor's theorem with integral remainder gives the claim. For (53), the Lagrange remainder is bounded by

$$\frac{z^{J+1}}{(J+1)!} \le \left(\frac{eH}{J+1}\right)^{J+1}.$$

Taking $J \ge A_{\exp}H$ with $A_{\exp} > 2e$ makes the last display at most $e^{-c_{\exp}J}$. □

E.2. Scale choices for tangent-cell large-noise approximation

On a large-noise slab $I_k = [t_{k-1}, t_k]$, the cell radius is chosen at the slab noise scale until it reaches the accuracy-limited radius required by the fixed-order tangent-cell expansion. In Theorem G.2 this is implemented as $r_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$, so the per-slab cell count is controlled by $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$. The centered residual structure then permits the sharper scaled residual tolerance $n^{-(\beta+1)/(d+2\beta)}$, so the score error is $\sigma_t^{-2}n^{-2(\beta+1)/(d+2\beta)}$ rather than $\sigma_t^{-2}n^{-2\beta/(d+2\beta)}$. Under Theorem 2.5, the denominator lower bound in Theorem E.4 stabilizes the ratio step; the remaining accuracy dependence enters through the explicit finite-order cell-radius cap. Consequently the switched large-noise network has total leading intrinsic size $\sum_k \sigma_{t_{k-1}}^{-d}$ plus the accuracy-cap contribution. The latter is explicitly propagated through the learned-score and W_1 bounds rather than hidden inside a polynomial factor.

Proposition E.3 (Tangent-cell chart approximation with separated cell scale and accuracy)**.** *Assume Theorems 2.2 and 2.4. Let q_{geom} and $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$ be as in Theorem B.5. Fix a cell-scale parameter $0 < \varepsilon_{\mathrm{cell}} \le 1/2$, a denominator floor $q_{\mathrm{floor}} \in (0,1]$, and an accuracy parameter $0 < \varepsilon_{\mathrm{rel}} \le 1/16$. Put*

$$r_{\mathrm{cell}} := \varepsilon_{\mathrm{cell}}^{1/\beta}, \qquad \mathcal{I}_{\mathrm{lg}}^{\mathrm{cell}} := \{t \in [\underline{t}, \bar{t}] : \sigma_t \ge c_{\mathrm{lg}} r_{\mathrm{cell}}\},$$

and define

$$H_{\mathrm{lg}}^{\mathrm{dec}} := 1 + \log\Big(C_{\mathrm{lg}}\Gamma_{\mathcal{M},q_{\mathrm{geom}}} B_0^{C_{\mathrm{lg}}}(D \vee \log n)^{C_{\mathrm{lg}}}\sigma_{\underline{t}}^{-d} q_{\mathrm{floor}}^{-1}\varepsilon_{\mathrm{rel}}^{-1}\Big). \tag{55}$$

Assume the fixed-order geometry remainder is below the requested tolerance and the linear tangent phase is kept in the small-argument regime, for a sufficiently small numerical constant $c_{\mathrm{ph,cell}} > 0$:

$$\Gamma_{\mathcal{M},q_{\mathrm{geom}}} r_{\mathrm{cell}}(H_{\mathrm{lg}}^{\mathrm{dec}})^{1/2} \le c_{\mathrm{ph,cell}}, \qquad \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} r_{\mathrm{cell}}^{q_{\mathrm{geom}}-2}(H_{\mathrm{lg}}^{\mathrm{dec}})^{C} \le c\varepsilon_{\mathrm{rel}}. \tag{56}$$

Then, for every chart i, there exist ReLU networks $\widetilde{\mathcal{Q}}_i^{\mathrm{lg}}$ and $\widetilde{\mathcal{U}}_{i,k}^{\mathrm{lg}}$, $k = 1, \dots, D$, such that, uniformly for $t \in \mathcal{I}_{\mathrm{lg}}^{\mathrm{cell}}$ and $\boldsymbol{x} \in A_t$,

$$|\mathcal{Q}_i(\boldsymbol{x},t) - \widetilde{\mathcal{Q}}_i^{\mathrm{lg}}(\boldsymbol{x},t)| \vee \max_{1\le k\le D} |\mathcal{U}_{i,k}(\boldsymbol{x},t) - \widetilde{\mathcal{U}}_{i,k}^{\mathrm{lg}}(\boldsymbol{x},t)| \le \varepsilon_{\mathrm{rel}}\big(\mathcal{Q}_i(\boldsymbol{x},t) + q_{\mathrm{floor}}/C_{\mathcal{M}}\big). \tag{57}$$

Moreover, the depth, log-weight bound, width, and sparsity satisfy the following bounds, with the direct ambient powers displayed explicitly.

$$\begin{aligned} L &\le C_{\lg}\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{C} (H^{\mathrm{dec}}_{\lg})^{C}, \\ \log B &\le C_{\lg}\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{C} (H^{\mathrm{dec}}_{\lg})^{C}, \\ \|\boldsymbol{W}\|_\infty &\le C_{\lg}\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{C} D^{4(q_{\mathrm{geom}}+2)} (D \vee \log n)^{4(q_{\mathrm{geom}}+2)} \varepsilon_{\mathrm{cell}}^{-d/\beta} (H^{\mathrm{dec}}_{\lg})^{C}, \\ S &\le C_{\lg}\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{C} D^{4(q_{\mathrm{geom}}+2)} (D \vee \log n)^{4(q_{\mathrm{geom}}+2)} \varepsilon_{\mathrm{cell}}^{-d/\beta} (H^{\mathrm{dec}}_{\lg})^{C}. \end{aligned} \tag{58}$$

Thus, for this separated tangent-cell moment approximation, the cell count is determined by r_{cell}, and the conclusion is valid exactly on the range where (56) holds. If the slab-scale radius does not satisfy that condition, the cell radius must be capped by the accuracy-limited radius used in Theorem G.2. When r_{cell} is specialized to $c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$ in the large-noise estimation theorem, the resulting additional term $r_{\star,n}^{-d}$ is recorded explicitly in the learned-score stochastic bound.

Proof. Fix a chart i. Choose once and for all a fixed number $c_h > 0$ such that

$$c_h \le 1 \wedge c_{\mathrm{ph}} c_{\lg} \wedge \min_{1\le j\le C_{\mathcal{M}}} \frac{\delta_j^{\mathrm{cell}}}{2\sqrt{d}}.$$

This choice depends only on the fixed atlas, q_{geom}, and the branch constants, and is independent of $\varepsilon_{\mathrm{cell}}$, t, $\boldsymbol{x}$, and n. Partition the fixed compact coefficient support $K_i^{\mathrm{cell}} \Subset \square_i$ into coordinate cubes $\{Q_\nu : \nu \in \mathcal{V}_i\}$ of side length $h = c_h r_{\mathrm{cell}}$. Here ν is a cell label in the finite index set $\mathcal{V}_i$. The number of cells is $N_{\mathrm{cell}} \le C\varepsilon_{\mathrm{cell}}^{-d/\beta}$. For every $t \in \mathcal{I}_{\lg}^{\mathrm{cell}}$, the definition of c_h gives $h = c_h r_{\mathrm{cell}} \le c_{\mathrm{ph}}\sigma_t$, so Theorem E.1 applies on every retained cell. The same choice gives $h \le \delta_i^{\mathrm{cell}}/(2\sqrt{d})$, so every cell whose closure intersects K_i^{cell} is contained in $\square_i$. Let $\boldsymbol{u}_\nu$ be the cell center, $\boldsymbol{y}_\nu = \boldsymbol{z}_i(\boldsymbol{u}_\nu)$, and $\mathsf{P}_\nu^{\tan}$ the orthogonal projection onto $T_{\boldsymbol{y}_\nu}\mathcal{M}$. Define

$$\mathsf{T}_\nu(\boldsymbol{x}, t) = m_t \boldsymbol{y}_\nu + \mathsf{P}_\nu^{\tan}(\boldsymbol{x} - m_t \boldsymbol{y}_\nu).$$

For $\boldsymbol{y} \in \mathcal{M}$, the identity

$$\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2^2 = \|\boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t)\|_2^2 + 2T_\nu(\boldsymbol{x},\boldsymbol{y},t) + m_t^2 D_\nu(\boldsymbol{x},\boldsymbol{y},t)$$

holds with

$$T_\nu = \langle \boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t), \mathsf{T}_\nu(\boldsymbol{x},t) - m_t\boldsymbol{y}\rangle, \qquad D_\nu = \|\mathsf{T}_\nu(\boldsymbol{x},t)/m_t - \boldsymbol{y}\|_2^2.$$

For $\boldsymbol{y} = \boldsymbol{z}_i(\boldsymbol{u})$, write the exact cell exponential as

$$E_\nu(\boldsymbol{x},\boldsymbol{u},t) := \exp\left(-\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\right) \tag{59}$$

and the tangent Gaussian envelope as

$$G_\nu(\boldsymbol{x},t) := \exp\left(-\frac{\|\boldsymbol{x} - \mathsf{T}_\nu(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right). \tag{60}$$

The distance identity gives the exact factorization

$$E_\nu(\boldsymbol{x},\boldsymbol{u},t) = G_\nu(\boldsymbol{x},t)\exp\left(-\frac{T_\nu(\boldsymbol{x},\boldsymbol{z}_i(\boldsymbol{u}),t)}{\sigma_t^2}\right)\exp\left(-\frac{m_t^2 D_\nu(\boldsymbol{x},\boldsymbol{z}_i(\boldsymbol{u}),t)}{2\sigma_t^2}\right). \tag{61}$$

Hence the exact chart-cell pieces are

$$\begin{aligned} \mathcal{Q}_{i,\nu}(\boldsymbol{x},t) &:= \sigma_t^{-d}\int_{Q_\nu} E_\nu(\boldsymbol{x},\boldsymbol{u},t) a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}, \\ \mathcal{U}_{i,\nu,k}(\boldsymbol{x},t) &:= \sigma_t^{-d}\int_{Q_\nu} \frac{m_t z_{i,k}(\boldsymbol{u}) - x_k}{\sigma_t} E_\nu(\boldsymbol{x},\boldsymbol{u},t) a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}. \end{aligned} \tag{62}$$

Step 1: fixed gates and tails. Use the center-phase proxy

$$\Phi^0_\nu(\boldsymbol{x},t) = \frac{\|\boldsymbol{x} - m_t \boldsymbol{y}_\nu\|_2^2}{2\sigma_t^2}.$$

Let $\theta_{\mathrm{lg}} : \mathbb{R} \to [0,1]$ be the fixed piecewise-affine ReLU ramp with $\theta_{\mathrm{lg}}(u) = 1$ for $u \le A_0$ and $\theta_{\mathrm{lg}}(u) = 0$ for $u \ge 2A_0$, where $A_0 \ge A_{\mathrm{ph}}$ is a large numerical constant. Define the exact cell gate

$$\Theta^{\mathrm{ex}}_\nu(\boldsymbol{x},t) := \theta_{\mathrm{lg}}\big(\Phi^0_\nu(\boldsymbol{x},t)/H^{\mathrm{dec}}_{\mathrm{lg}}\big)\,.$$

The network gate Θ_ν is obtained by approximating m_t, σ_t^{-2}, the squared norm, and the final ramp. Its arithmetic accuracy is chosen so that, outside the ramp transition band,

$$|\Theta_\nu - \Theta^{\mathrm{ex}}_\nu| \le c\varepsilon_{\mathrm{rel}}(N_{\mathrm{cell}}C_{\mathcal{M}})^{-1}(H^{\mathrm{dec}}_{\mathrm{lg}})^{-C}.$$

Inside the transition band, the exact proxy is between $A_0 H^{\mathrm{dec}}_{\mathrm{lg}}$ and $2A_0 H^{\mathrm{dec}}_{\mathrm{lg}}$, so the true cell contribution is controlled by the explicit tail estimate below; the gate implementation error is assigned to that tail budget. If the exact proxy is above $A_0 H^{\mathrm{dec}}_{\mathrm{lg}}$, Theorem E.1 and the Gaussian tail estimate give

$$\sigma_t^{-d}\int_{Q_\nu} e^{-\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|^2/(2\sigma_t^2)} a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \le C\sigma_t^{-d}h^d e^{-cA_0H^{\mathrm{dec}}_{\mathrm{lg}}}.$$

The centered factor in $\mathcal{U}_{i,k}$ is bounded on the non-negligible tail by $C(D \vee \log n)(H^{\mathrm{dec}}_{\mathrm{lg}})^{1/2}$. Summing over all cells and using $\sigma_t^{-d} \le \sigma_{\underline{t}}^{-d}$, $N_{\mathrm{cell}}h^d \le C$, and (55), with A_0 and C_{lg} large enough, gives the explicit exponential budget

$$C(D \vee \log n)(H^{\mathrm{dec}}_{\mathrm{lg}})^{1/2}\sigma_{\underline{t}}^{-d}\exp(-cA_0H^{\mathrm{dec}}_{\mathrm{lg}}) \le c\varepsilon_{\mathrm{rel}}q_{\mathrm{floor}}/C_{\mathcal{M}}.$$

Thus all gated-away and gate-transition contributions are bounded by $c\varepsilon_{\mathrm{rel}}q_{\mathrm{floor}}/C_{\mathcal{M}}$.

Step 2: fixed-order geometric replacement. On cells with $\Theta^{\mathrm{ex}}_\nu \ne 0$, Theorem E.1 yields, for every $\boldsymbol{u} \in Q_\nu$,

$$\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2/\sigma_t \le C(H^{\mathrm{dec}}_{\mathrm{lg}})^{1/2},$$

$$|T_\nu|/\sigma_t^2 \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}h(H^{\mathrm{dec}}_{\mathrm{lg}})^{1/2}, \qquad m_t^2D_\nu/(2\sigma_t^2) \le CH^{\mathrm{dec}}_{\mathrm{lg}}.$$

Replace $\boldsymbol{z}_i$ in T_ν, D_ν, and in the centered factor $(m_t z_{i,k}(\boldsymbol{u}) - x_k)/\sigma_t$, by its Taylor polynomial at $\boldsymbol{u}_\nu$ of fixed degree $q_{\mathrm{geom}} - 2$. The exponent and centered-factor perturbations are bounded by

$$C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}r_{\mathrm{cell}}^{q_{\mathrm{geom}}-2}(H^{\mathrm{dec}}_{\mathrm{lg}})^C,$$

which is at most $c\varepsilon_{\mathrm{rel}}$ by (56). The phase-smallness part of the same condition gives $|T_\nu|/\sigma_t^2 \le 1/2$, while $m_t^2D_\nu/(2\sigma_t^2) \ge 0$; hence the derivative of the non-Gaussian exponential factor is bounded by an absolute constant on the interpolation between the exact and polynomialized phases. The mean-value theorem for e^{-z} then gives a cell error bounded by

$$c\varepsilon_{\mathrm{rel}}\,\sigma_t^{-d}e^{-\|\boldsymbol{x}-\mathsf{T}_\nu(\boldsymbol{x},t)\|^2/(2\sigma_t^2)}\int_{Q_\nu} a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + c\varepsilon_{\mathrm{rel}}q_{\mathrm{floor}}(C_{\mathcal{M}}N_{\mathrm{cell}})^{-1}.$$

From this point through Step 4, T_ν, D_ν, and the centered factor are understood to be their polynomialized versions; the exact-to-polynomialized replacement error is the displayed error above.

Step 3: scalar Taylor expansions. For the polynomialized geometry, expand the first non-Gaussian exponential in (61),

$$\exp\big(-T_\nu/\sigma_t^2\big)\,,$$

to the fixed order $q_{\mathrm{geom}} - 3$. The small-argument hypothesis needed for this Taylor expansion is the first, explicit phase-smallness inequality in (56): using (50) and the choice of the cell side length,

$$|T_\nu|/\sigma_t^2 \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}r_{\mathrm{cell}}(H^{\mathrm{dec}}_{\mathrm{lg}})^{1/2} \le 1/2$$

after decreasing $c_{\mathrm{ph,cell}}$. The Taylor remainder is controlled by the second, high-order geometry inequality in (56); indeed, after enlarging the harmless exponent C,

$$\left|\frac{T_\nu}{\sigma_t^2}\right|^{q_{\mathrm{geom}}-2} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} r_{\mathrm{cell}}^{q_{\mathrm{geom}}-2}(H_{\mathrm{lg}}^{\mathrm{dec}})^C \le c\varepsilon_{\mathrm{rel}}.$$

Hence (54) gives the explicit Taylor remainder

$$\left|\exp\bigl(-T_\nu/\sigma_t^2\bigr) - \sum_{\ell=0}^{q_{\mathrm{geom}}-3} \frac{(-T_\nu/\sigma_t^2)^\ell}{\ell!}\right| \le c\varepsilon_{\mathrm{rel}}.$$

Multiplication by the explicit Gaussian envelope $G_\nu(\boldsymbol{x},t)$ gives, for $\mathcal{Q}$, a contribution bounded by

$$c\varepsilon_{\mathrm{rel}}\,\sigma_t^{-d} G_\nu(\boldsymbol{x},t)\int_{Q_\nu} a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u},$$

while for $\mathcal{U}$ the centered factor is bounded on gated cells by $C(D\vee\log n)(H_{\mathrm{lg}}^{\mathrm{dec}})^{1/2}$ and is absorbed into the $(H_{\mathrm{lg}}^{\mathrm{dec}})^C$ budgets above. Expand the second non-Gaussian exponential in (61),

$$\exp\left(-\frac{m_t^2 D_\nu}{2\sigma_t^2}\right),$$

to order $J=\lceil C_J H_{\mathrm{lg}}^{\mathrm{dec}}\rceil$. By Theorem E.2, choosing C_J large enough makes the remainder explicitly

$$\left|\exp\left(-\frac{m_t^2 D_\nu}{2\sigma_t^2}\right) - \sum_{j=0}^{J}\frac{[-m_t^2 D_\nu/(2\sigma_t^2)]^j}{j!}\right| \le e^{-c_{\exp}J}.$$

Since $J=\lceil C_J H_{\mathrm{lg}}^{\mathrm{dec}}\rceil$, $N_{\mathrm{cell}}h^d\le C$, and $\sigma_t^{-d}\le\sigma_{\underline{t}}^{-d}$, (55) gives, for C_J and C_{lg} large enough,

$$\sigma_t^{-d}h^d e^{-c_{\exp}J} \le c\varepsilon_{\mathrm{rel}} q_{\mathrm{floor}}(C_{\mathcal{M}}N_{\mathrm{cell}})^{-1}(H_{\mathrm{lg}}^{\mathrm{dec}})^{-C}.$$

This is the absolute cell budget used in the final summation.

Step 4: stored moments and neural implementation. After the two scalar expansions, each cell contribution is a finite sum of atoms consisting of powers of m_t, powers of σ_t^{-1}, the tangent coordinates $\mathsf{P}_\nu^{\tan}(\boldsymbol{x}-m_t\boldsymbol{y}_\nu)$, fixed-degree normal residual terms, and the explicit Gaussian envelope

$$G_\nu(\boldsymbol{x},t)=\exp\left(-\frac{\|\boldsymbol{x}-\mathsf{T}_\nu(\boldsymbol{x},t)\|^2}{2\sigma_t^2}\right).$$

The coefficients are fixed moments of the form

$$\int_{Q_\nu}(\boldsymbol{u}-\boldsymbol{u}_\nu)^\alpha a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$$

for $\mathcal{Q}_i$, and for $\mathcal{U}_{i,k}$ the same moments together with

$$\int_{Q_\nu}(\boldsymbol{u}-\boldsymbol{u}_\nu)^\alpha z_{i,\nu,k}^{(q)}(\boldsymbol{u})a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u},$$

where $z_{i,\nu,k}^{(q)}$ is the fixed Taylor polynomial. The term $-x_k/\sigma_t$ multiplies the $\mathcal{Q}$-moment family. Hence the centered numerator is expanded directly and no amplitude $a_i^{(\mathcal{U}_k)}$ is introduced. The magnitudes of these stored moments are bounded by fixed powers of B_0 and the geometry constants through Theorem D.5. This is why (55) and (58) display the density-amplitude factor explicitly.

There are at most $(H_{\mathrm{lg}}^{\mathrm{dec}})^C$ atoms per cell; the dependence on the fixed order $q_{\mathrm{geom}}-3$ is absorbed into the fixed polynomial powers of D, $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$, and $H_{\mathrm{lg}}^{\mathrm{dec}}$. More explicitly, for one cell atom:

- the affine projection $\mathsf{P}_\nu^{\tan}(\boldsymbol{x} - m_t \boldsymbol{y}_\nu)$, squared norm, and coordinate summations contribute a multiplicative factor at most D^4;
- the fixed Taylor replacement of order at most q_{geom} contributes a multiplicative factor at most $D^{4q_{\mathrm{geom}}}$, while all intrinsic monomial counts depend only on d and q_{geom};
- the input ranges on A_t, the centered numerator factor, and the scalar arithmetic subnetworks contribute a multiplicative factor at most $(D \vee \log n)^{4(q_{\mathrm{geom}}+2)}$, with the remaining dependence absorbed into $(H_{\mathrm{lg}}^{\mathrm{dec}})^C$.

Thus the whole cellwise implementation is bounded by

$$D^{4q_{\mathrm{geom}}+4}(D \vee \log n)^{4(q_{\mathrm{geom}}+2)}(H_{\mathrm{lg}}^{\mathrm{dec}})^C$$

per cell after harmless enlargement of constants, which is covered by the exponent choice in the proposition statement. Reciprocal, schedule-power, scalar exponential, monomial, and product networks are assigned relative accuracy $c\varepsilon_{\mathrm{rel}}(H_{\mathrm{lg}}^{\mathrm{dec}})^{-C}$ and absolute accuracy $c\varepsilon_{\mathrm{rel}} q_{\mathrm{floor}}(C_{\mathcal{M}} N_{\mathrm{cell}})^{-1}(H_{\mathrm{lg}}^{\mathrm{dec}})^{-C}$. Summing the atom errors, the gate tails, the geometric replacement error, and the Taylor remainders gives (57). Parallelizing over N_{cell} cells gives (58). □

Lemma E.4 (Large-noise denominator lower bound)**.** *Assume Theorems 2.2, 2.4 and 2.5. There are constants $c_{\mathrm{lb}}, C_{\mathrm{lb}} > 0$, depending only on the displayed finite geometry controls, such that*

$$c_{\mathrm{lb}}^{-1} \vee C_{\mathrm{lb}} \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}$$

after enlarging the finite-order geometry constant from Theorem B.5. For every $t \in [\underline{t}, \overline{t}]$ and every

$$\boldsymbol{x} \in A_t = \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le C_\varrho \sigma_t \sqrt{D \vee \log n}\},$$

$$\mathcal{Q}(\boldsymbol{x}, t) \ge q_{\mathrm{lb}} := c_{\mathrm{lb}} p_{\min} \exp\{-C_{\mathrm{lb}} C_\varrho^2 (D \vee \log n)\}. \tag{63}$$

Moreover

$$\frac{\|\mathcal{U}(\boldsymbol{x}, t)\|_2}{\mathcal{Q}(\boldsymbol{x}, t)} \le R_{\mathrm{lb}} := 1 + \Big[2\log\big(e\sigma_{\underline{t}}^{-d} q_{\mathrm{lb}}^{-1}\big) + C_{\mathrm{lb}}(D \vee \log n)\Big]^{1/2}. \tag{64}$$

Proof. Step 1: lower-bound the local mass in one intrinsic ball. Fix $(\boldsymbol{x}, t)$ in the displayed tube and choose $\boldsymbol{y}_0 \in \mathcal{M}$ with $\|\boldsymbol{x} - m_t \boldsymbol{y}_0\|_2 = \mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M})$. Let $a_{\mathrm{lb}} \in (0, 1]$ be smaller than the injectivity and coordinate-buffer radii of the projection atlas. By the finite-order geometry convention and Theorem F.3, it can be chosen so that $a_{\mathrm{lb}}^{-1} \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}$, and the small-ball volume bound

$$\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(\boldsymbol{y}, a_{\mathrm{lb}} r)) \ge C^{-1}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C} r^d, \qquad 0 < r \le 1,$$

holds uniformly in $\boldsymbol{y} \in \mathcal{M}$. Since $0 < \sigma_t \le 1$, the ball $B_{\mathcal{M}}(\boldsymbol{y}_0, a_{\mathrm{lb}}\sigma_t)$ remains in a single normal coordinate patch. For $\boldsymbol{y} \in B_{\mathcal{M}}(\boldsymbol{y}_0, a_{\mathrm{lb}}\sigma_t)$,

$$\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2 \le C\sigma_t \sqrt{D \vee \log n},$$

after decreasing a_{lb}, because $m_t \le 1$ and $\boldsymbol{x} \in A_t$. The density lower bound and the uniform atlas/Jacobian bounds therefore give

$$\mathcal{Q}(\boldsymbol{x}, t) = \sigma_t^{-d} \int_{\mathcal{M}} e^{-\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2/(2\sigma_t^2)} p_0(\boldsymbol{y}) \mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y}) \ge c p_{\min} e^{-CC_\varrho^2(D \vee \log n)} \sigma_t^{-d} \mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(\boldsymbol{y}_0, a_{\mathrm{lb}}\sigma_t)).$$

The displayed small-ball volume lower bound gives $\mathrm{vol}_{\mathcal{M}}(B_{\mathcal{M}}(\boldsymbol{y}_0, a_{\mathrm{lb}}\sigma_t)) \ge C^{-1}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C}\sigma_t^d$. This proves (63) with $c_{\mathrm{lb}}^{-1} \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}$.

Step 2: bound the centered ratio. The identity $s^*(\boldsymbol{x}, t) = \sigma_t^{-1}\mathcal{U}(\boldsymbol{x}, t)/\mathcal{Q}(\boldsymbol{x}, t)$ and

$$p_t(\boldsymbol{x}) = (2\pi\sigma_t^2)^{-D/2}\sigma_t^d \mathcal{Q}(\boldsymbol{x}, t)$$

give the required bound on the score norm as follows. Let $\boldsymbol{Y} \sim P_0$, put $\boldsymbol{z}(\boldsymbol{y}) = (\boldsymbol{x} - m_t\boldsymbol{y})/\sigma_t$, and define the posterior measure

$$\mathrm{d}\nu_{\boldsymbol{x},t}(\boldsymbol{y}) = \frac{\exp(-\|\boldsymbol{z}(\boldsymbol{y})\|_2^2/2)}{\int_{\mathcal{M}} \exp(-\|\boldsymbol{z}(\boldsymbol{y}')\|_2^2/2)\mathrm{d}P_0(\boldsymbol{y}')}\mathrm{d}P_0(\boldsymbol{y}).$$

If $Z_{\boldsymbol{x},t}$ denotes the denominator in this display, then

$$\sigma_t s^*(\boldsymbol{x},t) = -\int_{\mathcal{M}} \boldsymbol{z}(\boldsymbol{y})\mathrm{d}\nu_{\boldsymbol{x},t}(\boldsymbol{y}).$$

Moreover

$$\mathrm{KL}(\nu_{\boldsymbol{x},t}\|P_0) = -\frac{1}{2}\int_{\mathcal{M}} \|\boldsymbol{z}(\boldsymbol{y})\|_2^2\mathrm{d}\nu_{\boldsymbol{x},t}(\boldsymbol{y}) - \log Z_{\boldsymbol{x},t} \geq 0,$$

and hence, by Jensen's inequality,

$$\|\sigma_t s^*(\boldsymbol{x},t)\|_2^2 \leq \int_{\mathcal{M}} \|\boldsymbol{z}(\boldsymbol{y})\|_2^2\mathrm{d}\nu_{\boldsymbol{x},t}(\boldsymbol{y}) \leq 2\log(Z_{\boldsymbol{x},t}^{-1}).$$

Since $Z_{\boldsymbol{x},t} = (2\pi\sigma_t^2)^{D/2}p_t(\boldsymbol{x}) = \sigma_t^d\mathcal{Q}(\boldsymbol{x},t)$, we obtain

$$\frac{\|\mathcal{U}(\boldsymbol{x},t)\|_2^2}{\mathcal{Q}(\boldsymbol{x},t)^2} = \sigma_t^2\|s^*(\boldsymbol{x},t)\|_2^2 \leq 2\log\left(\frac{(2\pi\sigma_t^2)^{-D/2}}{p_t(\boldsymbol{x})}\right) + 1 = 2\log\big(\sigma_t^{-d}\mathcal{Q}(\boldsymbol{x},t)^{-1}\big) + 1.$$

Using $\sigma_t \geq \sigma_{\underline{t}}$, $\mathcal{Q}(\boldsymbol{x},t) \geq q_{\mathrm{lb}}$, and enlarging C_{lb} gives (64). □

Theorem E.5 (Global tangent-cell large-noise score approximation)**.** *Assume Theorems 2.2, 2.4 and 2.5. Fix* $0 < \varepsilon_{\mathrm{cell}} \leq 1/2$ *and* $0 < \varepsilon_{\mathrm{sc}} \leq 1$. *Let* q_{geom} *be as in (1), and let* $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$ *be the finite geometry constant from Theorem B.5. Let* $r_{\mathrm{cell}} = \varepsilon_{\mathrm{cell}}^{1/\beta}$ *and*

$$\mathcal{I}_{\mathrm{lg}}^{\mathrm{cell}} := \{t \in [\underline{t},\overline{t}] : \sigma_t \geq c_{\mathrm{lg}}r_{\mathrm{cell}}\}.$$

Let q_{lb} *and* R_{lb} *be defined in (63) and (64), and set*

$$C_{\mathrm{den}} := q_{\mathrm{lb}}/8, \qquad \varepsilon_{\mathrm{rel}} := c_{\mathrm{glob}}\frac{\varepsilon_{\mathrm{sc}}}{1+R_{\mathrm{lb}}},$$

$$H_{\mathrm{lg}}^{\mathrm{lb}} := 1 + \log\Big(C_{\mathrm{lg}}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}B_0^{C_{\mathrm{lg}}}(D \vee \log n)^{C_{\mathrm{lg}}}\sigma_{\underline{t}}^{-d}C_{\mathrm{den}}^{-1}(1+R_{\mathrm{lb}})\varepsilon_{\mathrm{sc}}^{-1}\Big).$$

After decreasing c_{lb}, *we assume* $q_{\mathrm{lb}} \leq 1$, *so* $C_{\mathrm{den}} \in (0,1]$. *Choose the numerical constant* $c_{\mathrm{glob}} \leq 1/32$, *so that* $\varepsilon_{\mathrm{rel}} \leq 1/16$ *and the separated tangent-cell proposition is applicable. Assume the large-noise cell scale obeys the phase-smallness and fixed-order geometry conditions, with the same sufficiently small numerical constant* $c_{\mathrm{ph,cell}}$*:*

$$\Gamma_{\mathcal{M},q_{\mathrm{geom}}}r_{\mathrm{cell}}(H_{\mathrm{lg}}^{\mathrm{lb}})^{1/2} \leq c_{\mathrm{ph,cell}}, \qquad \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}r_{\mathrm{cell}}^{q_{\mathrm{geom}}-2}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C} \leq c\varepsilon_{\mathrm{rel}}. \tag{65}$$

Then there exists a ReLU network $\widetilde{s}_{\mathrm{lb}}^{\mathrm{lg}}$ *such that, uniformly over* $t \in \mathcal{I}_{\mathrm{lg}}^{\mathrm{cell}}$ *and* $\boldsymbol{x} \in A_t$,

$$\sigma_t\|\widetilde{s}_{\mathrm{lb}}^{\mathrm{lg}}(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_\infty \leq \varepsilon_{\mathrm{sc}}. \tag{66}$$

Its depth L, *maximum width* $\|\boldsymbol{W}\|_\infty$, *sparsity* S, *and log-weight bound* $\log B$ *satisfy*

$$\begin{aligned} L &\leq C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}B_0^{C}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C}, \\ \log B &\leq C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}B_0^{C}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C}, \\ \|\boldsymbol{W}\|_\infty &\leq CD^{4(q_{\mathrm{geom}}+2)}(D \vee \log n)^{4(q_{\mathrm{geom}}+2)}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}B_0^{C}\varepsilon_{\mathrm{cell}}^{-d/\beta}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C}, \\ S &\leq CD^{4(q_{\mathrm{geom}}+2)}(D \vee \log n)^{4(q_{\mathrm{geom}}+2)}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}B_0^{C}\varepsilon_{\mathrm{cell}}^{-d/\beta}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C}. \end{aligned} \tag{67}$$

In particular, the width and sparsity contain the intrinsic tangent-cell count $\varepsilon_{\mathrm{cell}}^{-d/\beta}$. *In the slabwise application below, the fixed-order geometry condition forces an accuracy-dependent cap on* $\varepsilon_{\mathrm{cell}}$*:* $\varepsilon_{\mathrm{cell}}^{-d/\beta} = r_k^{-d} \leq C(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d})$, *and the* $r_{\star,n}^{-d}$ *term is carried explicitly through the large-noise oracle and* W_1 *bounds.*

Proof. Step 1: approximate the unrefactored chart integrals. Apply Theorem E.3 with $q_{\text{floor}} = C_{\text{den}}$ and ε_{rel} as above. The definition of $H_{\text{lg}}^{\text{lb}}$ dominates the logarithmic scale $H_{\text{lg}}^{\text{dec}}$, and (65) implies (56). After summing over the fixed number of charts,

$$|\widehat{\mathcal{Q}}(\boldsymbol{x},t) - \mathcal{Q}(\boldsymbol{x},t)| \vee \max_k |\widehat{\mathcal{U}}_k(\boldsymbol{x},t) - \mathcal{U}_k(\boldsymbol{x},t)| \le C\varepsilon_{\text{rel}}\big(\mathcal{Q}(\boldsymbol{x},t) + C_{\text{den}}\big).$$

By (63), $C_{\text{den}} \le \mathcal{Q}/8$, so after decreasing c_{glob},

$$|\widehat{\mathcal{Q}} - \mathcal{Q}| \vee \|\widehat{\mathcal{U}} - \mathcal{U}\|_\infty \le c\varepsilon_{\text{rel}}\mathcal{Q}. \tag{68}$$

Step 2: pass from integral approximation to ratio approximation. The map $q \mapsto q \vee C_{\text{den}}$ is ReLU-realizable, and on the present domain $\mathcal{Q} \ge 8C_{\text{den}}$. Thus $\widehat{\mathcal{Q}} \vee C_{\text{den}} \ge \mathcal{Q}/2$. Using (64) and (68),

$$\left\| \frac{\widehat{\mathcal{U}}}{\widehat{\mathcal{Q}} \vee C_{\text{den}}} - \frac{\mathcal{U}}{\mathcal{Q}} \right\|_\infty \le C(1 + R_{\text{lb}})\varepsilon_{\text{rel}} \le \varepsilon_{\text{sc}}/4.$$

The reciprocal network is applied only on the interval $[C_{\text{den}}, 2M_0]$, where M_0 is the uniform chart-integral bound from Theorem D.6; the preceding approximation and $\mathcal{Q} \le M_0$ put $\widehat{\mathcal{Q}} \vee C_{\text{den}}$ in this interval after decreasing c_{glob}. Its accuracy is chosen

$$c\varepsilon_{\text{sc}}(H_{\text{lg}}^{\text{lb}})^{-C}$$

and its depth, sparsity, and log-weight are absorbed into the displayed powers of $H_{\text{lg}}^{\text{lb}}$. The final multiplication by σ_t^{-1} is implemented with the same arithmetic accuracy. This proves (66).

Step 3: record size. The chart networks have the size in (58) with $H_{\text{lg}}^{\text{dec}} \le H_{\text{lg}}^{\text{lb}}$. The reciprocal, products, clipping, and the final schedule multiplier add only fixed powers of $H_{\text{lg}}^{\text{lb}}$ and do not introduce any additional intrinsic net. Therefore the width and sparsity scale as the cell count

$$N_{\text{cell}} \le C\varepsilon_{\text{cell}}^{-d/\beta},$$

times the explicit ambient and logarithmic polynomial factor, giving (67). □

F. Score Approximation in the Small-Noise Regime

With a positive density lower bound, the small-noise denominator can be controlled uniformly after de-Gaussianization. This route uses the projection-centered Laplace expansion and the neural approximation of the nearest-point projection. This section packages the whole small-noise approximation pipeline. It has four submodules: active-chart geometry in Sec. F.1, the dimension-explicit projection-network primitive in Sec. F.2, the projection-centered Laplace expansion in Sec. F.3, and the final network assembly in Sec. F.4.

F.1. Proof architecture and active-chart geometry

Small-noise proof architecture. The target result is Theorem F.29. The proof first localizes to active charts and discards inactive Gaussian tails. It then uses reach to make the nearest projection single-valued, expands the de-Gaussianized chart integrals in projection-centered tangent coordinates, implements the projection coordinate by finite-anchor Gauss–Newton networks, and finally builds the stable ratio $\bar{\mathcal{U}}/\bar{\mathcal{Q}}$. The main denominator step is Theorem F.27; the main network-assembly step is Theorem F.28, while Theorem F.25 records the local chart-level Laplace network on its certified active set.

F.1.1. Active-chart localization and inactive-tail bounds

The first task is to decide when a chart can matter. The compact set $\mathcal{S}_i$ was fixed in (19); it is the part of the manifold on which the i-th partition function is active. If no point of $m_t\mathcal{S}_i$ lies near $\boldsymbol{x}$, the Gaussian factor is uniformly small on the whole chart support, and the corresponding chart integral can be discarded at the target accuracy.

Let

$$C_{\mathrm{tail}}^{\max} := 1 \vee 2B_p^{\mathrm{proj}} B^{\mathrm{Jac}} \max_{1\le i\le C_{\mathcal{M}}} \sup_{0<\sigma\le 1,\ \boldsymbol{x}\in\mathbb{R}^D} \sigma^{-d} \int_{\phi_i(U_i)} \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{8\sigma^2}\right) \mathrm{d}\boldsymbol{u}. \tag{69}$$

This constant is finite by the chart bi-Lipschitz bounds and the annular estimate used in the proof of Theorem F.1; importantly, it is independent of $\sigma_{\underline{t}}$. Choose the default localization constant $c_\star \ge 1$ large enough for the finite collection of inactive Gaussian-tail estimates used below, and also so that $c_\star \ge 2C_\varrho$. More precisely, whenever a bound below has the form

$$C_{\mathrm{tail}}^{\max} C\big(1+\log(e/\varepsilon)\big)^M \varepsilon^{c_\star^2/C}, \qquad 0<\varepsilon\le 1/2,$$

with one of the finitely many constants C, M arising from the fixed atlas, amplitudes, and retained Laplace monomials, our choice of $c_\star$ makes this quantity at most $\varepsilon/12$. The same choice is also large enough that, for every $H \ge 1$, the H-localized de-Gaussianized inactive-tail bounds below have the form

$$C_{\mathrm{tail}}^{\max} C(1+H)^M e^{-cc_\star^2 H} \le e^{-H}$$

for the finite collection of constants C, M, c used there. Such a choice is possible because $(1+\log(e/\varepsilon))^M\varepsilon^\gamma \le C_{\gamma,M}\varepsilon$ on $(0,1/2]$ for all sufficiently large γ, and because, for each fixed triple C, M, c,

$$\sup_{H\ge 1} C_{\mathrm{tail}}^{\max} C(1+H)^M \exp\{-(cc_\star^2-1)H\} \to 0 \qquad \text{as } c_\star\to\infty.$$

The collection of such triples is finite, so one $c_\star$ satisfies all of them. No lower-noise prefactor appears in the annular tail estimate. For any auxiliary radius multiplier $c_a > 0$ and any $0<\varepsilon<1$, define the parameterized active region

$$\mathcal{K}_i^{(c_a)}(\varepsilon) := \big\{(\boldsymbol{x},t) : t\in[\underline{t},\overline{t}],\ \boldsymbol{x}\in A_t, \exists \boldsymbol{y}\in\mathcal{S}_i \text{ s.t. } \|\boldsymbol{x}-m_t\boldsymbol{y}\|_2 \le c_a\sigma_t\sqrt{\log\varepsilon^{-1}}\big\}. \tag{70}$$

The default active set is

$$\mathcal{K}_i(\varepsilon) := \mathcal{K}_i^{(c_\star)}(\varepsilon). \tag{71}$$

Thus $c_\star$ is used for the default active set throughout; c_a only denotes a temporary enlargement factor in statements involving $\mathcal{K}_i^{(c_a)}(\varepsilon)$.

Lemma F.1 (Inactive-chart Gaussian tail bounds for $\mathcal{P}_i$ and $\mathcal{Q}_i$)**.** *Fix any chart $i\in\{1,\dots,C_{\mathcal{M}}\}$ and every $0<\varepsilon<1$. Let $c_a>0$, $t\in[\underline{t},\overline{t}]$, and $\boldsymbol{x}\in A_t$. If $(\boldsymbol{x},t)\notin\mathcal{K}_i^{(c_a)}(\varepsilon)$, then for every $\boldsymbol{u}\in\phi_i(U_i)$ with $\bar{\rho}_i(\boldsymbol{u})\neq 0$,*

$$\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2 > c_a\sigma_t\sqrt{\log\varepsilon^{-1}}. \tag{72}$$

Consequently, for a constant C_{tail} depending only on the fixed atlas, amplitude bounds, d, and the schedule bound $\underline{m}$, but not on $\sigma_{\underline{t}}$,

$$|\mathcal{Q}_i(\boldsymbol{x},t)| + \max_{1\le k\le D} |\mathcal{P}_{i,k}(\boldsymbol{x},t)| \le C_{\text{tail}}\big(1 + c_a\sqrt{\log(e/\varepsilon)}\big)^d \varepsilon^{c_a^2/8}. \tag{73}$$

Proof of Theorem F.1. If $\bar{\rho}_i(\boldsymbol{u}) \neq 0$, then $\boldsymbol{z}_i(\boldsymbol{u}) \in \mathcal{S}_i$. Failure of membership in $\mathcal{K}_i^{(c_a)}(\varepsilon)$ means that no $\boldsymbol{y} \in \mathcal{S}_i$ satisfies the active distance condition. This gives (72).

Put $R := c_a\sqrt{\log \varepsilon^{-1}}$. The chart maps are uniformly bi-Lipschitz on their compact supports and $m_t \ge \underline{m} > 0$. Hence there is a constant C_i, independent of t, σ_t, and $\boldsymbol{x}$, such that for every $r > 0$,

$$\mathrm{vol}_d\{\boldsymbol{u} \in \phi_i(U_i) : \|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2 \le r\} \le C_i(1 + r/\sigma_t)^d \sigma_t^d. \tag{74}$$

For $r \le 1$, this follows from the local bi-Lipschitz bounds after covering $m_t\mathcal{S}_i$ by a bounded number of coordinate balls; for $r > 1$, the right side dominates the fixed chart volume after increasing C_i.

Decompose the support into annuli

$$A_\ell := \{\boldsymbol{u} :\ R + \ell < \|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2/\sigma_t \le R + \ell + 1\}, \qquad \ell = 0, 1, 2, \dots.$$

By (74), $\mathrm{vol}_d(A_\ell) \le C_i\sigma_t^d(1 + R + \ell)^d$. Therefore,

$$\begin{aligned}\sigma_t^{-d}\int_{\phi_i(U_i)} \exp\left(-\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\right)\mathbb{1}\{\bar{\rho}_i(\boldsymbol{u}) \neq 0\}\mathrm{d}\boldsymbol{u} &\le C_i\sum_{\ell\ge 0}(1 + R + \ell)^d e^{-(R+\ell)^2/2} \\ &\le C_i'(1 + R)^d e^{-R^2/8}.\end{aligned}$$

The last inequality is the standard Gaussian-annulus summation bound; it follows, for instance, by comparing the sum with $e^{-R^2/8}\sum_{\ell\ge 0}(1 + \ell)^d e^{-\ell^2/8}$. Since $|a_{i,k}^{(\mathcal{P})}| \le |a_i^{(\mathcal{Q})}| \le B_p^{\text{proj}} B^{\text{Jac}}$ by (31) and $\|\boldsymbol{z}_i\|_\infty \le 1$, the same estimate applies to $\mathcal{Q}_i$ and to every coordinate of $\mathcal{P}_i$. Because $e^{-R^2/8} = \varepsilon^{c_a^2/8}$ and $1 + R \le 1 + c_a\sqrt{\log(e/\varepsilon)}$, we obtain (73). □

By the fixed choice of $c_\star$, the bound in (73) with $c_a = c_\star$, and also with any larger fixed multiplier, is below the allocated $\mathcal{O}(\varepsilon)$ tail budget after multiplication by the finitely many logarithmic factors used later. Consequently, the projection–Laplace approximation only has to be constructed on the active region $\mathcal{K}_i(\varepsilon)$. It is used only during the *small-noise regime* $t \in \mathcal{I}_{\text{sm}}(\varepsilon)$, where $\sigma_t \lesssim \varepsilon^{1/\beta}$.

F.1.2. Projection geometry on active small-noise charts

The next lemmas turn the active-distance condition into usable local coordinates. The buffered boxes keep all projected active points away from chart boundaries, while the reach-tube estimates ensure that the nearest-point projection and the projection-centered objective have stable curvature. These facts are the analytic foundation for both the Laplace expansion and the neural Gauss–Newton projection network.

Buffered coordinate set. The compact support $\mathcal{S}_i$ and buffered boxes $K_i^\circ \Subset \square_i$ were fixed in (19) to (21). Since $\mathcal{S}_i$ is compact and $\phi_i(\mathcal{S}_i) \subset \mathrm{int}(K_i^\circ)$, we can define the strictly positive geometric buffer

$$B_i^{\text{buff}} := 1 \wedge \mathrm{dist}_{\mathbb{R}^D}\Big(\mathcal{S}_i,\ \mathcal{M} \setminus \phi_i^{-1}\big(\mathrm{int}(K_i^\circ)\big)\Big) > 0. \tag{75}$$

with the convention $\mathrm{dist}(A, \emptyset) = +\infty$.

Auxiliary compact chart domains. Set

$$d_i^\square := \mathrm{dist}(K_i^\circ, \partial\square_i) > 0. \tag{76}$$

On this single compact set $\square_i$, define

$$L_i^{(+)} := \sup_{\boldsymbol{u}\in\square_i} \|\boldsymbol{J}_i(\boldsymbol{u})\|_{\mathrm{op}}, \qquad L_i^{(-)} := \Big(\sup_{\boldsymbol{y}\in\phi_i^{-1}(\square_i)} \|\mathrm{D}\phi_i(\boldsymbol{y})\|_{\mathrm{op}}\Big)^{-1}, \tag{77}$$

$$\lambda_i^{(-)} := \inf_{\boldsymbol{u}\in\square_i} \lambda_{\min}\big(\boldsymbol{G}_i(\boldsymbol{u})\big), \qquad \lambda_i^{(+)} := \sup_{\boldsymbol{u}\in\square_i} \lambda_{\max}\big(\boldsymbol{G}_i(\boldsymbol{u})\big), \tag{78}$$

$$\mathsf{M}_i^{(2)} := 1 \vee \sup_{\boldsymbol{u}\in\square_i} \sum_{\ell=1}^{D} \|\nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u})\|_{\mathrm{op}}. \tag{79}$$

Since $\boldsymbol{z}_i = \phi_i^{-1}$ is smooth on $\phi_i(U_i)$ and $\square_i \Subset \phi_i(U_i)$, all these constants are finite. Moreover, because $\boldsymbol{z}_i$ is an immersion, $\boldsymbol{G}_i(\boldsymbol{u}) = \boldsymbol{J}_i(\boldsymbol{u})^\top \boldsymbol{J}_i(\boldsymbol{u})$ is symmetric positive definite for every $\boldsymbol{u} \in \square_i$. Hence

$$0 < L_i^{(-)} \le L_i^{(+)} < \infty, \qquad 0 < \lambda_i^{(-)} \le \lambda_i^{(+)} < \infty.$$

Define

$$\boldsymbol{A}_i(\boldsymbol{u}) := \boldsymbol{G}_i(\boldsymbol{u})^{-1}\boldsymbol{J}_i(\boldsymbol{u})^\top, \qquad \boldsymbol{u} \in \square_i. \tag{80}$$

Again, because $\boldsymbol{z}_i$ is smooth on $\square_i$ and $\boldsymbol{G}_i(\boldsymbol{u})$ is uniformly invertible there, $\boldsymbol{A}_i$ is smooth on $\square_i$. Hence

$$L_{A,i} := 1 \vee \sup_{\boldsymbol{u}\in\square_i} \|\mathrm{D}\boldsymbol{A}_i(\boldsymbol{u})\|_{\mathrm{op}} < \infty. \tag{81}$$

Lemma F.2 (Uniform chart geometry on compact coordinate patches)**.** *For each chart index i, for all $\boldsymbol{u}, \boldsymbol{v} \in \square_i$,*

$$L_i^{(-)}\|\boldsymbol{u}-\boldsymbol{v}\|_2 \le \|\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_i(\boldsymbol{v})\|_2 \le L_i^{(+)}\|\boldsymbol{u}-\boldsymbol{v}\|_2, \tag{82}$$

and for all $\boldsymbol{u} \in \square_i$,

$$\lambda_i^{(-)}\boldsymbol{I}_d \preceq \boldsymbol{G}_i(\boldsymbol{u}) \preceq \lambda_i^{(+)}\boldsymbol{I}_d. \tag{83}$$

Proof. For the upper Lipschitz bound, by the fundamental theorem of calculus,

$$\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_i(\boldsymbol{v}) = \int_0^1 \boldsymbol{J}_i\big(\boldsymbol{v} + s(\boldsymbol{u}-\boldsymbol{v})\big)(\boldsymbol{u}-\boldsymbol{v})\mathrm{d}s.$$

Since $\square_i$ is convex, $\boldsymbol{v} + \theta(\boldsymbol{u}-\boldsymbol{v}) \in \square_i$ for every $\theta \in [0,1]$. Therefore,

$$\begin{aligned}\|\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_i(\boldsymbol{v})\|_2 &\le \int_0^1 \|\boldsymbol{J}_i(\boldsymbol{v}+\theta(\boldsymbol{u}-\boldsymbol{v}))\|_{\mathrm{op}}\|\boldsymbol{u}-\boldsymbol{v}\|_2\mathrm{d}\theta \\ &\le L_i^{(+)}\|\boldsymbol{u}-\boldsymbol{v}\|_2.\end{aligned}$$

For the lower Lipschitz bound, apply the same argument to the inverse map ϕ_i on the compact set $\phi_i^{-1}(\square_i) \subset U_i$:

$$\begin{aligned}\|\boldsymbol{u}-\boldsymbol{v}\|_2 &= \|\phi_i(\boldsymbol{z}_i(\boldsymbol{u})) - \phi_i(\boldsymbol{z}_i(\boldsymbol{v}))\|_2 \\ &\le \sup_{\boldsymbol{y}\in\phi_i^{-1}(\square_i)} \|\mathrm{D}\phi_i(\boldsymbol{y})\|_{\mathrm{op}}\|\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_i(\boldsymbol{v})\|_2.\end{aligned}$$

By the definition of $L_i^{(-)}$, this gives

$$L_i^{(-)}\|\boldsymbol{u}-\boldsymbol{v}\|_2 \le \|\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_i(\boldsymbol{v})\|_2.$$

Finally, $\boldsymbol{G}_i(\boldsymbol{u})$ is continuous and symmetric positive definite on the compact set $\square_i$. Hence its smallest and largest eigenvalues attain finite extrema on $\square_i$, and the definitions of $\lambda_i^{(-)}$ and $\lambda_i^{(+)}$ give (83). $\square$

Lemma F.3 (Finite-order geometry from compact smoothness)**.** *Assume Theorem 2.2. Fix an integer $q \geq 1$. For the finite projection atlas $\{(U_i, \phi_i)\}_{i=1}^{C_{\mathcal{M}}}$, the partition of unity $\{\rho_i\}$, the compact coordinate boxes $\square_i$, and the compact tubular domains used in the construction, there exists a finite constant $\Gamma_{\mathcal{M},q} \geq e$ such that, uniformly over the chart index i,*

$$\max_{|\boldsymbol{\alpha}| \leq q+4} \|\partial^{\boldsymbol{\alpha}} \boldsymbol{z}_i\|_{L^\infty(\square_i)} \leq \Gamma_{\mathcal{M},q}, \tag{84}$$

$$\max_{|\boldsymbol{\alpha}| \leq q+4} \|\partial^{\boldsymbol{\alpha}} \bar{\rho}_i\|_{L^\infty(\square_i)} + \max_{|\boldsymbol{\alpha}| \leq q+4} \|\partial^{\boldsymbol{\alpha}} \sqrt{\det \boldsymbol{G}_i}\|_{L^\infty(\square_i)} \leq \Gamma_{\mathcal{M},q}. \tag{85}$$

After enlarging $\Gamma_{\mathcal{M},q}$, the same bound holds for transition maps between overlapping projection charts, for $\boldsymbol{A}_i = \boldsymbol{G}_i^{-1} \boldsymbol{J}_i^\top$, for the nearest-projection coordinate maps

$$\boldsymbol{x} \mapsto \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}))$$

on the compact reach tubes where they are used, and for the corresponding projection residual maps. Schedule-dependent maps such as $(\boldsymbol{x}, t) \mapsto \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t))$ are controlled by composing these geometric maps with the separately approximated schedule m_t^{-1}; their constants therefore depend on $\Gamma_{\mathcal{M},q}$ and the fixed lower bound $\underline{m}$, not on any new geometry assumption. The constant can also be chosen to dominate the finite atlas cardinality, fixed coordinate buffers, tube constants, chart-volume constants, and their reciprocals. With $\mathcal{K}_{\mathcal{M},q}$ denoting the high-order atlas envelope from Theorem B.5, it may be chosen so that

$$\Gamma_{\mathcal{M},q} \leq C_{d,q} (1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}(q)} (1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}(q)} \mathcal{K}_{\mathcal{M},q}^{\mathfrak{a}_{\mathrm{geom}}(q)},$$

where $C_{d,q}$ and $\mathfrak{a}_{\mathrm{geom}}(q)$ depend only on d and q.

Proof. For each $\boldsymbol{y} \in \mathcal{M}$, let $\boldsymbol{V}_{\boldsymbol{y}}$ be an orthonormal frame for $T_{\boldsymbol{y}}\mathcal{M}$. The map $\phi_{\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{V}_{\boldsymbol{y}}^\top (\boldsymbol{z} - \boldsymbol{y})$, restricted to $\mathcal{M}$, has differential equal to the identity at $\boldsymbol{y}$. The inverse function theorem therefore gives a C^∞ projection chart near $\boldsymbol{y}$. Compactness of $\mathcal{M}$ gives a finite subcover, and the construction in Sec. D.3.1 shrinks these charts to compact coordinate boxes $\square_i \Subset \phi_i(U_i)$.

Each inverse chart $\boldsymbol{z}_i = \phi_i^{-1}$ is C^∞. Hence every partial derivative $\partial^{\boldsymbol{\alpha}} \boldsymbol{z}_i$, $|\boldsymbol{\alpha}| \leq q+4$, is continuous on the compact set $\square_i$ and is bounded. Taking the maximum over the finitely many charts proves (84). The functions $\bar{\rho}_i$ are smooth because the partition of unity is smooth, and $\sqrt{\det \boldsymbol{G}_i}$ is smooth because $\boldsymbol{G}_i = \boldsymbol{J}_i^\top \boldsymbol{J}_i$ is smooth and uniformly positive definite on $\square_i$. Their derivatives up to order $q+4$ are therefore bounded on compact boxes, proving (85) after increasing $\Gamma_{\mathcal{M},q}$.

Transition maps and $\boldsymbol{A}_i = \boldsymbol{G}_i^{-1} \boldsymbol{J}_i^\top$ are compositions, products, and matrix inverses of smooth maps on compact sets where the relevant determinants are bounded away from zero, so their derivatives up to order $q+4$ are bounded. Since $\mathcal{M}$ has positive reach κ, the nearest projection $\Pi_{\mathcal{M}}$ is C^∞ on every closed tube of radius strictly smaller than κ. The construction only evaluates $\Pi_{\mathcal{M}}$ on such compact tubes. Therefore $\boldsymbol{x} \mapsto \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}))$ and the associated residual maps have bounded derivatives up to order $q+4$ on those tubes. If the argument is $\boldsymbol{x}/m_t$, the chain rule only adds fixed powers of m_t^{-1}, which are bounded by $\underline{m}^{-1}$ on $[\underline{t}, \overline{t}]$. The atlas cardinality and tube constant obey the reach-volume bounds

$$C_{\mathcal{M}} \leq C_d S_{\mathcal{M}} (1 \vee \kappa^{-1})^d, \qquad C_{\mathrm{tube}} \leq C_d (1 \vee S_{\mathcal{M}})(1 \vee \kappa^{-1})^d.$$

All remaining auxiliary constants are among the finite C^{q+4}-norms and buffer reciprocals dominated by $\mathcal{K}_{\mathcal{M},q}$, or are obtained from them by finitely many products, compositions, and matrix inverses of dimension at most d. Enlarging $C_{d,q}$ and $\mathfrak{a}_{\mathrm{geom}}(q)$ gives the displayed explicit bound for $\Gamma_{\mathcal{M},q}$, and completes the proof. □

Let

$$\delta_{\mathrm{act}}^{(c_a)}(\varepsilon, t) := c_a \frac{\sigma_t}{m_t} \sqrt{\log \varepsilon^{-1}}, \qquad \boldsymbol{\xi}(\boldsymbol{x}, t) := \frac{\boldsymbol{x}}{m_t}. \tag{86}$$

We write

$$\delta_{\mathrm{act}}(\varepsilon, t) := \delta_{\mathrm{act}}^{(c_\star)}(\varepsilon, t) \tag{87}$$

Since the atlas is finite, define the uniform projection and Gauss–Newton margin

$$\eta_\Pi := \min_{1 \leq i \leq C_{\mathcal{M}}} \min\Big\{\kappa, \frac{1}{2} B_i^{\mathrm{buff}}, \frac{\lambda_i^{(-)}}{\mathsf{M}_i^{(2)}}, \frac{1}{2L_{A,i}}\Big\} > 0. \tag{88}$$

All projection-centered arguments below are used only for $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$. Choose $\varepsilon > 0$ small enough that, for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$,

$$\delta_{\mathrm{act}}(\varepsilon, t) < \frac{1}{2}\eta_{\Pi}. \tag{89}$$

For a fixed $c_a > 0$, the corresponding c_a-active localization condition is the same display with $\delta_{\mathrm{act}}^{(c_a)}$ in place of δ_{act}.

For the Laplace localization argument, define the compact ambient set

$$\mathcal{X}_i := \{\boldsymbol{\zeta} \in \mathbb{R}^D : \mathrm{dist}_{\mathbb{R}^D}(\boldsymbol{\zeta}, \mathcal{S}_i) \le \eta_{\Pi}/2\}. \tag{90}$$

Since $\boldsymbol{z}_i$ is smooth on $\square_i$ and $\mathcal{X}_i$ is compact, there is a finite constant $H_i \ge 1$ such that, for $\Phi_i(\boldsymbol{u}; \boldsymbol{\zeta}) := \frac{1}{2}\|\boldsymbol{\zeta} - \boldsymbol{z}_i(\boldsymbol{u})\|_2^2$,

$$\left\|\nabla_{\boldsymbol{u}}^2 \Phi_i(\boldsymbol{u}; \boldsymbol{\zeta}) - \nabla_{\boldsymbol{u}}^2 \Phi_i(\boldsymbol{v}; \boldsymbol{\zeta})\right\|_{\mathrm{op}} \le H_i \|\boldsymbol{u} - \boldsymbol{v}\|_2 \tag{91}$$

for all $\boldsymbol{u}, \boldsymbol{v} \in \square_i$ and $\boldsymbol{\zeta} \in \mathcal{X}_i$. Define

$$r_i := \min\Big\{d_i^{\square}/2,\ \lambda_i^{(-)}/[4(1 + H_i)]\Big\} > 0. \tag{92}$$

For each chart i, define the fixed chart-geometry constant

$$\begin{aligned}\mathfrak{C}_i := 1 \vee L_i^{(+)} \vee (L_i^{(-)})^{-1} \vee \lambda_i^{(+)} \vee (\lambda_i^{(-)})^{-1} \vee \mathsf{M}_i^{(2)} \vee L_{A,i} \vee H_i \vee (d_i^{\square})^{-1} \vee r_i^{-1} \\ \vee \max_{1 \le \ell \le D} \|\boldsymbol{z}_{i,\ell}\|_{C^{\lfloor \beta \rfloor + 3}(\square_i)} \vee \|\boldsymbol{G}_i\|_{C^{\lfloor \beta \rfloor + 2}(\square_i)} \vee \|\boldsymbol{G}_i^{-1}\|_{C^{\lfloor \beta \rfloor + 2}(\square_i)}.\end{aligned}$$

Since $\square_i \Subset \phi_i(U_i)$ and $\boldsymbol{z}_i$ is smooth, this quantity is finite. Since the atlas is finite, set

$$\mathfrak{C}_{\mathcal{M}} := \max_{1 \le i \le C_{\mathcal{M}}} \mathfrak{C}_i < \infty. \tag{93}$$

Lemma F.4 (Nearest-point projection on the active small-noise chart)**.** *Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and (89). If $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, then the global nearest-point projection $\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))$ is well-defined and unique. Moreover, $\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t)) \in U_i$, and*

$$\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) := \phi_i\big(\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\big) \in K_i^{\circ}. \tag{94}$$

Furthermore, with

$$\boldsymbol{r}_i(\boldsymbol{x}, t) := \boldsymbol{x} - m_t \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t)),$$

we have the orthogonality relation

$$\boldsymbol{J}_i\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\big)^{\top} \boldsymbol{r}_i(\boldsymbol{x}, t) = 0. \tag{95}$$

Proof. Write $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{x}, t)$. Since $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, there exists $\boldsymbol{y} \in \mathcal{S}_i$ such that

$$\|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \le \delta_{\mathrm{act}}(\varepsilon, t). \tag{96}$$

Because $\boldsymbol{y} \in \mathcal{M}$, we have

$$\mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \le \delta_{\mathrm{act}}(\varepsilon, t) < \kappa. \tag{97}$$

Thus $\boldsymbol{\xi}$ lies inside the reach tube of $\mathcal{M}$, so the nearest-point projection $\Pi_{\mathcal{M}}(\boldsymbol{\xi})$ is well-defined and unique.

Next, since $\Pi_{\mathcal{M}}(\boldsymbol{\xi})$ is the closest point on $\mathcal{M}$ to $\boldsymbol{\xi}$,

$$\|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 = \mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \|\boldsymbol{\xi} - \boldsymbol{y}\|_2.$$

Therefore,

$$\begin{aligned}\|\Pi_{\mathcal{M}}(\boldsymbol{\xi}) - \boldsymbol{y}\|_2 &\le \|\Pi_{\mathcal{M}}(\boldsymbol{\xi}) - \boldsymbol{\xi}\|_2 + \|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \\ &= \mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) + \|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \\ &\le 2\|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \le 2\delta_{\mathrm{act}}(\varepsilon, t) < \frac{1}{2} B_i^{\mathrm{buff}}.\end{aligned}$$

Hence

$$\mathrm{dist}_{\mathbb{R}^D}(\Pi_{\mathcal{M}}(\boldsymbol{\xi}), \mathcal{S}_i) < \frac{1}{2} B_i^{\mathrm{buff}}.$$

By the definition of B_i^{buff}, this implies $\Pi_{\mathcal{M}}(\boldsymbol{\xi}) \in \phi_i^{-1}(\mathrm{int}(K_i^\circ))$. Hence

$$\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) = \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{\xi})) \in K_i^\circ.$$

In particular, $\Pi_{\mathcal{M}}(\boldsymbol{\xi}) \in U_i$.

It remains to prove the orthogonality statement. Since $\boldsymbol{\xi}$ lies inside the reach tube, the metric projection residual is normal to the tangent space:

$$\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) \perp T_{\Pi_{\mathcal{M}}(\boldsymbol{\xi})}\mathcal{M}.$$

The columns of $\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi}) = \nabla \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi})$ form a basis of this tangent space. Therefore,

$$\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top \big(\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\big) = 0.$$

Since $\boldsymbol{r}_i(\boldsymbol{x}, t) = \boldsymbol{x} - m_t \Pi_{\mathcal{M}}(\boldsymbol{\xi}) = m_t\big(\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\big)$, multiplying the previous display by m_t gives

$$\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top \boldsymbol{r}_i(\boldsymbol{x}, t) = 0.$$

□

For $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, define

$$\Phi_i(\boldsymbol{u}; \boldsymbol{x}, t) := \frac{1}{2}\|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{z}_i(\boldsymbol{u})\|_2^2, \qquad \boldsymbol{g}_i(\boldsymbol{u}; \boldsymbol{x}, t) := \nabla_{\boldsymbol{u}} \Phi_i(\boldsymbol{u}; \boldsymbol{x}, t). \tag{98}$$

Since

$$\Phi_i(\boldsymbol{u}; \boldsymbol{x}, t) = \frac{1}{2} \sum_{\ell=1}^{D} \big(\xi_\ell(\boldsymbol{x}, t) - z_{i,\ell}(\boldsymbol{u})\big)^2,$$

we have

$$\boldsymbol{g}_i(\boldsymbol{u}; \boldsymbol{x}, t) = \boldsymbol{J}_i(\boldsymbol{u})^\top \big(\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{\xi}(\boldsymbol{x}, t)\big). \tag{99}$$

Differentiating once more gives

$$\nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}; \boldsymbol{x}, t) = \boldsymbol{G}_i(\boldsymbol{u}) - \sum_{\ell=1}^{D} \big(\xi_\ell(\boldsymbol{x}, t) - z_{i,\ell}(\boldsymbol{u})\big) \nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u}). \tag{100}$$

At $\boldsymbol{u} = \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)$, we have $\boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)) = \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))$. Hence, using (95),

$$\begin{aligned} \boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t); \boldsymbol{x}, t) &= \boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top \big(\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t)) - \boldsymbol{\xi}(\boldsymbol{x}, t)\big) \\ &= -m_t^{-1} \boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top \boldsymbol{r}_i(\boldsymbol{x}, t) = 0. \end{aligned} \tag{101}$$

Lemma F.5 (Uniform invertibility near the projection)**.** *Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and (89). If $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, then*

$$\nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t); \boldsymbol{x}, t) \succeq \frac{1}{2} \lambda_i^{(-)} \boldsymbol{I}_d. \tag{102}$$

Proof. Write $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{x}, t)$ and $\boldsymbol{u}_i^{\Pi} = \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)$. By Theorem F.4,

$$\boldsymbol{u}_i^{\Pi} \in K_i^\circ \subset \square_i, \qquad \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi}) = \Pi_{\mathcal{M}}(\boldsymbol{\xi}).$$

Fix any $\boldsymbol{w} \in \mathbb{R}^d$. By (100),

$$\boldsymbol{w}^\top \nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}; \boldsymbol{x}, t) \boldsymbol{w} = \boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u}_i^{\Pi}) \boldsymbol{w} - \sum_{\ell=1}^{D} \big(\xi_\ell - z_{i,\ell}(\boldsymbol{u}_i^{\Pi})\big) \boldsymbol{w}^\top \nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u}_i^{\Pi}) \boldsymbol{w}. \tag{103}$$

Because $\boldsymbol{u}_i^{\Pi} \in \square_i$, (83) gives $\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u}_i^{\Pi})\boldsymbol{w} \geq \lambda_i^{(-)}\|\boldsymbol{w}\|_2^2$. For the perturbation term, use $\boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi}) = \Pi_{\mathcal{M}}(\boldsymbol{\xi})$ and the definition of $\mathsf{M}_i^{(2)}$:

$$\begin{aligned}\left|\sum_{\ell=1}^{D}\big(\xi_\ell - z_{i,\ell}(\boldsymbol{u}_i^{\Pi})\big)\boldsymbol{w}^\top \nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u}_i^{\Pi})\boldsymbol{w}\right| &\leq \sum_{\ell=1}^{D}|\xi_\ell - z_{i,\ell}(\boldsymbol{u}_i^{\Pi})|\|\nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u}_i^{\Pi})\|_{\mathrm{op}}\|\boldsymbol{w}\|_2^2 \\ &\leq \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 \sum_{\ell=1}^{D}\|\nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u}_i^{\Pi})\|_{\mathrm{op}}\|\boldsymbol{w}\|_2^2 \\ &\leq \mathrm{dist}(\boldsymbol{\xi}, \mathcal{M})\,\mathsf{M}_i^{(2)}\|\boldsymbol{w}\|_2^2. \end{aligned} \tag{104}$$

Since $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, there exists $\boldsymbol{y} \in \mathcal{S}_i \subset \mathcal{M}$ such that $\|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \leq \delta_{\mathrm{act}}(\varepsilon, t)$. Therefore $\mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \leq \delta_{\mathrm{act}}(\varepsilon, t)$. By (88) and (89),

$$\delta_{\mathrm{act}}(\varepsilon, t)\mathsf{M}_i^{(2)} < \frac{1}{2}\lambda_i^{(-)}.$$

Combining this with (103) and (104) and $\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u}_i^{\Pi})\boldsymbol{w} \geq \lambda_i^{(-)}\|\boldsymbol{w}\|_2^2$, we obtain

$$\boldsymbol{w}^\top \nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}; \boldsymbol{x}, t)\boldsymbol{w} \geq \frac{1}{2}\lambda_i^{(-)}\|\boldsymbol{w}\|_2^2.$$

Since this holds for every $\boldsymbol{w} \in \mathbb{R}^d$, the matrix inequality

$$\nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}; \boldsymbol{x}, t) \succeq \lambda_i^{(-)}\boldsymbol{I}_d/2$$

□

F.2. Finite-anchor projection networks

This section isolates the neural implementation of the nearest-point projection. It is the only nonlinear geometric approximation step that could otherwise create a hidden ambient-dimensional cost. The construction works strictly in intrinsic chart coordinates: finitely many anchors initialize local Gauss–Newton iterations, and a ReLU objective gate selects near-minimizing candidates. Later network-assembly sections reuse these projection-coordinate networks together with standard schedule, product, reciprocal, and exponential subnetworks.

Throughout this section, every construction involving m_t, σ_t, or powers of their reciprocals uses Theorem I.2 and the derived schedule lemmas stated below.

For orientation, the projection-centered network assembly uses the following objects in this order:

(1) $\boldsymbol{\xi}(\boldsymbol{x}, t) = \boldsymbol{x}/m_t$,

(2) the global nearest-point projection $\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))$,

(3) the local coordinate $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) = \phi_i\big(\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\big)$,

(4) the normalized residual $\boldsymbol{\nu}_i(\boldsymbol{x}, t) = \sigma_t^{-1}\boldsymbol{r}_i = \sigma_t^{-1}\big(\boldsymbol{x} - m_t\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\big)$,

(5) the Gaussian factor $\exp(-\|\boldsymbol{\nu}_i\|_2^2/2)$,

(6) and the coefficient functions $A_{i,q,\boldsymbol{\lambda}}^{(R)}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t), t\big)$.

The projection-coordinate networks are the only part that uses the manifold geometry in a substantial way. Once the projection center is available, the residual, Gaussian factor, and Laplace coefficients are obtained by combining standard schedule, Hölder, multiplication, reciprocal, and exponential subnetworks with carefully allocated errors.

In the following, we construct subnetworks to approximate the components of (189):

Function	Approximation lemma
$t \mapsto m_t$ $t \mapsto \sigma_t$	Theorem I.2
$t \mapsto \sigma_t^k$	Theorem I.8
$t \mapsto 1/\sigma_t$ $t \mapsto 1/\sigma_t^k$	Theorem I.9
$t \mapsto 1/m_t$ $(\boldsymbol{x}, t) \mapsto \boldsymbol{\xi}(\boldsymbol{x}, t) := \boldsymbol{x}/m_t$	Theorem I.6 Theorem I.7
$(\boldsymbol{x}, t) \to \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))$ $(\boldsymbol{x}, t) \to \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) := \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t)))$	Theorem F.14 Theorem F.11
$(\boldsymbol{x}, t) \mapsto \boldsymbol{r}_i(\boldsymbol{x}, t) := \boldsymbol{x} - m_t \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))$ $(\boldsymbol{x}, t) \mapsto \boldsymbol{\nu}_i(\boldsymbol{x}, t) := \sigma_t^{-1} \boldsymbol{r}_i(\boldsymbol{x}, t)$	Theorem F.17
$(\boldsymbol{x}, t) \mapsto \exp\big(-\frac{\lVert \boldsymbol{r}_i(\boldsymbol{x}, t) \rVert_2^2}{2\sigma_t^2}\big)$	Theorem F.18
$(\boldsymbol{x}, t) \to A_{i,q,\boldsymbol{\lambda}}^{(\mathcal{Q})}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t), t\big)$ $(\boldsymbol{x}, t) \to A_{i,q,\boldsymbol{\lambda}}^{(\mathcal{P})}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t), t\big)$	Theorems F.20 and F.21
active chart expansion for $\mathcal{Q}_i$ active chart expansion for $\mathcal{P}_i$	Theorem F.25

F.2.1. Gauss–Newton networks for nearest-point projection

We approximate the nearest-point projection in local intrinsic coordinates. The nonlinear unknown is the chart coordinate $\boldsymbol{u} \in \square_i \subset \mathbb{R}^d$, whereas the ambient variable enters the iteration only through $\boldsymbol{\xi}(\boldsymbol{x}, t) = \boldsymbol{x}/m_t$ and matrix-vector operations. The lemmas in this block first prove a stable intrinsic Gauss–Newton step, then initialize it from finitely many anchors, and finally combine all chart-anchor candidates by an objective gate. This produces coordinate-valued and ambient-valued projection networks without ever applying a chart map to an approximate ambient point.

We define the coordinate-wise clipping map:

$$\Pi_i^{(\mathrm{box})}(\boldsymbol{u}) := \big(a_{i,j}^{\square} + \mathsf{ReLU}(u_j - a_{i,j}^{\square}) - \mathsf{ReLU}(u_j - b_{i,j}^{\square})\big)_{j=1}^{d}. \tag{105}$$

$\Pi_i^{(\mathrm{box})}$ explicitly translates standard $\min / \max$ bounding operations into exact ReLU arithmetic. Consequently, $\Pi_i^{(\mathrm{box})} : \mathbb{R}^d \to \square_i$ is exactly representable by a shallow ReLU network without any approximation error. Furthermore, because it is a projection onto a convex set, it is inherently 1-Lipschitz.

Lemma F.6 (Local Gauss–Newton contraction for nearest-point projection)**.** *Define the intrinsic Gauss–Newton map*

$$\mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t) := \Pi_i^{(\mathrm{box})}\big(\boldsymbol{u} + \boldsymbol{A}_i(\boldsymbol{u})\big(\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{z}_i(\boldsymbol{u})\big)\big), \qquad \textit{where } \boldsymbol{A}_i(\boldsymbol{u}) = \boldsymbol{G}_i(\boldsymbol{u})^{-1} \boldsymbol{J}_i(\boldsymbol{u})^{\top}. \tag{106}$$

Set

$$r_i^{\mathrm{GN}} := \min\Big\{1, \frac{1}{2}\mathrm{dist}(K_i^{\circ}, \partial\square_i), \frac{1}{4L_{A,i}L_i^{(+)}}\Big\}. \tag{107}$$

Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$, $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, *and* (89)*. Then*

$$\lVert \mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \rVert_2 \le \frac{1}{2} \lVert \boldsymbol{u} - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \rVert_2 \tag{108}$$

whenever

$$\boldsymbol{u} \in \mathcal{B}(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t), r_i^{\mathrm{GN}}) \cap \square_i.$$

Proof. Write $\boldsymbol{\xi} := \boldsymbol{\xi}(\boldsymbol{x}, t), \boldsymbol{u}^{\Pi} := \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)$. Before clipping, define

$$\overline{\mathcal{T}}_i(\boldsymbol{u}; \boldsymbol{x}, t) := \boldsymbol{u} + \boldsymbol{A}_i(\boldsymbol{u})\big(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\big).$$

Since

$$\boldsymbol{A}_i(\boldsymbol{u})\boldsymbol{J}_i(\boldsymbol{u}) = \boldsymbol{G}_i(\boldsymbol{u})^{-1}\boldsymbol{J}_i(\boldsymbol{u})^{\top}\boldsymbol{J}_i(\boldsymbol{u}) = \boldsymbol{I}_d,$$

differentiating with respect to $\boldsymbol{u}$ gives, for every $\boldsymbol{h} \in \mathbb{R}^d$,

$$\begin{aligned} \mathrm{D}_{\boldsymbol{u}}\overline{\mathcal{T}}_i(\boldsymbol{u};\boldsymbol{x},t)[\boldsymbol{h}] &= \boldsymbol{h} + \mathrm{D}\boldsymbol{A}_i(\boldsymbol{u})[\boldsymbol{h}]\big(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\big) - \boldsymbol{A}_i(\boldsymbol{u})\boldsymbol{J}_i(\boldsymbol{u})\boldsymbol{h} \\ &= \mathrm{D}\boldsymbol{A}_i(\boldsymbol{u})[\boldsymbol{h}]\big(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\big). \end{aligned} \tag{109}$$

By Theorem F.4,

$$\boldsymbol{z}_i(\boldsymbol{u}^\Pi) = \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \qquad \boldsymbol{J}_i(\boldsymbol{u}^\Pi)^\top\big(\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\big) = 0.$$

Therefore

$$\boldsymbol{A}_i(\boldsymbol{u}^\Pi)\big(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u}^\Pi)\big) = \boldsymbol{G}_i(\boldsymbol{u}^\Pi)^{-1}\boldsymbol{J}_i(\boldsymbol{u}^\Pi)^\top\big(\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\big) = 0,$$

and hence

$$\overline{\mathcal{T}}_i(\boldsymbol{u}^\Pi;\boldsymbol{x},t) = \boldsymbol{u}^\Pi. \tag{110}$$

Since $\boldsymbol{u}^\Pi \in K_i^\circ \subset \square_i$, the clipping map fixes $\boldsymbol{u}^\Pi$:

$$\Pi_i^{(\mathrm{box})}(\boldsymbol{u}^\Pi) = \boldsymbol{u}^\Pi. \tag{111}$$

Now let $\boldsymbol{u} \in \mathcal{B}(\boldsymbol{u}^\Pi, r_i^{\mathrm{GN}}) \cap \square_i$. For $s \in [0,1]$, set

$$\boldsymbol{u}_s := \boldsymbol{u}^\Pi + s(\boldsymbol{u} - \boldsymbol{u}^\Pi).$$

Because $\square_i$ is convex and both $\boldsymbol{u}^\Pi, \boldsymbol{u} \in \square_i$, we have $\boldsymbol{u}_s \in \square_i$. Also,

$$\|\boldsymbol{u}_s - \boldsymbol{u}^\Pi\|_2 \le \|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \le r_i^{\mathrm{GN}}.$$

Using (89), the definition of η_Π, and the fact that $(\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)$, we have

$$\mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta_{\mathrm{act}}(\varepsilon,t) < \frac{1}{4L_{A,i}}. \tag{112}$$

Therefore, by the chart Lipschitz bound (82),

$$\begin{aligned} \|\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 &\le \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 + \|\Pi_{\mathcal{M}}(\boldsymbol{\xi}) - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 \\ &= \mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) + \|\boldsymbol{z}_i(\boldsymbol{u}^\Pi) - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 \\ &\le \delta_{\mathrm{act}}(\varepsilon,t) + L_i^{(+)}\|\boldsymbol{u}_s - \boldsymbol{u}^\Pi\|_2 \\ &< \frac{1}{4L_{A,i}} + L_i^{(+)} r_i^{\mathrm{GN}} \\ &\le \frac{1}{4L_{A,i}} + \frac{1}{4L_{A,i}} = \frac{1}{2L_{A,i}}. \end{aligned} \tag{113}$$

Combining (81), (109) and (113), we get

$$\|\mathrm{D}_{\boldsymbol{u}}\overline{\mathcal{T}}_i(\boldsymbol{u}_s;\boldsymbol{x},t)\|_{\mathrm{op}} \le L_{A,i}\|\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 \le \frac{1}{2}.$$

By the fundamental theorem of calculus and (110),

$$\begin{aligned} \big\|\overline{\mathcal{T}}_i(\boldsymbol{u};\boldsymbol{x},t) - \boldsymbol{u}^\Pi\big\|_2 &= \big\|\overline{\mathcal{T}}_i(\boldsymbol{u};\boldsymbol{x},t) - \overline{\mathcal{T}}_i(\boldsymbol{u}^\Pi;\boldsymbol{x},t)\big\|_2 \\ &\le \int_0^1 \|\mathrm{D}_{\boldsymbol{u}}\overline{\mathcal{T}}_i(\boldsymbol{u}_s;\boldsymbol{x},t)\|_{\mathrm{op}}\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \mathrm{d}s \\ &\le \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2. \end{aligned}$$

Finally, $\Pi_i^{(\mathrm{box})}$ is 1-Lipschitz and fixes $\boldsymbol{u}^\Pi$. Thus,

$$\begin{aligned} \|\mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t) - \boldsymbol{u}^\Pi\|_2 &= \big\|\Pi_i^{(\mathrm{box})}(\overline{\mathcal{T}}_i(\boldsymbol{u};\boldsymbol{x},t)) - \Pi_i^{(\mathrm{box})}(\boldsymbol{u}^\Pi)\big\|_2 \\ &\le \big\|\overline{\mathcal{T}}_i(\boldsymbol{u};\boldsymbol{x},t) - \boldsymbol{u}^\Pi\big\|_2 \le \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2. \end{aligned}$$

□

Lemma F.7 (ReLU anchor networks for projection initialization)**.** *For every chart i, there exists a finite set*

$$\mathcal{A}_i^{\mathrm{anc}} = \{\boldsymbol{a}_{i,1}, \dots, \boldsymbol{a}_{i,N_i}\} \subset K_i^\circ$$

depending only on the fixed chart geometry such that

$$\forall \boldsymbol{u} \in K_i^\circ, \qquad \min_{\boldsymbol{a} \in \mathcal{A}_i^{\mathrm{anc}}} \|\boldsymbol{a} - \boldsymbol{u}\|_2 \le \frac{1}{2} r_i^{\mathrm{GN}}. \tag{114}$$

Moreover, writing $\ell_{i,r}^\circ := b_{i,r}^\circ - a_{i,r}^\circ$,

$$N_i \le \prod_{r=1}^{d} \left(1 + \left\lceil \frac{\sqrt{d}\,\ell_{i,r}^\circ}{r_i^{\mathrm{GN}}} \right\rceil\right) \le \prod_{r=1}^{d} \left(2 + \frac{\sqrt{d}\,\ell_{i,r}^\circ}{r_i^{\mathrm{GN}}}\right) \le C_i^{\mathrm{anc}} (r_i^{\mathrm{GN}})^{-d}, \qquad C_i^{\mathrm{anc}} := \prod_{r=1}^{d} \left(2 + \sqrt{d}\,\ell_{i,r}^\circ\right). \tag{115}$$

In particular, $\mathcal{A}_i^{\mathrm{anc}}$ is fixed once and for all and is independent of ε, ε_Π, and $(\boldsymbol{x}, t)$.

Proof. For each coordinate r, let $m_{i,r} := \left\lceil \sqrt{d}\,\ell_{i,r}^\circ / r_i^{\mathrm{GN}} \right\rceil$. Choose the tensor-product grid

$$\mathcal{A}_i^{\mathrm{anc}} := \prod_{r=1}^{d} \left\{ a_{i,r}^\circ + \frac{k_r}{m_{i,r}} \ell_{i,r}^\circ : k_r = 0, \dots, m_{i,r} \right\} \subset K_i^\circ.$$

For any $\boldsymbol{u} \in K_i^\circ$, choose the nearest grid point in each coordinate. The coordinatewise error is at most $\ell_{i,r}^\circ/(2m_{i,r}) \le r_i^{\mathrm{GN}}/(2\sqrt{d})$, and therefore the Euclidean error is at most $r_i^{\mathrm{GN}}/2$. This proves (114). The grid cardinality is

$$|\mathcal{A}_i^{\mathrm{anc}}| = \prod_{r=1}^{d} (m_{i,r} + 1) = \prod_{r=1}^{d} \left(1 + \left\lceil \frac{\sqrt{d}\,\ell_{i,r}^\circ}{r_i^{\mathrm{GN}}} \right\rceil\right),$$

and the final inequality in (115) follows from $1 + \lceil x \rceil \le 2 + x$, together with $r_i^{\mathrm{GN}} \le 1$. □

Define the aggregate anchor complexity

$$A_{\mathcal{M}} := \sum_{i=1}^{C_{\mathcal{M}}} |\mathcal{A}_i^{\mathrm{anc}}|. \tag{116}$$

By (115),

$$A_{\mathcal{M}} \le \sum_{i=1}^{C_{\mathcal{M}}} C_i^{\mathrm{anc}} (r_i^{\mathrm{GN}})^{-d}. \tag{117}$$

Lemma F.8 (Error propagation for anchored approximate Gauss–Newton iteration)**.** *Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$, $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, and (89). Assume $\boldsymbol{a} \in K_i^\circ$ satisfies*

$$\|\boldsymbol{a} - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\|_2 \le \frac{1}{2} r_i^{\mathrm{GN}}. \tag{118}$$

Let $\widetilde{\mathcal{T}}_i$ be an approximate Gauss–Newton step satisfying

$$\sup_{\boldsymbol{u} \in \square_i,\ t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon),\ (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)} \left\| \widetilde{\mathcal{T}}_i(\boldsymbol{u}; \boldsymbol{x}, t) - \mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t) \right\|_\infty \le \varepsilon_{\mathrm{step}}. \tag{119}$$

Assume moreover that

$$\varepsilon_{\mathrm{step}} \le \frac{r_i^{\mathrm{GN}}}{8\sqrt{d}}. \tag{120}$$

Define the approximate iterates by

$$\widetilde{\boldsymbol{u}}_0 := \boldsymbol{a}, \qquad \widetilde{\boldsymbol{u}}_{k+1} := \widetilde{\mathcal{T}}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t). \tag{121}$$

Then, every iterate stays in the Gauss–Newton contraction basin:

$$\widetilde{\boldsymbol{u}}_k \in \mathcal{B}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), r_i^{\mathrm{GN}}\big) \cap \square_i, \qquad k \ge 0. \tag{122}$$

Moreover, for every $K \ge 0$,

$$\left\| \widetilde{\boldsymbol{u}}_K - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t) \right\|_2 \le 2^{-(K+1)} r_i^{\mathrm{GN}} + 2\sqrt{d}\,\varepsilon_{\mathrm{step}}. \tag{123}$$

Proof. Write $\boldsymbol{u}^{\Pi} := \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t), e_k := \|\widetilde{\boldsymbol{u}}_k - \boldsymbol{u}^{\Pi}\|_2$. By the anchor assumption (118),

$$e_0 = \|\boldsymbol{a} - \boldsymbol{u}^{\Pi}\|_2 \le \frac{1}{2} r_i^{\mathrm{GN}}.$$

We first prove that the iterates stay inside the contraction basin. Assume inductively that

$$e_k \le \frac{3}{4} r_i^{\mathrm{GN}}.$$

Then $\widetilde{\boldsymbol{u}}_k \in \mathcal{B}(\boldsymbol{u}^{\Pi}, r_i^{\mathrm{GN}}) \cap \square_i$, so the exact Gauss–Newton contraction lemma (Theorem F.6) applies:

$$\|\mathcal{T}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t) - \boldsymbol{u}^{\Pi}\|_2 \le \frac{1}{2} e_k.$$

Using the approximate-step error and the norm comparison $\|\boldsymbol{v}\|_2 \le \sqrt{d}\,\|\boldsymbol{v}\|_\infty$, we obtain

$$\begin{aligned} e_{k+1} &= \|\widetilde{\mathcal{T}}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t) - \boldsymbol{u}^{\Pi}\|_2 \\ &\le \left\|\widetilde{\mathcal{T}}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t) - \mathcal{T}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t)\right\|_2 + \|\mathcal{T}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t) - \boldsymbol{u}^{\Pi}\|_2 \\ &\le \sqrt{d}\varepsilon_{\mathrm{step}} + \frac{1}{2} e_k \\ &\le \sqrt{d}\varepsilon_{\mathrm{step}} + \frac{3}{8} r_i^{\mathrm{GN}}. \end{aligned}$$

By (120),

$$\sqrt{d}\,\varepsilon_{\mathrm{step}} \le \frac{1}{8} r_i^{\mathrm{GN}}.$$

Therefore

$$e_{k+1} \le \frac{1}{8} r_i^{\mathrm{GN}} + \frac{3}{8} r_i^{\mathrm{GN}} = \frac{1}{2} r_i^{\mathrm{GN}} < \frac{3}{4} r_i^{\mathrm{GN}}.$$

Since $e_0 \le r_i^{\mathrm{GN}}/2$, induction proves

$$e_k \le \frac{3}{4} r_i^{\mathrm{GN}} < r_i^{\mathrm{GN}} \qquad \text{for all } k \ge 0.$$

Also, because the approximate step $\widetilde{\mathcal{T}}_i$ is assumed to be $\square_i$-valued, every $\widetilde{\boldsymbol{u}}_k$ lies in $\square_i$. Hence (122) holds.

Now that the exact contraction applies at every step, we have the recursion

$$e_{k+1} \le \frac{1}{2} e_k + \sqrt{d}\,\varepsilon_{\mathrm{step}}. \tag{124}$$

Iterating (124) gives

$$\begin{aligned} e_K &\le 2^{-K} e_0 + \sqrt{d}\,\varepsilon_{\mathrm{step}} \sum_{j=0}^{K-1} 2^{-j} \\ &\le 2^{-(K+1)} r_i^{\mathrm{GN}} + 2\sqrt{d}\,\varepsilon_{\mathrm{step}}. \end{aligned}$$

□

Lemma F.9 (Neural approximation of one Gauss–Newton step)**.** *Fix a localization parameter $0 < \varepsilon \le 1$. For every $\varepsilon_{\mathcal{T}} \in (0, 1/2]$, there exists a ReLU network $\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with input $(\boldsymbol{u}, \boldsymbol{x}, t) \in \mathbb{R}^d \times \mathbb{R}^D \times [\underline{t}, \overline{t}]$ and output in $\square_i$ such that*

$$\sup_{\boldsymbol{u} \in \square_i, t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon), (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)} \left\|\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}}(\boldsymbol{u}, \boldsymbol{x}, t) - \mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t)\right\|_\infty \le \varepsilon_{\mathcal{T}}. \tag{125}$$

Moreover, for a finite chart-dependent constant $C_{i,\mathcal{T}} \ge 1$,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^2(C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1}), \\ \|\boldsymbol{W}\|_\infty &\lesssim dD(d \vee \beta)(C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1})^{\frac{d}{2(\beta \vee 1)}} + D \log^3(C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1}), \\ S &\lesssim dD(d + 2(\beta \vee 1))^{d+3} (C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1})^{\frac{d}{2(\beta \vee 1)}} \log(C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1}) + dD \log^4(C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1}), \\ \log B &\lesssim \log^2(C_{i,\mathcal{T}} \varepsilon_{\mathcal{T}}^{-1}). \end{aligned} \tag{126}$$

Proof. Recall that

$$\mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t) = \Pi_i^{(\mathrm{box})}\Big(\boldsymbol{u} + \boldsymbol{A}_i(\boldsymbol{u})\big(\boldsymbol{\xi}(\boldsymbol{x},t) - \boldsymbol{z}_i(\boldsymbol{u})\big)\Big), \qquad \boldsymbol{A}_i(\boldsymbol{u}) = \boldsymbol{G}_i(\boldsymbol{u})^{-1}\boldsymbol{J}_i(\boldsymbol{u})^\top. \tag{127}$$

The clipping map $\Pi_i^{(\mathrm{box})}$ is exactly representable by a shallow ReLU network and is 1-Lipschitz.

Step 1: Uniform bounds. Since $\square_i$ is compact and $\boldsymbol{z}_i, \boldsymbol{A}_i$ are C^∞ on $\square_i$, there exists a finite constant $B_{i,\mathcal{T}} \geq 1$ such that every scalar component of $\boldsymbol{z}_i$ and $\boldsymbol{A}_i$ belongs to $\mathcal{H}^{2(\beta\vee 1)}(\square_i, B_{i,\mathcal{T}})$, and

$$\|\boldsymbol{z}_i(\boldsymbol{u})\|_\infty + \|\boldsymbol{A}_i(\boldsymbol{u})\|_\infty \leq B_{i,\mathcal{T}}, \qquad \boldsymbol{u} \in \square_i. \tag{128}$$

Here $\|\boldsymbol{A}_i(\boldsymbol{u})\|_\infty$ denotes the maximum absolute entry of the $d \times D$ matrix $\boldsymbol{A}_i(\boldsymbol{u})$.

Moreover, on the active small-noise region there exists a finite constant $M_{x,i} \geq 1$ such that

$$\|\boldsymbol{x}\|_\infty \leq M_{x,i}, \qquad \|\boldsymbol{\xi}(\boldsymbol{x},t)\|_\infty = \Big\|\frac{\boldsymbol{x}}{m_t}\Big\|_\infty \leq M_{x,i}\underline{m}^{-1}. \tag{129}$$

Indeed, if $(\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)$, then $\boldsymbol{x}$ is within $\mathcal{O}(\sigma_t\sqrt{\log\varepsilon^{-1}})$ of $m_t\mathcal{S}_i$; on the small-noise branch this is uniformly bounded, and $\mathcal{S}_i$ is compact.

Set

$$C_{i,\times} := 1 \vee B_{i,\mathcal{T}} \vee M_{x,i}\underline{m}^{-1}. \tag{130}$$

Step 2: Approximate the ingredients. Let $\varepsilon_{\mathrm{aux}} \in (0, 1/2]$ be chosen later.

First, by Theorem I.7 with $M_x = M_{x,i}$, there exists a vector-valued ReLU network $\widetilde{\boldsymbol{\xi}}$ such that

$$\big\|\widetilde{\boldsymbol{\xi}}(\boldsymbol{x},t) - \boldsymbol{\xi}(\boldsymbol{x},t)\big\|_\infty \leq \varepsilon_{\mathrm{aux}} \tag{131}$$

uniformly over the active small-noise region. Its size is bounded by

$$\begin{aligned} L_\xi, \log B_\xi &\lesssim \log^2\varepsilon_{\mathrm{aux}}^{-1} + \log^2(C_{i,\times}), \\ \|\boldsymbol{W}_\xi\|_\infty &\lesssim D\big(\log^3\varepsilon_{\mathrm{aux}}^{-1} + \log^3(C_{i,\times})\big), \\ S_\xi &\lesssim D\big(\log^4\varepsilon_{\mathrm{aux}}^{-1} + \log^4(C_{i,\times})\big). \end{aligned} \tag{132}$$

Second, apply Theorem I.1 to the scalar components of $\boldsymbol{z}_i$ and $\boldsymbol{A}_i$ on $\square_i$ after a fixed affine rescaling of $\square_i$ to $[0,1]^d$. Thus there exist ReLU networks $\widetilde{\boldsymbol{z}}_i$ and $\widetilde{\boldsymbol{A}}_i$ such that

$$\|\widetilde{\boldsymbol{z}}_i - \boldsymbol{z}_i\|_{L^\infty(\square_i)} \leq \varepsilon_{\mathrm{aux}}, \qquad \|\widetilde{\boldsymbol{A}}_i - \boldsymbol{A}_i\|_{L^\infty(\square_i)} \leq \varepsilon_{\mathrm{aux}}. \tag{133}$$

Stacking the $D + dD$ scalar approximants gives size

$$\begin{aligned} L_H &\lesssim \log(e(d\vee\beta))\log(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1}), \\ \|\boldsymbol{W}_H\|_\infty &\lesssim dD(d\vee\beta)(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1})^{\frac{d}{2(\beta\vee 1)}}, \\ S_H &\lesssim dD(d + 2(\beta\vee 1))^{d+3}(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1})^{\frac{d}{2(\beta\vee 1)}}\log(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1}), \\ B_H &\leq 1. \end{aligned} \tag{134}$$

Define

$$\boldsymbol{v}(\boldsymbol{u},\boldsymbol{x},t) := \boldsymbol{\xi}(\boldsymbol{x},t) - \boldsymbol{z}_i(\boldsymbol{u}), \qquad \widetilde{\boldsymbol{v}}(\boldsymbol{u},\boldsymbol{x},t) := \widetilde{\boldsymbol{\xi}}(\boldsymbol{x},t) - \widetilde{\boldsymbol{z}}_i(\boldsymbol{u}).$$

By (131) and (133),

$$\|\widetilde{\boldsymbol{v}} - \boldsymbol{v}\|_\infty \leq 2\varepsilon_{\mathrm{aux}}. \tag{135}$$

Also, by the bounds above,

$$|A_{i,j\ell}(\boldsymbol{u})| \leq C_{i,\times}, \qquad |v_\ell(\boldsymbol{u},\boldsymbol{x},t)| \leq 2C_{i,\times}.$$

Increasing $C_{i,\times}$ by an absolute factor if necessary, we may assume $C_{i,\times} \geq 1$ and both true inputs to every multiplication below lie in $[-C_{i,\times}, C_{i,\times}]$.

Step 3: Approximate the matrix-vector product. For each $j = 1, \ldots, d$ and $\ell = 1, \ldots, D$, apply Theorem I.3 with two inputs, bound $C_{i,\times}$, intrinsic multiplication error ε_{aux}, and input perturbation tolerance $2\varepsilon_{\text{aux}}$. This gives a multiplication network satisfying

$$\begin{aligned} &\left|\phi_{\text{mult},j,\ell}\big(\widetilde{A}_{i,j,\ell}(\boldsymbol{u}), \widetilde{v}_\ell(\boldsymbol{u},\boldsymbol{x},t)\big) - A_{i,j,\ell}(\boldsymbol{u}) v_\ell(\boldsymbol{u},\boldsymbol{x},t)\right| \\ &\qquad \leq \varepsilon_{\text{aux}} + 4C_{i,\times}\varepsilon_{\text{aux}} \leq 5C_{i,\times}\varepsilon_{\text{aux}}. \end{aligned} \tag{136}$$

The size of each multiplication network is

$$\begin{aligned} L_{\text{mult}} &\lesssim \log(\varepsilon_{\text{aux}}^{-1}) + \log(C_{i,\times}), \\ \|\boldsymbol{W}_{\text{mult}}\|_\infty &\leq 96, \\ S_{\text{mult}} &\lesssim \log(\varepsilon_{\text{aux}}^{-1}) + \log(C_{i,\times}), \\ \log B_{\text{mult}} &\lesssim \log(C_{i,\times}). \end{aligned} \tag{137}$$

Define the approximate pre-clipped update by

$$\widetilde{q}_j(\boldsymbol{u},\boldsymbol{x},t) := u_j + \sum_{\ell=1}^{D} \phi_{\text{mult},j\ell}\big(\widetilde{A}_{i,j\ell}(\boldsymbol{u}), \widetilde{v}_\ell(\boldsymbol{u},\boldsymbol{x},t)\big), \qquad j = 1, \ldots, d. \tag{138}$$

The exact pre-clipped update is

$$q_j(\boldsymbol{u},\boldsymbol{x},t) := u_j + \sum_{\ell=1}^{D} A_{i,j\ell}(\boldsymbol{u}) v_\ell(\boldsymbol{u},\boldsymbol{x},t).$$

Summing (136) over ℓ gives

$$\|\widetilde{\boldsymbol{q}}(\boldsymbol{u},\boldsymbol{x},t) - \boldsymbol{q}(\boldsymbol{u},\boldsymbol{x},t)\|_\infty \leq 5DC_{i,\times}\varepsilon_{\text{aux}}. \tag{139}$$

Now define the final network by exact clipping:

$$\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}}(\boldsymbol{u},\boldsymbol{x},t) := \Pi_i^{(\text{box})}\big(\widetilde{\boldsymbol{q}}(\boldsymbol{u},\boldsymbol{x},t)\big). \tag{140}$$

Because $\Pi_i^{(\text{box})}$ is 1-Lipschitz,

$$\begin{aligned} \left\|\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}}(\boldsymbol{u},\boldsymbol{x},t) - \mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t)\right\|_\infty &\leq \|\widetilde{\boldsymbol{q}}(\boldsymbol{u},\boldsymbol{x},t) - \boldsymbol{q}(\boldsymbol{u},\boldsymbol{x},t)\|_\infty \\ &\leq 5DC_{i,\times}\varepsilon_{\text{aux}}. \end{aligned}$$

Choose

$$\varepsilon_{\text{aux}} := \frac{\varepsilon_{\mathcal{T}}}{5DC_{i,\times}}. \tag{141}$$

Then (125) follows.

Step 4: Network size. The network is obtained by composing the following pieces: the $\boldsymbol{\xi} = \boldsymbol{x}/m_t$ network from Theorem I.7, the d-dimensional Hölder networks for $\boldsymbol{z}_i$ and $\boldsymbol{A}_i$ from Theorem I.1, the dD multiplication networks from Theorem I.3, exact affine summations, and the exact clipping map.

Substituting (141) into (132), (134) and (137) and absorbing the fixed chart constants into a single constant $C_{i,\mathcal{T}} \geq 1$ gives (126). This completes the proof. □

Lemma F.10 (Anchored neural approximation of the local projection coordinate)**.** *Fix an anchor $\boldsymbol{a} \in K_i^\circ$. Fix a localization parameter $0 < \varepsilon_{\text{loc}} \leq 1$ such that (89) holds for every $t \in \mathcal{I}_{\text{sm}}(\varepsilon_{\text{loc}})$. For every accuracy $\varepsilon_u \in (0, 1]$, there exists a ReLU network $\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\text{loc}}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with input $(\boldsymbol{x}, t) \in \mathbb{R}^D \times [\underline{t}, \overline{t}]$ and output in $\square_i$ such that, for every $t \in \mathcal{I}_{\text{sm}}(\varepsilon_{\text{loc}})$ and every $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon_{\text{loc}})$ satisfying*

$$\|\boldsymbol{a} - \boldsymbol{u}_i^\Pi(\boldsymbol{x},t)\|_2 \leq \frac{1}{2} r_i^{\text{GN}},$$

$$\left\|\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\rm loc}}(\boldsymbol{x},t)-\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t)\right\|_\infty \le \varepsilon_u. \tag{142}$$

Moreover, for a finite chart-dependent constant $C_{i,\Pi}\ge 1$,

$$\begin{aligned} L &\lesssim \log(e(d\vee\beta))\log^3(C_{i,\Pi}\varepsilon_u^{-1}),\\ \|\boldsymbol{W}\|_\infty &\lesssim dD(d\vee\beta)(C_{i,\Pi}\varepsilon_u^{-1})^{\frac{d}{2(\beta\vee 1)}}+D\log^3(C_{i,\Pi}\varepsilon_u^{-1}),\\ S &\lesssim dD(d+2(\beta\vee 1))^{d+3}(C_{i,\Pi}\varepsilon_u^{-1})^{\frac{d}{2(\beta\vee 1)}}\log^2(C_{i,\Pi}\varepsilon_u^{-1})+dD\log^5(C_{i,\Pi}\varepsilon_u^{-1}),\\ \log B &\lesssim \log^2(C_{i,\Pi}\varepsilon_u^{-1}). \end{aligned} \tag{143}$$

Proof. The construction uses a fixed anchor $\boldsymbol{a}$ as the initial condition and then unrolls approximate Gauss–Newton steps. Set

$$\varepsilon_{\rm step} := \min\Big\{\frac{r_i^{\rm GN}}{8\sqrt{d}},\frac{\varepsilon_u}{4\sqrt{d}}\Big\}. \tag{144}$$

By Theorem F.9, applied with localization parameter $\varepsilon_{\rm loc}$, there exists a ReLU network $\widetilde{\mathcal{T}}_{i,\varepsilon_{\rm step}}$ with output in $\square_i$ such that

$$\sup_{\boldsymbol{u}\in\square_i,\ t\in\mathcal{I}_{\rm sm}(\varepsilon_{\rm loc}),\ (\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon_{\rm loc})}\left\|\widetilde{\mathcal{T}}_{i,\varepsilon_{\rm step}}(\boldsymbol{u},\boldsymbol{x},t)-\mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t)\right\|_\infty \le \varepsilon_{\rm step}. \tag{145}$$

The choice (144) guarantees in particular

$$\varepsilon_{\rm step}\le \frac{r_i^{\rm GN}}{8\sqrt{d}},$$

which is exactly the basin condition (120) required in Theorem F.8.

Let

$$K_{\varepsilon_u} := \Big\lceil \log_2\Big(\frac{r_i^{\rm GN}}{\varepsilon_u}\Big)\Big\rceil_+,\qquad \lceil a\rceil_+ := \max\{\lceil a\rceil,0\}. \tag{146}$$

Then

$$2^{-(K_{\varepsilon_u}+1)}r_i^{\rm GN}\le\frac{\varepsilon_u}{2}. \tag{147}$$

Indeed, if $r_i^{\rm GN}\le\varepsilon_u$, then $K_{\varepsilon_u}=0$ and $r_i^{\rm GN}/2\le\varepsilon_u/2$. If $r_i^{\rm GN}>\varepsilon_u$, then $K_{\varepsilon_u}\ge\log_2(r_i^{\rm GN}/\varepsilon_u)$, so $2^{-K_{\varepsilon_u}}r_i^{\rm GN}\le\varepsilon_u$.

Step 1: Define the unrolled network. Define recursively

$$\widetilde{\boldsymbol{u}}_0(\boldsymbol{x},t):=\boldsymbol{a},\qquad \widetilde{\boldsymbol{u}}_{k+1}(\boldsymbol{x},t):=\widetilde{\mathcal{T}}_{i,\varepsilon_{\rm step}}\big(\widetilde{\boldsymbol{u}}_k(\boldsymbol{x},t),\boldsymbol{x},t\big),\qquad k=0,\dots,K_{\varepsilon_u}-1. \tag{148}$$

Finally set

$$\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\rm loc}}(\boldsymbol{x},t):=\widetilde{\boldsymbol{u}}_{K_{\varepsilon_u}}(\boldsymbol{x},t). \tag{149}$$

Because each one-step network is $\square_i$-valued, every iterate $\widetilde{\boldsymbol{u}}_k(\boldsymbol{x},t)$ also lies in $\square_i$.

Step 2: The approximation error. Assume now that $t\in\mathcal{I}_{\rm sm}(\varepsilon_{\rm loc})$, $(\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon_{\rm loc})$, and

$$\|\boldsymbol{a}-\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t)\|_2\le\frac{1}{2}r_i^{\rm GN}.$$

Then the hypotheses of Theorem F.8 are satisfied with $K=K_{\varepsilon_u}$. Hence

$$\begin{aligned}\left\|\widetilde{\boldsymbol{u}}_{K_{\varepsilon_u}}(\boldsymbol{x},t)-\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t)\right\|_2 &\le 2^{-(K_{\varepsilon_u}+1)}r_i^{\rm GN}+2\sqrt{d}\,\varepsilon_{\rm step}\\ &\le\frac{\varepsilon_u}{2}+2\sqrt{d}\cdot\frac{\varepsilon_u}{4\sqrt{d}}=\varepsilon_u.\end{aligned} \tag{150}$$

Since $\|\boldsymbol{v}\|_\infty\le\|\boldsymbol{v}\|_2$, (150) implies

$$\left\|\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\rm loc}}(\boldsymbol{x},t)-\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t)\right\|_\infty\le\varepsilon_u,$$

which proves (142).

Step 3: Network size. By Theorem F.9, the one-step network with accuracy $\varepsilon_{\text{step}}$ satisfies, for a chart-dependent constant $C_{i,\mathcal{T}} \geq 1$,

$$\begin{aligned} L_{\text{step}} &\lesssim \log(e(d \vee \beta)) \log^2(C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1}), \\ \|\boldsymbol{W}_{\text{step}}\|_\infty &\lesssim dD(d \vee \beta)(C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1})^{\frac{d}{2(\beta \vee 1)}} + D \log^3(C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1}), \\ S_{\text{step}} &\lesssim dD(d + 2(\beta \vee 1))^{d+3} (C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1})^{\frac{d}{2(\beta \vee 1)}} \log(C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1}) + dD \log^4(C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1}), \\ \log B_{\text{step}} &\lesssim \log^2(C_{i,\mathcal{T}} \varepsilon_{\text{step}}^{-1}). \end{aligned}$$

From (144), there exists a chart-dependent constant $C_{i,\Pi} \geq 1$ such that

$$\varepsilon_{\text{step}}^{-1} \leq C_{i,\Pi} \varepsilon_u^{-1}. \tag{151}$$

Also, by (146),

$$K_{\varepsilon_u} \leq C \log(C_{i,\Pi} \varepsilon_u^{-1}). \tag{152}$$

The unrolled network is the composition of K_{ε_u} copies of the one-step network, with the constant initialization $\boldsymbol{a}$ inserted by an affine layer. Therefore

$$L \leq K_{\varepsilon_u} L_{\text{step}} + \mathcal{O}(1), \qquad S \leq K_{\varepsilon_u} S_{\text{step}} + \mathcal{O}(1),$$

while the width and weight bound are controlled by the maximum of the corresponding quantities of the one-step networks. Combining these relations with (151) and (152) gives exactly (143). This completes the proof. □

Lemma F.11 (Neural approximation of the active chart coordinate)**.** *Fix a chart i and a localization parameter $0 < \varepsilon \leq 1$ such that (89) holds for every $t \in \mathcal{I}_{\text{sm}}(\varepsilon)$. For every $\varepsilon_u \in (0, 1]$, there exists a ReLU network*

$$\widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon} : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to \square_i$$

such that

$$\sup_{\substack{t \in \mathcal{I}_{\text{sm}}(\varepsilon) \\ (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)}} \left\| \widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \right\|_\infty \leq \varepsilon_u. \tag{153}$$

Moreover, for a finite chart-dependent constant $C_{i,\Pi}^{\text{coord}} \geq 1$,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3\big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big), \\ \|\boldsymbol{W}\|_\infty &\lesssim C_i^{\text{anc}} (r_i^{\text{GN}})^{-d} dD(d \vee \beta) \big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big)^{d/(2(\beta \vee 1))} + C_i^{\text{anc}} (r_i^{\text{GN}})^{-d} D \log^3\big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big), \\ S &\lesssim C_i^{\text{anc}} (r_i^{\text{GN}})^{-d} dD(d + 2(\beta \vee 1))^{d+3} \big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big)^{d/(2(\beta \vee 1))} \log^2\big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big) \\ &\qquad + C_i^{\text{anc}} (r_i^{\text{GN}})^{-d} dD \log^5\big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big), \\ \log B &\lesssim \log^2\big(C_{i,\Pi}^{\text{coord}} \varepsilon_u^{-2}\big). \end{aligned} \tag{154}$$

Proof. The construction averages projection-coordinate candidates from the fixed chart i. Let $L_{\phi,i} := (L_i^{(-)})^{-1}$ be a Lipschitz constant for ϕ_i on $\phi_i^{-1}(\square_i)$, and choose

$$\varepsilon_{y,i}^{\text{loc}} := \frac{L_i^{(-)} \varepsilon_u}{8}, \qquad \varepsilon_u^{\text{loc}} := 1 \wedge \frac{\varepsilon_{y,i}^{\text{loc}}}{\sqrt{d}\, L_i^{(+)}}.$$

For each anchor $\boldsymbol{a} \in \mathcal{A}_i^{\text{anc}}$, let $\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\text{loc}}}^{\Pi,\varepsilon}$ be the network from Theorem F.10, and set

$$\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) := \boldsymbol{z}_i\Big(\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\text{loc}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t)\Big), \qquad \mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) := \|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t)\|_2^2.$$

Approximate each objective $\mathcal{D}_{i,\boldsymbol{a}}$, uniformly over

$$t \in \mathcal{I}_{\text{sm}}(\varepsilon), \qquad (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon),$$

by a ReLU network $\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}$ with accuracy

$$\varepsilon_D := c(L_i^{(-)}\varepsilon_u)^2,$$

where $c > 0$ is a sufficiently small numerical constant fixed below. This objective network is constructed directly as follows. Approximate $\boldsymbol{\xi}(\boldsymbol{x}, t) = \boldsymbol{x}/m_t$ by Theorem I.7. Since each coordinate candidate network is $\square_i$-valued and $\boldsymbol{z}_i$ is smooth on the compact box $\square_i$, approximate $\boldsymbol{z}_i$ on $\square_i$ by Theorem I.1. Then square and sum the D coordinate differences with Theorem I.3. All variables are uniformly bounded on the displayed domain, so the internal accuracies can be chosen so that

$$|\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x}, t)| \le \varepsilon_D \tag{155}$$

for every anchor $\boldsymbol{a} \in \mathcal{A}_i^{\mathrm{anc}}$. Let

$$\widetilde{\mathcal{D}}_{\min} := \min_{\boldsymbol{a}\in\mathcal{A}_i^{\mathrm{anc}}} \widetilde{\mathcal{D}}_{i,\boldsymbol{a}}, \qquad \varepsilon_{D,i}^{\mathrm{gate}} := c(L_i^{(-)}\varepsilon_u)^2,$$

and define

$$\omega_{i,\boldsymbol{a}} := \mathsf{ReLU}\big(\varepsilon_{D,i}^{\mathrm{gate}} - (\widetilde{\mathcal{D}}_{i,\boldsymbol{a}} - \widetilde{\mathcal{D}}_{\min})\big), \qquad W_i := \sum_{\boldsymbol{a}\in\mathcal{A}_i^{\mathrm{anc}}} \omega_{i,\boldsymbol{a}}.$$

Because the minimum is attained by one of the finitely many anchors, $\varepsilon_{D,i}^{\mathrm{gate}} \le W_i \le N_i\varepsilon_{D,i}^{\mathrm{gate}}$.

We now prove the positive-weight candidate bound. Fix

$$t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon), \qquad (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon),$$

and write

$$\boldsymbol{\xi} := \boldsymbol{\xi}(\boldsymbol{x}, t), \qquad \boldsymbol{y}^{\Pi} := \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \qquad \mathcal{D}_\star := \|\boldsymbol{\xi} - \boldsymbol{y}^{\Pi}\|_2^2.$$

By Theorem F.12, $\mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta_\Pi$, so Theorem F.13 applies. Also Theorem F.4 gives $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \in K_i^\circ$. Hence Theorem F.7 provides an anchor $\boldsymbol{a}_\star \in \mathcal{A}_i^{\mathrm{anc}}$ in the Gauss–Newton basin, and Theorem F.10 gives

$$\|\boldsymbol{y}_{i,\boldsymbol{a}_\star} - \boldsymbol{y}^{\Pi}\|_2 \le \varepsilon_{y,i}^{\mathrm{loc}}.$$

Consequently, by the upper bound in (159),

$$\mathcal{D}_{i,\boldsymbol{a}_\star} - \mathcal{D}_\star \le \left(1 + \frac{\delta_\Pi}{\kappa}\right)(\varepsilon_{y,i}^{\mathrm{loc}})^2.$$

If $\omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t) > 0$, then $\widetilde{\mathcal{D}}_{i,\boldsymbol{a}} < \widetilde{\mathcal{D}}_{\min} + \varepsilon_{D,i}^{\mathrm{gate}}$. Using (155) and $\widetilde{\mathcal{D}}_{\min} \le \widetilde{\mathcal{D}}_{i,\boldsymbol{a}_\star}$, we obtain

$$\mathcal{D}_{i,\boldsymbol{a}} - \mathcal{D}_\star \le \left(1 + \frac{\delta_\Pi}{\kappa}\right)(\varepsilon_{y,i}^{\mathrm{loc}})^2 + \varepsilon_{D,i}^{\mathrm{gate}} + 2\varepsilon_D.$$

Since $\varepsilon_{y,i}^{\mathrm{loc}} = L_i^{(-)}\varepsilon_u/8$ and $\varepsilon_{D,i}^{\mathrm{gate}} = \varepsilon_D = c(L_i^{(-)}\varepsilon_u)^2$, choose $c > 0$ small enough that the right-hand side is at most $(L_i^{(-)}\varepsilon_u)^2/8$. The lower bound in (159) then shows that every positive-weight candidate satisfies

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\|_2 \le \frac{L_i^{(-)}\varepsilon_u}{2}.$$

Therefore

$$\left\|\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\mathrm{loc}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\right\|_2 \le L_{\phi,i}\,\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\|_2 \le \varepsilon_u/2$$

for all positive-weight anchors.

Define the ideal normalized coordinate output

$$\boldsymbol{u}_{\mathrm{gate}}(\boldsymbol{x}, t) := \frac{\sum_{\boldsymbol{a}\in\mathcal{A}_i^{\mathrm{anc}}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t)\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\mathrm{loc}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t)}{W_i(\boldsymbol{x}, t)}.$$

It is a convex combination of points in the box $\square_i$, hence lies in $\square_i$, and the preceding display gives $\|\boldsymbol{u}_{\text{gate}} - \boldsymbol{u}_i^{\Pi}\|_\infty \le \varepsilon_u/2$. Approximate $1/W_i$ by Theorem I.4 and multiply by the numerator coordinates using Theorem I.3, with arithmetic accuracy $\varepsilon_u/2$, obtaining a raw coordinate network $\widehat{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}$ satisfying

$$\|\widehat{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon} - \boldsymbol{u}_i^{\Pi}\|_\infty \le \varepsilon_u.$$

Finally define

$$\widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon} := \Pi_i^{(\text{box})}\big(\widehat{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}\big).$$

This clipping is the final layer, applied after the reciprocal and multiplication arithmetic. The clipping map is exactly ReLU-realizable, maps into $\square_i$, is 1-Lipschitz, and fixes $\boldsymbol{u}_i^{\Pi} \in K_i^\circ \subset \square_i$. Hence the same ε_u-error bound holds after clipping, proving (153). Since $\varepsilon_{D,i}^{\text{gate}} \asymp \varepsilon_u^2$, the reciprocal and multiplication subnetworks add only logarithmic factors, and the objective approximations dominate the non-logarithmic cost. The final clipping layer has fixed size. Summing over the N_i anchors and using (115) gives (154). □

$$\mathcal{K}_{\text{act}}(\varepsilon) := \bigcup_{i=1}^{C_{\mathcal{M}}} \mathcal{K}_i(\varepsilon). \tag{156}$$

For a fixed auxiliary radius multiplier $c_a > 0$, also write

$$\mathcal{K}_{\text{act}}^{(c_a)}(\varepsilon) := \bigcup_{i=1}^{C_{\mathcal{M}}} \mathcal{K}_i^{(c_a)}(\varepsilon). \tag{157}$$

Lemma F.12 (Active chart centers lie within the reach tube)**.** *Let $\delta_\Pi := \eta_\Pi/2$. Assume $t \in \mathcal{I}_{\text{sm}}(\varepsilon)$ and (89). If $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$ for some $i \in \{1, \dots, C_{\mathcal{M}}\}$, then*

$$\operatorname{dist}\big(\boldsymbol{\xi}(\boldsymbol{x}, t), \mathcal{M}\big) \le \delta_\Pi. \tag{158}$$

Moreover, $\delta_\Pi \le \kappa/2$.

Proof. Since $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, there exists $\boldsymbol{y} \in \mathcal{S}_i \subset \mathcal{M}$ such that

$$\|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}\|_2 \le \delta_{\text{act}}(\varepsilon, t).$$

Therefore

$$\operatorname{dist}(\boldsymbol{\xi}(\boldsymbol{x}, t), \mathcal{M}) \le \|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}\|_2 \le \delta_{\text{act}}(\varepsilon, t).$$

By (89),

$$\delta_{\text{act}}(\varepsilon, t) < \frac{1}{2}\eta_\Pi = \delta_\Pi.$$

Hence (158) holds. Since $\eta_\Pi \le \kappa$ by (88), we also have $\delta_\Pi \le \kappa/2$. □

Lemma F.13 (Projection objective gap inside the reach tube)**.** *Assume $0 < \delta \le \kappa/2$. Then, for every $\boldsymbol{\xi} \in \mathbb{R}^D$ satisfying $\operatorname{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta$ and every $\boldsymbol{y} \in \mathcal{M}$,*

$$\frac{1}{2}\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \|\boldsymbol{\xi} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \Big(1 + \frac{\delta}{\kappa}\Big)\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2. \tag{159}$$

Proof. Since $\operatorname{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta \le \kappa/2$, the reach assumption implies that $\Pi_{\mathcal{M}}(\boldsymbol{\xi})$ is unique, $(\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})) \perp T_{\Pi_{\mathcal{M}}(\boldsymbol{\xi})}\mathcal{M}$, and $\|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 \le \delta$. For every $\boldsymbol{y} \in \mathcal{M}$,

$$\begin{aligned}\|\boldsymbol{\xi} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 &= \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) + \Pi_{\mathcal{M}}(\boldsymbol{\xi}) - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \\ &= \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 - 2\langle \boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\rangle.\end{aligned} \tag{160}$$

The reach bound gives the tangent-ball inequality

$$\langle \boldsymbol{v}, \boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\rangle \le \frac{\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2}{2\kappa}, \qquad \boldsymbol{y} \in \mathcal{M}, \quad \boldsymbol{v} \in N_{\Pi_{\mathcal{M}}}(\boldsymbol{\xi})\mathcal{M}, \quad \|\boldsymbol{v}\|_2 = 1. \tag{161}$$

If $\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) = 0$, the claim follows immediately from (160). Otherwise, applying (161) to $\boldsymbol{v} = (\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}))/\|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2$ gives

$$\langle \boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) \rangle \le \frac{\|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2}{2\kappa} \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \frac{\delta}{2\kappa} \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2.$$

Applying the same inequality to $-(\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}))/\|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2$ gives

$$\langle \boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) \rangle \ge -\frac{\delta}{2\kappa} \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2.$$

Substituting these two bounds into (160) yields

$$\left(1 - \frac{\delta}{\kappa}\right) \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \|\boldsymbol{\xi} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \left(1 + \frac{\delta}{\kappa}\right) \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2.$$

Since $\delta \le \kappa/2$, the lower coefficient is at least $1/2$. This proves (159). □

Theorem F.14 is the step that avoids black-box high-order approximation of the ambient map $\Pi_{\mathcal{M}}$. The point to check is that the network searches over finitely many intrinsic anchors, performs local Gauss–Newton updates in chart coordinates, and uses an objective gap to select near-minimizing projection candidates. Thus the non-logarithmic covering cost is intrinsic in the active chart coordinates; the remaining direct ambient factors are displayed explicitly. The proof has five logically separate checks:

(i) the active-tube and reach conditions make $\Pi_{\mathcal{M}}(\boldsymbol{\xi})$ unique and place its coordinate in a buffered chart box;

(ii) a finite intrinsic anchor net contains an anchor inside the Gauss–Newton contraction basin of this coordinate;

(iii) an approximate Gauss–Newton step is realized by ReLU networks using only d-dimensional chart variables and ambient matrix–vector arithmetic;

(iv) a ReLU minimum and objective gate discard all candidates whose squared-distance objective is not within $\mathcal{O}(\varepsilon_\Pi^2)$ of the best candidate;

(v) the gated average is normalized by a reciprocal network whose denominator is bounded below by the positive gate mass.

Only the third and fourth checks carry non-logarithmic complexity; their covering cost is the intrinsic anchor count times d-dimensional chart approximation, with the remaining D dependence coming from explicit coordinatewise ambient arithmetic.

Theorem F.14 (Neural approximation of the active projection center)**.** *Fix a localization parameter* $0 < \varepsilon < 1$ *such that* (89) *holds for every* $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$. *For every* $\varepsilon_\Pi \in (0, 1]$, *there exists a ReLU network*

$$\widetilde{\Pi}_{\varepsilon_\Pi} : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to \mathbb{R}^D$$

such that

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x}, t) \in \mathcal{K}_{\mathrm{act}}(\varepsilon)}} \left\| \widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}\big(\boldsymbol{\xi}(\boldsymbol{x}, t)\big) \right\|_\infty \le \varepsilon_\Pi. \tag{162}$$

Let

$$N_{\mathrm{cand}} := \sum_{i=1}^{C_{\mathcal{M}}} |\mathcal{A}_i^{\mathrm{anc}}|.$$

Then $N_{\mathrm{cand}} \le A_{\mathcal{M}}$, *with* $A_{\mathcal{M}}$ *bounded in* (117). *For a finite geometry-dependent constant* $C_{\mathrm{proj}} \ge 1$, $\widetilde{\Pi}_{\varepsilon_\Pi} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ *with*

$$\begin{aligned}
L &\lesssim \log(e(d \vee \beta)) \log^3(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}), \\
\|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta)(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2})^{\frac{d}{2(\beta \vee 1)}} + A_{\mathcal{M}} D \log^3(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}), \\
S &\lesssim A_{\mathcal{M}} d D (d + 2(\beta \vee 1))^{d+3} (C_{\mathrm{proj}} D \varepsilon_\Pi^{-2})^{\frac{d}{2(\beta \vee 1)}} \log^2(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}) + A_{\mathcal{M}} d D \log^5(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}), \\
\log B &\lesssim \log^2(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}).
\end{aligned} \tag{163}$$

Proof. It suffices to prove the result for sufficiently small ε_Π. For larger ε_Π, use the construction at a fixed small ε_Π and absorb constants into C_{proj}.

Let

$$\mathcal{J} := \{(i, \boldsymbol{a}) : 1 \le i \le C_{\mathcal{M}}, \boldsymbol{a} \in \mathcal{A}_i^{\mathrm{anc}}\}.$$

For each $(i, \boldsymbol{a}) \in \mathcal{J}$, we construct a candidate projection center, then use a continuous objective gate to average only near-minimizing candidates.

Step 1: Existence of a good chart-anchor candidate. Fix $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $(\boldsymbol{x}, t) \in \mathcal{K}_{\mathrm{act}}(\varepsilon)$. Write

$$\boldsymbol{\xi} := \boldsymbol{\xi}(\boldsymbol{x}, t), \qquad \mathcal{D}_\star := \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2.$$

By Theorem F.12, $\mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta_\Pi$ and $\delta_\Pi \le \kappa/2$. Hence, Theorem F.13 applies with $\delta = \delta_\Pi$. For convenience write

$$C_{\mathrm{glob},\Pi}^{(+)} := \Big(1 + \frac{\delta_\Pi}{\kappa}\Big).$$

Since $(\boldsymbol{x}, t) \in \mathcal{K}_{\mathrm{act}}(\varepsilon)$, there exists an active chart $i_\star$ such that $(\boldsymbol{x}, t) \in \mathcal{K}_{i_\star}(\varepsilon)$. By Theorem F.4,

$$\Pi_{\mathcal{M}}(\boldsymbol{\xi}) \in U_{i_\star}, \qquad \boldsymbol{u}_{i_\star}^{\Pi}(\boldsymbol{x}, t) := \phi_{i_\star}(\Pi_{\mathcal{M}}(\boldsymbol{\xi})) \in K_{i_\star}^{\circ}.$$

By Theorem F.7, there exists $\boldsymbol{a}_\star \in \mathcal{A}_{i_\star}^{\mathrm{anc}}$ such that

$$\|\boldsymbol{a}_\star - \boldsymbol{u}_{i_\star}^{\Pi}(\boldsymbol{x}, t)\|_2 \le \frac{1}{2} r_{i_\star}^{\mathrm{GN}}. \tag{164}$$

Let

$$L_z^{\max} := \max_{1 \le i \le C_{\mathcal{M}}} L_i^{(+)}.$$

Choose

$$\varepsilon_y^{\mathrm{glob}} := \min\Big\{\frac{\varepsilon_\Pi}{8}, \sqrt{\frac{1}{32 C_{\mathrm{glob},\Pi}^{(+)}}}\varepsilon_\Pi\Big\}, \qquad \varepsilon_u^{\mathrm{glob}} := \frac{\varepsilon_y^{\mathrm{glob}}}{\sqrt{d}\, L_z^{\max}}. \tag{165}$$

For every $(i, \boldsymbol{a}) \in \mathcal{J}$, let $\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\mathrm{glob}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t)$ be the anchored local-coordinate network from Theorem F.10, with localization parameter ε and accuracy $\varepsilon_u^{\mathrm{glob}}$. Define the exact manifold-valued candidate

$$\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) := \boldsymbol{z}_i\Big(\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\mathrm{glob}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t)\Big). \tag{166}$$

Since the coordinate network is $\square_i$-valued and $\square_i \subset \phi_i(U_i)$, we have $\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) \in \mathcal{M}$.

For the good pair $(i_\star, \boldsymbol{a}_\star)$, (164) and Theorem F.10 give

$$\Big\|\widetilde{\boldsymbol{u}}_{i_\star,\boldsymbol{a}_\star,\varepsilon_u^{\mathrm{glob}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_{i_\star}^{\Pi}(\boldsymbol{x}, t)\Big\|_\infty \le \varepsilon_u^{\mathrm{glob}}.$$

Therefore

$$\begin{aligned}\|\boldsymbol{y}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 &\le L_{i_\star}^{(+)}\Big\|\widetilde{\boldsymbol{u}}_{i_\star,\boldsymbol{a}_\star,\varepsilon_u^{\mathrm{glob}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_{i_\star}^{\Pi}(\boldsymbol{x}, t)\Big\|_2 \\ &\le L_z^{\max}\sqrt{d}\,\varepsilon_u^{\mathrm{glob}} = \varepsilon_y^{\mathrm{glob}}.\end{aligned} \tag{167}$$

Define

$$\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) := \|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t)\|_2^2. \tag{168}$$

By the upper bound in (159),

$$\begin{aligned}\mathcal{D}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) - \mathcal{D}_\star &\le C_{\mathrm{glob},\Pi}^{(+)}\|\boldsymbol{y}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \\ &\le C_{\mathrm{glob},\Pi}^{(+)}(\varepsilon_y^{\mathrm{glob}})^2 \le \frac{1}{32}\varepsilon_\Pi^2.\end{aligned} \tag{169}$$

Step 2: Approximate candidate outputs and objectives. For each $(i, \boldsymbol{a}) \in \mathcal{J}$, construct a ReLU network $\widetilde{\boldsymbol{y}}^{\text{out}}_{i,\boldsymbol{a}}$ satisfying

$$\left\| \widetilde{\boldsymbol{y}}^{\text{out}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) \right\|_2 \le \frac{\varepsilon_\Pi}{8}. \tag{170}$$

This is obtained by composing $\widetilde{\boldsymbol{u}}^{\Pi,\varepsilon}_{i,\boldsymbol{a},\varepsilon_u^{\text{glob}}}$ with a $\mathcal{H}^{2(\beta\vee 1)}$ approximation of $\boldsymbol{z}_i$ on $\square_i$ to coordinate accuracy $\varepsilon_\Pi/(8\sqrt{D})$ using Theorem I.1.

Next, construct ReLU networks $\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}$ such that

$$\left| \widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) \right| \le \varepsilon_D, \qquad \varepsilon_D := \frac{1}{32}\varepsilon_\Pi^2. \tag{171}$$

Indeed, approximate $\boldsymbol{\xi}(\boldsymbol{x}, t)$ by Theorem I.7, approximate $\boldsymbol{z}_i$ on $\square_i$ by Theorem I.1, then use Theorem I.3 to square and sum the D coordinate differences. All variables are uniformly bounded on the active small-noise region and on the compact boxes, so coordinate and arithmetic accuracies of order ε_D/D make the total error at most ε_D. This is the source of the D factor inside the Hölder approximation term in (163).

Step 3: Continuous objective gate. Define the finite minimum

$$\widetilde{\mathcal{D}}_{\min}(\boldsymbol{x}, t) := \min_{(i,\boldsymbol{a})\in\mathcal{J}} \widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t).$$

This is exactly ReLU-realizable using

$$\min\{a, b\} = a - \mathsf{ReLU}(a - b).$$

Let

$$\varepsilon_\Pi^{\text{gate}} := \frac{1}{32}\varepsilon_\Pi^2.$$

For each $(i, \boldsymbol{a}) \in \mathcal{J}$, define

$$\omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t) := \mathsf{ReLU}\Big(\varepsilon_\Pi^{\text{gate}} - \big(\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \widetilde{\mathcal{D}}_{\min}(\boldsymbol{x}, t)\big)\Big), \tag{172}$$

and

$$W_{\text{glob}}(\boldsymbol{x}, t) := \sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t).$$

At least one index attains $\widetilde{\mathcal{D}}_{\min}$, and its weight equals $\varepsilon_\Pi^{\text{gate}}$. Hence

$$\varepsilon_\Pi^{\text{gate}} \le W_{\text{glob}}(\boldsymbol{x}, t) \le N_{\text{cand}}\varepsilon_\Pi^{\text{gate}}. \tag{173}$$

We now show that every positive-weight exact candidate is close to $\Pi_{\mathcal{M}}(\boldsymbol{\xi})$. Suppose $\omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t) > 0$. Then

$$\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) < \widetilde{\mathcal{D}}_{\min}(\boldsymbol{x}, t) + \varepsilon_\Pi^{\text{gate}}.$$

Since $(i_\star, \boldsymbol{a}_\star)$ is an available candidate,

$$\widetilde{\mathcal{D}}_{\min}(\boldsymbol{x}, t) \le \widetilde{\mathcal{D}}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) \le \mathcal{D}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) + \varepsilon_D.$$

Therefore, by (169),

$$\begin{aligned}
\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) &\le \widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) + \varepsilon_D \\
&< \widetilde{\mathcal{D}}_{\min}(\boldsymbol{x}, t) + \varepsilon_\Pi^{\text{gate}} + \varepsilon_D \\
&\le \mathcal{D}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) + \varepsilon_\Pi^{\text{gate}} + 2\varepsilon_D \\
&\le \mathcal{D}_\star + \varepsilon_\Pi^{\text{gate}} + 3\varepsilon_D.
\end{aligned} \tag{174}$$

By the definitions of $\varepsilon_\Pi^{\text{gate}}$ and ε_D,

$$\varepsilon_\Pi^{\text{gate}} + 3\varepsilon_D = \varepsilon_\Pi^2\Big(\frac{1}{32} + \frac{3}{32}\Big) = \frac{1}{2}\Big(\frac{\varepsilon_\Pi}{2}\Big)^2.$$

Thus

$$\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \mathcal{D}_\star < \frac{1}{2}\Big(\frac{\varepsilon_\Pi}{2}\Big)^2.$$

By the lower bound in (159),

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le 2(\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \mathcal{D}_\star).$$

Therefore

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 < \frac{\varepsilon_\Pi}{2} \qquad \text{whenever } \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t) > 0. \tag{175}$$

Step 4: Normalize and average. Define the ideal gated output

$$\Pi_{\text{gate}}(\boldsymbol{x},t) := \frac{\sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t)\widetilde{\boldsymbol{y}}^{\text{out}}_{i,\boldsymbol{a}}(\boldsymbol{x},t)}{W_{\text{glob}}(\boldsymbol{x},t)}.$$

The normalized weights are nonnegative and sum to one. Combining (170) and (175), for every positive-weight candidate,

$$\big\|\widetilde{\boldsymbol{y}}^{\text{out}}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\big\|_2 \le \frac{\varepsilon_\Pi}{8} + \frac{\varepsilon_\Pi}{2} = \frac{5}{8}\varepsilon_\Pi.$$

Hence

$$\|\Pi_{\text{gate}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x},t))\|_2 \le \frac{5}{8}\varepsilon_\Pi. \tag{176}$$

It remains to implement the division by W_{glob}. The scalar weights $\omega_{i,\boldsymbol{a}}$ are computed in parallel for all $(i,\boldsymbol{a}) \in \mathcal{J}$, and the sum

$$W_{\text{glob}}(\boldsymbol{x},t) = \sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t)$$

is represented exactly by one affine layer. Thus, forming W_{glob} introduces no additional approximation error. By (173), W_{glob} lies in $[\varepsilon_\Pi^{\text{gate}}, N_{\text{cand}}\varepsilon_\Pi^{\text{gate}}]$.

Apply Theorem I.4 to approximate $1/W_{\text{glob}}$ on this interval, and then use Theorem I.3 to multiply the reciprocal by each numerator coordinate

$$N_k(\boldsymbol{x},t) := \sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t)\widetilde{y}^{\text{out}}_{i,\boldsymbol{a},k}(\boldsymbol{x},t), \qquad k = 1,\dots,D.$$

Since N_{cand} is finite, all candidate outputs are uniformly bounded, and $\varepsilon_\Pi^{\text{gate}} \asymp \varepsilon_\Pi^2$, the reciprocal and multiplication accuracies can be chosen with only logarithmic overhead in ε_Π^{-1} so that the total arithmetic error is at most $3\varepsilon_\Pi/8$ in ℓ_2. This yields a ReLU network $\widetilde{\Pi}_{\varepsilon_\Pi}$ satisfying

$$\|\widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x},t) - \Pi_{\text{gate}}(\boldsymbol{x},t)\|_2 \le \frac{3}{8}\varepsilon_\Pi.$$

Combining this with (176) gives

$$\|\widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x},t))\|_2 \le \varepsilon_\Pi.$$

Since $\|\cdot\|_\infty \le \|\cdot\|_2$, this proves (162).

Step 5: Network size. The dominant cost comes from approximating the chart maps inside the objective networks to coordinate accuracy of order $\varepsilon_D/D \asymp \varepsilon_\Pi^2/D$. By Theorem I.1, this gives the intrinsic rate $(C_{\text{proj}}D\varepsilon_\Pi^{-2})^{\frac{d}{2(\beta\vee 1)}}$ up to logarithmic factors. The local coordinate networks, finite ReLU minimum, ReLU gates, affine summation layers, reciprocal network, and multiplication networks add only finite atlas factors and logarithmic factors. Indeed, $W_{\text{glob}} \ge \varepsilon_\Pi^{\text{gate}} \asymp \varepsilon_\Pi^2$, so approximating $1/W_{\text{glob}}$ only contributes logarithmic factors of order $\log(C_{\text{proj}}D\varepsilon_\Pi^{-2})$. Stacking over the N_{cand} chart-anchor candidates gives the displayed $A_{\mathcal{M}}$ factors in (163). This completes the proof.

□

F.3. Projection-centered Laplace expansion

This block derives the exact small-noise asymptotic object that the neural network will later approximate. The phase is expanded at $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t)$, the coordinate of the nearest point on the manifold. The normal residual is kept in the Gaussian factor, while the remaining smooth terms are expanded into finitely many powers of σ_t and $\boldsymbol{\nu}_i = \sigma_t^{-1}\boldsymbol{r}_i$. The final proposition in this block records the uniform remainder bound and fixes the small-noise constants used later.

On the active region, we center the phase exactly at the projection coordinate $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t)$ from Theorem F.4. For fixed $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $(\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)$, write

$$\boldsymbol{\xi} := \boldsymbol{\xi}(\boldsymbol{x},t), \qquad \Pi_{\mathcal{M}}(\boldsymbol{\xi}) := \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x},t)), \qquad \boldsymbol{u}_i^{\Pi} := \boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),$$

and

$$\boldsymbol{r}_i := \boldsymbol{r}_i(\boldsymbol{x},t) = \boldsymbol{x} - m_t\Pi_{\mathcal{M}}(\boldsymbol{\xi}), \qquad \boldsymbol{\nu}_i := \sigma_t^{-1}\boldsymbol{r}_i, \qquad \boldsymbol{G}_i^{\Pi} := \boldsymbol{G}_i(\boldsymbol{u}_i^{\Pi}).$$

Lemma F.15 (Projection-centered Gaussian phase expansion)**.** *For each $m \geq 1$, define*

$$\boldsymbol{z}_{i,m}(\boldsymbol{u},\boldsymbol{w}) := \frac{1}{m!}\mathrm{D}^m\boldsymbol{z}_i(\boldsymbol{u})[\boldsymbol{w}^{\otimes m}] \in \mathbb{R}^D.$$

Fix $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $(\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)$, and assume (89). Let $\boldsymbol{w} \in \mathbb{R}^d$ satisfy $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t) + \sigma_t\boldsymbol{w} \in \square_i$. Then

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}_i^{\Pi} + \sigma_t\boldsymbol{w})}{\sigma_t^2} = \frac{\|\boldsymbol{r}_i\|_2^2}{2\sigma_t^2} + \frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{G}_i^{\Pi}\boldsymbol{w} + \sum_{\ell=1}^{\lfloor\beta\rfloor}\sigma_t^\ell\Psi_{i,\ell}\big(\boldsymbol{u}_i^{\Pi},\boldsymbol{\nu}_i,\boldsymbol{w},t\big) + \sigma_t^{\lfloor\beta\rfloor+1}\mathrm{Rem}_i^{(F)}\big(\boldsymbol{u}_i^{\Pi},\boldsymbol{\nu}_i,\boldsymbol{w},t\big), \tag{177}$$

where for $\ell \geq 1$,

$$\Psi_{i,\ell}(\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{w},t) := -m_t\,\boldsymbol{\nu}^\top\boldsymbol{z}_{i,\ell+1}(\boldsymbol{u},\boldsymbol{w}) + \frac{m_t^2}{2}\sum_{\substack{m_1+m_2=\ell+2\\ m_1,m_2\geq 1}}\boldsymbol{z}_{i,m_1}(\boldsymbol{u},\boldsymbol{w})^\top\boldsymbol{z}_{i,m_2}(\boldsymbol{u},\boldsymbol{w}). \tag{178}$$

Moreover, for constants independent of $(\boldsymbol{x},t,\boldsymbol{w})$,

$$\begin{aligned}\big|\Psi_{i,\ell}(\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{w},t)\big| &\leq C_\ell(1+\|\boldsymbol{\nu}\|_2)(1+\|\boldsymbol{w}\|_2^{\ell+2}),\\ \big|\mathrm{Rem}_i^{(F)}(\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{w},t)\big| &\leq C(1+\|\boldsymbol{\nu}\|_2)(1+\|\boldsymbol{w}\|_2^{2\lfloor\beta\rfloor+6}).\end{aligned} \tag{179}$$

Each $\Psi_{i,\ell}$ is affine in $\boldsymbol{\nu}$.

Proof. **Step 1: Taylor expansion of the chart map.** Because $\square_i \Subset \phi_i(U_i)$ and $\boldsymbol{z}_i$ is smooth on $\phi_i(U_i)$, all derivatives of $\boldsymbol{z}_i$ up to order $\lfloor\beta\rfloor + 3$ are uniformly bounded on $\square_i$. Since both $\boldsymbol{u}_i^{\Pi}$ and $\boldsymbol{u}_i^{\Pi} + \sigma_t\boldsymbol{w}$ lie in $\square_i$, Taylor's theorem gives

$$\begin{aligned}\boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi} + \sigma_t\boldsymbol{w}) &= \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi}) + \sum_{m=1}^{\lfloor\beta\rfloor+2}\sigma_t^m\boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w}) + \sigma_t^{\lfloor\beta\rfloor+3}\mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t)\\ &= \Pi_{\mathcal{M}}(\boldsymbol{\xi}) + \sum_{m=1}^{\lfloor\beta\rfloor+2}\sigma_t^m\boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w}) + \sigma_t^{\lfloor\beta\rfloor+3}\mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t).\end{aligned}$$

The remainder satisfies

$$\|\mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t)\|_2 \leq C(1+\|\boldsymbol{w}\|_2^{\lfloor\beta\rfloor+3}). \tag{180}$$

Define

$$\Delta_i(\boldsymbol{w}) := \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi} + \sigma_t\boldsymbol{w}) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}). \tag{181}$$

Then

$$\Delta_i(\boldsymbol{w}) = \sum_{m=1}^{\lfloor\beta\rfloor+2}\sigma_t^m\boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w}) + \sigma_t^{\lfloor\beta\rfloor+3}\mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t). \tag{182}$$

Step 2: Exact algebraic expansion of the phase. By the definition of $F_{i,\boldsymbol{x},t}$,

$$\begin{aligned}2F_{i,\boldsymbol{x},t}(\boldsymbol{u}_i^{\Pi}+\sigma_t\boldsymbol{w}) &= \|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi}+\sigma_t\boldsymbol{w})\|_2^2\\ &= \|\boldsymbol{x}-m_t\Pi_{\mathcal{M}}(\boldsymbol{\xi})-m_t\Delta_i(\boldsymbol{w})\|_2^2\\ &= \|\boldsymbol{r}_i-m_t\Delta_i(\boldsymbol{w})\|_2^2\\ &= \|\boldsymbol{r}_i\|_2^2-2m_t\boldsymbol{r}_i^\top\Delta_i(\boldsymbol{w})+m_t^2\|\Delta_i(\boldsymbol{w})\|_2^2.\end{aligned}$$

Therefore

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}_i^{\Pi}+\sigma_t\boldsymbol{w})}{\sigma_t^2}=\frac{\|\boldsymbol{r}_i\|_2^2}{2\sigma_t^2}-\frac{m_t}{\sigma_t^2}\boldsymbol{r}_i^\top\Delta_i(\boldsymbol{w})+\frac{m_t^2}{2\sigma_t^2}\|\Delta_i(\boldsymbol{w})\|_2^2. \tag{183}$$

The first Taylor term is

$$\boldsymbol{z}_{i,1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})=\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})\boldsymbol{w}.$$

By (95),

$$\boldsymbol{r}_i^\top\boldsymbol{z}_{i,1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})=\boldsymbol{r}_i^\top\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})\boldsymbol{w}=\boldsymbol{w}^\top\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top\boldsymbol{r}_i=0. \tag{184}$$

Using $\boldsymbol{r}_i=\sigma_t\boldsymbol{\nu}_i$ and (182) and (184), we get

$$\begin{aligned}-\frac{m_t}{\sigma_t^2}\boldsymbol{r}_i^\top\Delta_i(\boldsymbol{w}) &= -m_t\sum_{m=2}^{\lfloor\beta\rfloor+2}\sigma_t^{m-1}\boldsymbol{\nu}_i^\top\boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})-m_t\sigma_t^{\lfloor\beta\rfloor+2}\boldsymbol{\nu}_i^\top\mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t)\\ &= -m_t\sum_{\ell=1}^{\lfloor\beta\rfloor}\sigma_t^{\ell}\boldsymbol{\nu}_i^\top\boldsymbol{z}_{i,\ell+1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})+\sigma_t^{\lfloor\beta\rfloor+1}\mathrm{Rem}_i^{(\mathrm{cr})}(\boldsymbol{u}_i^{\Pi},\boldsymbol{\nu}_i,\boldsymbol{w},t).\end{aligned} \tag{185}$$

Here the term with $m=\lfloor\beta\rfloor+2$ has order $\sigma_t^{\lfloor\beta\rfloor+1}$ and is included in the remainder, together with the Taylor remainder. Hence, using (180) and $\sigma_t\le 1$,

$$|\mathrm{Rem}_i^{(\mathrm{cr})}(\boldsymbol{u},\boldsymbol{\nu},\boldsymbol{w},t)|\le C(1+\|\boldsymbol{\nu}\|_2)(1+\|\boldsymbol{w}\|_2^{\lfloor\beta\rfloor+3}). \tag{186}$$

Step 3: Expansion of the quadratic term. Using (182),

$$\frac{m_t^2}{2\sigma_t^2}\|\Delta_i(\boldsymbol{w})\|_2^2=\frac{m_t^2}{2}\sum_{m_1,m_2=1}^{\lfloor\beta\rfloor+2}\sigma_t^{m_1+m_2-2}\boldsymbol{z}_{i,m_1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})^\top\boldsymbol{z}_{i,m_2}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})+\sigma_t^{\lfloor\beta\rfloor+1}\mathrm{Rem}_i^{(\mathrm{quad})}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t).$$

The pair $(m_1,m_2)=(1,1)$ gives

$$\frac{m_t^2}{2}\boldsymbol{z}_{i,1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})^\top\boldsymbol{z}_{i,1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})=\frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top\boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})\boldsymbol{w}=\frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{G}_i^{\Pi}\boldsymbol{w}.$$

All other principal pairs satisfy $m_1+m_2-2\ge 1$. Reindexing by $\ell=m_1+m_2-2$ and retaining powers $\ell=1,\dots,\lfloor\beta\rfloor$ gives

$$\begin{aligned}\frac{m_t^2}{2\sigma_t^2}\|\Delta_i(\boldsymbol{w})\|_2^2 &= \frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{G}_i^{\Pi}\boldsymbol{w}+\frac{m_t^2}{2}\sum_{\ell=1}^{\lfloor\beta\rfloor}\sigma_t^{\ell}\sum_{\substack{m_1+m_2=\ell+2\\ m_1,m_2\ge 1}}\boldsymbol{z}_{i,m_1}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})^\top\boldsymbol{z}_{i,m_2}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w})\\ &\quad+\sigma_t^{\lfloor\beta\rfloor+1}\mathrm{Rem}_i^{(\mathrm{quad})}(\boldsymbol{u}_i^{\Pi},\boldsymbol{w},t).\end{aligned} \tag{187}$$

The omitted Taylor pairs have total power at least $\sigma_t^{\lfloor\beta\rfloor+1}$ after division by σ_t^2, and the terms involving $\mathrm{Rem}_i^{(z)}$ have even higher powers. Therefore

$$|\mathrm{Rem}_i^{(\mathrm{quad})}(\boldsymbol{u},\boldsymbol{w},t)|\le C(1+\|\boldsymbol{w}\|_2^{2\lfloor\beta\rfloor+6}). \tag{188}$$

Step 4: Collect the coefficients. Substituting (185) and (187) into (183) gives (177), with $\Psi_{i,\ell}$ defined by (178) and

$$\mathrm{Rem}_i^{(F)}:=\mathrm{Rem}_i^{(\mathrm{cr})}+\mathrm{Rem}_i^{(\mathrm{quad})}.$$

The bounds in (179) follow from the multilinearity of $\mathrm{D}^m\boldsymbol{z}_i(\boldsymbol{u})[\boldsymbol{w}^{\otimes m}]$, the uniform derivative bounds on $\square_i$, and (186) and (188). Finally, $\Psi_{i,\ell}$ is affine in $\boldsymbol{\nu}$ because its first term is linear in $\boldsymbol{\nu}$ and its second term is independent of $\boldsymbol{\nu}$. $\square$

By the convention in Theorem 2.3, every coefficient retained in the neural approximation has strictly positive regularity $\beta - q > 0$.

Proposition F.16 (Local Laplace expansion for de-Gaussianized chart integrals)**.** *Assume the small-noise branch is chosen so that, for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$, we have $\delta_{\mathrm{act}}(\varepsilon, t) < \eta_\Pi/2$. Fix $i \in \{1, \dots, C_{\mathcal{M}}\}$ and $R_i \in \{\mathcal{Q}_i, \mathcal{P}_{i,1}, \dots, \mathcal{P}_{i,D}\}$. Then, there exist coefficient functions*

$$A^{(R)}_{i,q,\boldsymbol{\lambda}} : K_i^\circ \times [\underline{t}, \overline{t}] \to \mathbb{R}, \qquad 0 \le q \le \lfloor \beta \rfloor, \qquad \boldsymbol{\lambda} \in \mathbb{N}^D, \quad \|\boldsymbol{\lambda}\|_1 \le q,$$

such that, for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and every $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$,

$$R_i(\boldsymbol{x}, t) = \exp\Big(-\frac{\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big) \sum_{q=0}^{\lfloor \beta \rfloor} \sigma_t^q \sum_{\|\boldsymbol{\lambda}\|_1 \le q} A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), t\big) \boldsymbol{\nu}_i(\boldsymbol{x}, t)^{\boldsymbol{\lambda}} + \overline{\mathrm{Rem}}_i^{(R)}(\boldsymbol{x}, t), \tag{189}$$

where

$$\big|\overline{\mathrm{Rem}}_i^{(R)}(\boldsymbol{x}, t)\big| \le C \sigma_t^\beta \exp\Big(-c \frac{\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2^2}{\sigma_t^2}\Big). \tag{190}$$

Moreover,

$$A^{(R)}_{i,q,\boldsymbol{\lambda}}(\cdot, t) \in \mathcal{H}^{\beta - q}(K_i^\circ, C_{\mathrm{coef}}), \qquad t \in [\underline{t}, \overline{t}],$$

where C_{coef} depends only on B_a, β, $\underline{m}, \overline{m}$, and $\mathfrak{C}_{\mathcal{M}}$.

Proof. We prove the result for a generic $R_i \in \{\mathcal{Q}_i, \mathcal{P}_{i,1}, \dots, \mathcal{P}_{i,D}\}$, and write $a_i^{(R)}$ for the corresponding scalar amplitude. Fix $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, and set

$$\boldsymbol{u}^\Pi := \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), \qquad \boldsymbol{r} := \boldsymbol{r}_i(\boldsymbol{x}, t), \qquad \boldsymbol{\nu} := \boldsymbol{\nu}_i(\boldsymbol{x}, t) = \sigma_t^{-1} \boldsymbol{r}, \qquad \boldsymbol{G}^\Pi := \boldsymbol{G}_i(\boldsymbol{u}^\Pi).$$

All constants below may depend on $\underline{m}, \overline{m}, \beta$, and $\mathfrak{C}_{\mathcal{M}}$, but not on $(\boldsymbol{x}, t)$ and ε.

By definition,

$$R_i(\boldsymbol{x}, t) = \sigma_t^{-d} \int_{\phi_i(U_i)} \exp\Big(-F_{i,\boldsymbol{x},t}(\boldsymbol{u})/\sigma_t^2\Big) a_i^{(R)}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}.$$

With $\boldsymbol{u} = \boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}$ and

$$\Omega_{i,\boldsymbol{x},t} := \sigma_t^{-1}\big(\phi_i(U_i) - \boldsymbol{u}^\Pi\big),$$

the factor σ_t^{-d} cancels and

$$R_i(\boldsymbol{x}, t) = \int_{\Omega_{i,\boldsymbol{x},t}} \exp\Big(-F_{i,\boldsymbol{x},t}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w})/\sigma_t^2\Big) a_i^{(R)}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}) \mathrm{d}\boldsymbol{w}. \tag{191}$$

Step 1: Uniform quadratic lower bound for the phase. Define

$$\boldsymbol{\xi} := \boldsymbol{x}/m_t, \qquad \Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) := \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\|_2^2.$$

Since $F_{i,\boldsymbol{x},t}(\boldsymbol{u}) = m_t^2 \Phi_i(\boldsymbol{u}; \boldsymbol{\xi})$ and $\boldsymbol{z}_i(\boldsymbol{u}^\Pi) = \Pi_{\mathcal{M}}(\boldsymbol{\xi})$,

$$F_{i,\boldsymbol{x},t}(\boldsymbol{u}^\Pi) = \frac{1}{2}\|\boldsymbol{r}\|_2^2.$$

By Theorem F.4, $\boldsymbol{u}^\Pi \in K_i^\circ \Subset \square_i$ and

$$\nabla_{\boldsymbol{u}} \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) = \boldsymbol{J}_i(\boldsymbol{u}^\Pi)^\top \big(\boldsymbol{z}_i(\boldsymbol{u}^\Pi) - \boldsymbol{\xi}\big) = 0.$$

Moreover, by Theorem F.5,

$$\nabla_{\boldsymbol{u}}^2 \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) \succeq \frac{1}{2} \lambda_i^{(-)} \boldsymbol{I}_d.$$

The active condition and $\delta_{\mathrm{act}}(\varepsilon, t) < \eta_\Pi/2$ imply that $\boldsymbol{\xi} \in \mathcal{X}_i$. Recall the constants H_i and r_i from (91) and (92). Then, whenever $\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \le r_i$,

$$\nabla^2_{\boldsymbol{u}} \Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) \succeq \frac{1}{4}\lambda_i^{(-)} \boldsymbol{I}_d.$$

Taylor's theorem gives

$$\Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) - \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) \ge \frac{\lambda_i^{(-)}}{8}\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2^2.$$

Substituting $\boldsymbol{u} = \boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}$ yields, for $\sigma_t \|\boldsymbol{w}\|_2 \le r_i$,

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w})}{\sigma_t^2} \ge \frac{\|\boldsymbol{r}\|_2^2}{2\sigma_t^2} + c_0 \|\boldsymbol{w}\|_2^2 \tag{192}$$

for some $c_0 > 0$.

Next let $S_i^{(R)} := \operatorname{supp} a_i^{(R)} \Subset \phi_i(U_i)$ and consider

$$\mathcal{X}_i^\circ := \{\boldsymbol{\zeta} \in \mathcal{X}_i : \Pi_{\mathcal{M}}(\boldsymbol{\zeta}) \in \phi_i^{-1}(K_i^\circ)\}.$$

This set is compact because $\mathcal{X}_i$ is compact, the reach projection is continuous on the reach tube, and $\phi_i^{-1}(K_i^\circ)$ is compact in U_i. On $\mathcal{X}_i^\circ$, write

$$\boldsymbol{u}_i^\Pi(\boldsymbol{\zeta}) := \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{\zeta})).$$

Define

$$\mathcal{C}_i^{(R)} := \Big\{(\boldsymbol{\zeta}, \boldsymbol{u}) \in \mathcal{X}_i^\circ \times S_i^{(R)} : \|\boldsymbol{u} - \boldsymbol{u}_i^\Pi(\boldsymbol{\zeta})\|_2 \ge r_i\Big\}.$$

If this compact set is nonempty, the continuous function

$$(\boldsymbol{\zeta}, \boldsymbol{u}) \mapsto \Phi_i(\boldsymbol{u}; \boldsymbol{\zeta}) - \Phi_i\big(\boldsymbol{u}_i^\Pi(\boldsymbol{\zeta}); \boldsymbol{\zeta}\big)$$

is strictly positive on it. Otherwise $\boldsymbol{z}_i(\boldsymbol{u})$ would be another nearest point on $\mathcal{M}$ to $\boldsymbol{\zeta}$, contradicting uniqueness of the metric projection in the reach tube. Hence there is $c_1 > 0$ such that, for active $\boldsymbol{\xi}$,

$$\Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) - \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) \ge c_1$$

whenever $\boldsymbol{u} \in S_i^{(R)}$ and $\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \ge r_i$. Therefore, on this part of the support,

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2} \ge \frac{\|\boldsymbol{r}\|_2^2}{2\sigma_t^2} + \frac{\underline{m}^2 c_1}{\sigma_t^2}. \tag{193}$$

Step 2: Split the integral into a local window and a tail. Let

$$R_\sigma := A\sqrt{\log(e/\sigma_t)},$$

where $A > 0$ will be chosen large enough. By decreasing the small-noise threshold if necessary, assume

$$\sigma_t R_\sigma \le r_i \le d_i^\square/2.$$

Since $\boldsymbol{u}^\Pi \in K_i^\circ$, this implies

$$\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w} \in \square_i \subset \phi_i(U_i), \qquad \|\boldsymbol{w}\|_2 \le R_\sigma.$$

Thus the local part of (191) is over the full ball $\{\|\boldsymbol{w}\|_2 \le R_\sigma\}$.

Write $R_i = I_{\mathrm{loc}} + I_{\mathrm{tail}}$, where I_{loc} is the integral over $\|\boldsymbol{w}\|_2 \le R_\sigma$. On $R_\sigma < \|\boldsymbol{w}\|_2 \le r_i/\sigma_t$, use (192); on $\|\boldsymbol{w}\|_2 > r_i/\sigma_t$, use (193) on the support of $a_i^{(R)}$. Since $a_i^{(R)}$ is bounded and $S_i^{(R)}$ has finite Lebesgue measure,

$$\begin{aligned}|I_{\mathrm{tail}}| &\le C e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} \int_{\|\boldsymbol{w}\|_2 > R_\sigma} e^{-c_0\|\boldsymbol{w}\|_2^2} \mathrm{d}\boldsymbol{w} + C\sigma_t^{-d} e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-c/\sigma_t^2} \\ &\le C e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-c_2 R_\sigma^2} + C\sigma_t^{-d} e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-c/\sigma_t^2}.\end{aligned} \tag{194}$$

Choose A so large that $e^{-c_2 R_\sigma^2} \le C\sigma_t^\beta$. Also, after decreasing the small-noise threshold if necessary, $\sigma_t^{-d} e^{-c/\sigma_t^2} \le C\sigma_t^\beta$. Hence

$$|I_{\text{tail}}| \le C\sigma_t^\beta \exp\Big(-c\frac{\|\boldsymbol{r}\|_2^2}{\sigma_t^2}\Big). \tag{195}$$

Step 3: Expand the phase on the local window. On $\|\boldsymbol{w}\|_2 \le R_\sigma$, Theorem F.15 and the $\lfloor\beta\rfloor$-convention give the truncated form

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w})}{\sigma_t^2} = \frac{\|\boldsymbol{r}\|_2^2}{2\sigma_t^2} + \frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}^\Pi \boldsymbol{w} + \sum_{\ell=1}^{\lfloor\beta\rfloor} \sigma_t^\ell \Psi_{i,\ell}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \widehat{\mathrm{Rem}}_i^{(F)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t). \tag{196}$$

Indeed, the remainder from Theorem F.15 has order $\sigma_t^{\lfloor\beta\rfloor+1}$. Since $\lfloor\beta\rfloor < \beta \le \lfloor\beta\rfloor + 1$ and $\sigma_t \le 1$, this is bounded by σ_t^β times the same polynomial; when $\beta \in \mathbb{N}_+$, the two powers coincide. Thus

$$|\widehat{\mathrm{Rem}}_i^{(F)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)| \le C(1 + \|\boldsymbol{\nu}\|_2)(1 + \|\boldsymbol{w}\|_2^{2\lfloor\beta\rfloor+6})$$

on the local window. Set

$$Z_\sigma := -\sum_{\ell=1}^{\lfloor\beta\rfloor} \sigma_t^\ell \Psi_{i,\ell}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) - \sigma_t^\beta \widehat{\mathrm{Rem}}_i^{(F)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t).$$

Then

$$\exp\Big(-F_{i,\boldsymbol{x},t}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w})/\sigma_t^2\Big) = e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}^\Pi \boldsymbol{w}} e^{Z_\sigma}.$$

Since $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, $\|\boldsymbol{\nu}\|_2 \le c_\star\sqrt{\log \varepsilon^{-1}}$. On $\mathcal{I}_{\text{sm}}(\varepsilon)$, $\log \varepsilon^{-1} \lesssim \log(e/\sigma_t)$. Combining this with $\|\boldsymbol{w}\|_2 \le R_\sigma$ and (179), we obtain $|Z_\sigma| \le 1/2$ after decreasing the small-noise threshold if necessary.

Step 4: Expand the exponential correction. Taylor expansion of e^{Z_σ}, collecting powers of σ_t up to order $\lfloor\beta\rfloor$, gives

$$e^{Z_\sigma} = \sum_{\ell=0}^{\lfloor\beta\rfloor} \sigma_t^\ell \mathcal{E}_{i,\ell}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \mathrm{Rem}_i^{(\exp)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t),$$

where $\mathcal{E}_{i,0} = 1$, and for $1 \le \ell \le \lfloor\beta\rfloor$,

$$\mathcal{E}_{i,\ell}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) := \sum_{r=1}^{\ell} \frac{(-1)^r}{r!} \sum_{\substack{\ell_1+\cdots+\ell_r=\ell \\ \ell_j \ge 1}} \Psi_{i,\ell_1}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) \cdots \Psi_{i,\ell_r}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t).$$

Moreover, on the local window,

$$|\mathrm{Rem}_i^{(\exp)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)| \le C(1 + \|\boldsymbol{\nu}\|_2^{P_E})(1 + \|\boldsymbol{w}\|_2^{M_E})$$

for finite constants P_E, M_E.

Step 5: Expand the amplitude and multiply expansions. Since $a_i^{(R)} \in \mathcal{H}^\beta(\phi_i(U_i), B_a)$, Taylor's theorem with the $\mathcal{H}^\beta$ remainder gives

$$a_i^{(R)}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}) = \sum_{m=0}^{\lfloor\beta\rfloor} \sigma_t^m \mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{w}) + \sigma_t^\beta \mathrm{Rem}_i^{(\text{amp})}(\boldsymbol{u}^\Pi, \boldsymbol{w}),$$

where

$$\mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}, \boldsymbol{w}) := \frac{1}{m!} \mathrm{D}^m a_i^{(R)}(\boldsymbol{u})[\boldsymbol{w}^{\otimes m}],$$

and

$$|\mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}, \boldsymbol{w})| \le C(1 + \|\boldsymbol{w}\|_2^m), \qquad |\mathrm{Rem}_i^{(\text{amp})}(\boldsymbol{u}, \boldsymbol{w})| \le C(1 + \|\boldsymbol{w}\|_2^\beta).$$

This is Taylor's theorem through order $\lfloor\beta\rfloor$ with the $\mathcal{H}^\beta$ remainder. When $\beta \in \mathbb{N}_+$, the convention $\lfloor\beta\rfloor = \beta - 1$ makes this the standard Lipschitz top-derivative remainder of order β. Multiplying gives

$$e^{Z_\sigma} a_i^{(R)}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}) = \sum_{q=0}^{\lfloor\beta\rfloor} \sigma_t^q \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \mathrm{Rem}_i^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t),$$

where

$$\mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) := \sum_{\substack{m+\ell=q \\ 0\le m,\ell\le\lfloor\beta\rfloor}} \mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}, \boldsymbol{w}) \mathcal{E}_{i,\ell}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t).$$

All omitted products contain either an explicit amplitude or exponential remainder, or have total algebraic order at least $\lfloor\beta\rfloor + 1$. Since $\lfloor\beta\rfloor + 1 \ge \beta$, every omitted product is bounded by σ_t^β times a fixed polynomial in $(\boldsymbol{\nu}, \boldsymbol{w})$ on $0 < \sigma_t \le 1$. Furthermore,

$$|\mathrm{Rem}_i^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)| \le C(1 + \|\boldsymbol{\nu}\|_2^{P_R})(1 + \|\boldsymbol{w}\|_2^{M_R})$$

on $\|\boldsymbol{w}\|_2 \le R_\sigma$.

Step 6: Identify the principal coefficients. Substituting the expansions into I_{loc} yields

$$I_{\mathrm{loc}} = e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} \sum_{q=0}^{\lfloor\beta\rfloor} \sigma_t^q \int_{\|\boldsymbol{w}\|_2 \le R_\sigma} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}^\Pi \boldsymbol{w}} \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) \mathrm{d}\boldsymbol{w} + E_{\mathrm{loc}}.$$

By the polynomial bound on $\mathrm{Rem}_i^{(R)}$ and uniform ellipticity of $\boldsymbol{G}^\Pi$,

$$|E_{\mathrm{loc}}| \le C\sigma_t^\beta e^{-\|\boldsymbol{\nu}\|_2^2/2}(1 + \|\boldsymbol{\nu}\|_2^{P_R}) \int_{\mathbb{R}^d} e^{-c_g\|\boldsymbol{w}\|_2^2}(1 + \|\boldsymbol{w}\|_2^{M_R}) \mathrm{d}\boldsymbol{w} \le C\sigma_t^\beta e^{-c\|\boldsymbol{\nu}\|_2^2}.$$

For $0 \le q \le \lfloor\beta\rfloor$, define the coefficient functions by collecting powers of $\boldsymbol{\nu}$ in

$$\int_{\mathbb{R}^d} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u})\boldsymbol{w}} \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) \mathrm{d}\boldsymbol{w} = \sum_{\|\boldsymbol{\lambda}\|_1 \le q} A_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}, t) \boldsymbol{\nu}^{\boldsymbol{\lambda}}.$$

This is valid because $\mathcal{U}_{i,q}^{(R)}$ is a polynomial in $\boldsymbol{\nu}$ of degree at most q.

The replacement of the local integral by the full integral costs only a tail error. Indeed,

$$|\mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)| \le C(1 + \|\boldsymbol{\nu}\|_2^q)(1 + \|\boldsymbol{w}\|_2^{M_U}),$$

so, after increasing A if necessary,

$$\begin{aligned} & e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} \sigma_t^q \left| \int_{\|\boldsymbol{w}\|_2 > R_\sigma} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}^\Pi \boldsymbol{w}} \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) \mathrm{d}\boldsymbol{w} \right| \\ & \le Ce^{-c\|\boldsymbol{\nu}\|_2^2} e^{-c_3 R_\sigma^2/2} \le C\sigma_t^\beta e^{-c\|\boldsymbol{\nu}\|_2^2}. \end{aligned}$$

Combining this extension error with the bounds for E_{loc} and I_{tail} proves (189) and (190), because $\|\boldsymbol{\nu}\|_2^2 = \|\boldsymbol{r}\|_2^2/\sigma_t^2$.

Step 7: Hölder regularity of the coefficients. Each $A_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}, t)$ is a finite linear combination of integrals of terms of the form

$$e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u})\boldsymbol{w}} \times \text{polynomial in } \boldsymbol{w} \times \text{smooth geometric factors in } \boldsymbol{u} \times \partial^\alpha a_i^{(R)}(\boldsymbol{u}), \qquad |\alpha| \le q.$$

The tensor $\boldsymbol{G}_i(\boldsymbol{u})$ is smooth and uniformly positive definite on K_i°, so derivatives falling on the Gaussian factor are dominated by $e^{-c\|\boldsymbol{w}\|_2^2}$ times a polynomial in $\|\boldsymbol{w}\|_2$, uniformly in $\boldsymbol{u} \in K_i^\circ$ and $t \in [\underline{t}, \overline{t}]$. Hence differentiation under the integral is justified up to the order allowed by the amplitude. Since $a_i^{(R)} \in \mathcal{H}^\beta(\phi_i(U_i), B_a)$, the worst regularity occurs when $|\alpha| = q$, giving $\partial^\alpha a_i^{(R)} \in \mathcal{H}^{\beta-q}$. Multiplication by smooth geometric factors and integration against the uniformly dominated Gaussian preserve this regularity. Therefore

$$A_{i,q,\boldsymbol{\lambda}}^{(R)}(\cdot, t) \in \mathcal{H}^{\beta-q}(K_i^\circ, C_{\mathrm{coef}}), \qquad t \in [\underline{t}, \overline{t}].$$

This completes the proof. □

Choice of noise-regime thresholds. All phrases above of the form "after decreasing the small-noise threshold if necessary" are collected in the single threshold $\bar{\sigma}_{\mathrm{Lap}}$ below. The branch constants C_{sm} and c_{lg} are chosen only after this collection is complete, so the small/large-noise cover and its overlap are not affected by the intermediate reductions. Since the atlas is finite, choose $C_{\mathrm{Lap}}^{\max} \geq 1$ to dominate the constants in the Laplace remainder bound (190), uniformly over all charts i and all $R \in \mathcal{R}$. Also choose $\bar{\sigma}_{\mathrm{Lap}} \in (0,1]$ small enough that every small-noise requirement invoked in the proof of Theorem F.16 holds whenever $0 < \sigma_t \leq \bar{\sigma}_{\mathrm{Lap}}$: in particular $\sigma_t R_\sigma \leq r_i$, the compact-away tail satisfies $\sigma_t^{-d} e^{-c/\sigma_t^2} \lesssim \sigma_t^{\beta}$, and the exponential correction obeys $|Z_\sigma| \leq 1/2$, uniformly over all charts and integrands. Such a number exists because the atlas and the retained coefficients are finite and because $\sigma^{-\ell} e^{-c/\sigma^2} \to 0$ and $\sigma^{\gamma}(\log(e/\sigma))^M \to 0$ as $\sigma \downarrow 0$ for every fixed $\ell, \gamma, M > 0$. From now on, the constants C_{sm} and c_{lg} in Theorem D.1 are fixed so that

$$C_{\mathrm{sm}} \leq \bar{\sigma}_{\mathrm{Lap}}, \qquad C_{\mathrm{Lap}}^{\max} C_{\mathrm{sm}}^{\beta} \leq \frac{1}{6}, \qquad 4c_{\mathrm{lg}} \leq C_{\mathrm{sm}}. \tag{197}$$

Thus, on $\mathcal{I}_{\mathrm{sm}}(\varepsilon)$, the analytic Laplace remainder is bounded by the target order $\mathcal{O}(\varepsilon)$, while $\mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $\mathcal{I}_{\mathrm{lg}}(\varepsilon)$ still have the overlap needed for branch interpolation.

F.4. Network assembly and final small-noise theorem

After the projection-coordinate networks are available, the remaining neural operations are algebraic or standard approximation tasks: residuals and normalized normals, Gaussian factors, Laplace coefficients, certified active-chart gates, and the final stable ratio. This section assembles those pieces into the small-noise score network.

F.4.1. Residual, normal-coordinate, and Gaussian-factor networks

After the projection network is available, the remaining geometric quantities are algebraic. The residual is $\boldsymbol{r}_i = \boldsymbol{x} - m_t \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$, the normalized residual is $\boldsymbol{\nu}_i = \sigma_t^{-1}\boldsymbol{r}_i$, and the Gaussian factor is $\exp(-\|\boldsymbol{\nu}_i\|_2^2/2)$. The following lemmas propagate the projection error through these operations and record the resulting network sizes.

Lemma F.17 (ReLU approximation of projection residuals and normalized normals)**.** *Fix $i \in \{1, \dots, C_{\mathcal{M}}\}$. Fix a localization parameter $0 < \varepsilon \leq 1$ such that (89) holds for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$. For $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, define*

$$\boldsymbol{\pi}(\boldsymbol{x}, t) := \Pi_{\mathcal{M}}\big(\boldsymbol{\xi}(\boldsymbol{x}, t)\big), \qquad \boldsymbol{r}_i(\boldsymbol{x}, t) := \boldsymbol{x} - m_t \boldsymbol{\pi}(\boldsymbol{x}, t), \qquad \boldsymbol{\nu}_i(\boldsymbol{x}, t) := \sigma_t^{-1}\boldsymbol{r}_i(\boldsymbol{x}, t).$$

Let

$$C_{\nu,\varepsilon} := 2\sigma_{\underline{t}}^{-1} + c_\star \sqrt{\log(e/\varepsilon)}.$$

For every $\varepsilon_{\mathrm{res}} \in (0, 1]$, define

$$C_{\mathrm{res}} := e \vee \big(4096\, C_{\mathrm{proj}} C_{\nu,\varepsilon}^2\big),$$

where C_{proj} is the geometry-dependent constant from Theorem F.14. Then there exist ReLU networks

$$\widetilde{\boldsymbol{r}}_{i,\varepsilon_{\mathrm{res}}}, \widetilde{\boldsymbol{\nu}}_{i,\varepsilon_{\mathrm{res}}} : \mathbb{R}^D \times [\underline{t}, \bar{t}] \to \mathbb{R}^D$$

such that

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)}} \big\| \widetilde{\boldsymbol{r}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x}, t) - \boldsymbol{r}_i(\boldsymbol{x}, t) \big\|_\infty \leq \varepsilon_{\mathrm{res}},$$

and

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)}} \big\| \widetilde{\boldsymbol{\nu}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x}, t) - \boldsymbol{\nu}_i(\boldsymbol{x}, t) \big\|_\infty \leq \varepsilon_{\mathrm{res}}.$$

Moreover,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big), \\ \|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta) \big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big)^{d/(2(\beta \vee 1))} + A_{\mathcal{M}} D \log^3\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big), \\ S &\lesssim A_{\mathcal{M}} d D (d + 2(\beta \vee 1))^{d+3} \big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big)^{d/(2(\beta \vee 1))} \log^2\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big) + A_{\mathcal{M}} d D \log^5\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big), \\ \log B &\lesssim \log^2\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big). \end{aligned} \tag{198}$$

Proof. **Step 1: Active-region bounds.** We work uniformly over $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$. Since $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, there exists $\boldsymbol{y} \in \mathcal{S}_i \subset \mathcal{M}$ such that

$$\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2 \le c_\star \sigma_t \sqrt{\log(e/\varepsilon)}.$$

Since $\boldsymbol{\pi}(\boldsymbol{x}, t)$ is the closest point on $\mathcal{M}$ to $\boldsymbol{\xi}(\boldsymbol{x}, t) = \boldsymbol{x}/m_t$,

$$\begin{aligned}\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2 &= m_t \mathrm{dist}\big(\boldsymbol{\xi}(\boldsymbol{x}, t), \mathcal{M}\big) \\ &\le m_t \|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}\|_2 = \|\boldsymbol{x} - m_t \boldsymbol{y}\|_2 \\ &\le c_\star \sigma_t \sqrt{\log(e/\varepsilon)}. \end{aligned} \tag{199}$$

Thus

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 = \sigma_t^{-1} \|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2 \le c_\star \sqrt{\log(e/\varepsilon)} \le C_{\nu,\varepsilon}.$$

Also, because $\sigma_t \le 1$,

$$|r_{i,j}(\boldsymbol{x}, t)| \le c_\star \sqrt{\log(e/\varepsilon)} \le C_{\nu,\varepsilon}.$$

Step 2: Approximate the projection center and residual. Set

$$\varepsilon_r := \frac{\varepsilon_{\mathrm{res}}}{8 C_{\nu,\varepsilon}}, \qquad \varepsilon_{\mathrm{cen}} := \varepsilon_r/8.$$

By Theorem F.14, applied with localization parameter ε and accuracy $\varepsilon_{\mathrm{cen}}$, there is a ReLU network $\widetilde{\boldsymbol{\pi}}_{\varepsilon_{\mathrm{cen}}}$ satisfying

$$\big\|\widetilde{\boldsymbol{\pi}}_{\varepsilon_{\mathrm{cen}}}(\boldsymbol{x}, t) - \boldsymbol{\pi}(\boldsymbol{x}, t)\big\|_\infty \le \varepsilon_{\mathrm{cen}}.$$

In Theorem F.14, we use the smoothness $2(\beta \vee 1)$ of the geometric chart maps, hence its non-logarithmic factor is controlled by

$$\big(C_{\mathrm{proj}} D \varepsilon_{\mathrm{cen}}^{-2}\big)^{d/(2(\beta \vee 1))}.$$

Since

$$\varepsilon_{\mathrm{cen}}^{-2} = 64 \varepsilon_r^{-2} = 4096 C_{\nu,\varepsilon}^2 \varepsilon_{\mathrm{res}}^{-2},$$

we have

$$C_{\mathrm{proj}} D \varepsilon_{\mathrm{cen}}^{-2} \le C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}.$$

By Theorem I.2, let $\widetilde{m}(t)$ approximate m_t with

$$|\widetilde{m}(t) - m_t| \le \varepsilon_m, \qquad \varepsilon_m := \varepsilon_r/8.$$

For each $j = 1, \dots, D$, use Theorem I.3 to multiply $\widetilde{m}(t)$ and $\widetilde{\pi}_{\varepsilon_{\mathrm{cen}},j}(\boldsymbol{x}, t)$ with multiplication accuracy $\varepsilon_{M,r} := \varepsilon_r/2$. Since $m_t \in [0, 1]$ and $\boldsymbol{\pi}(\boldsymbol{x}, t) \in \mathcal{M} \subset [0, 1]^D$, the true inputs lie in $[-1, 1]^2$, and the two input errors are at most $\varepsilon_r/8$. Therefore

$$|\widetilde{h}_j(\boldsymbol{x}, t) - m_t \pi_j(\boldsymbol{x}, t)| \le \varepsilon_{M,r} + 2 \cdot 1 \cdot \frac{\varepsilon_r}{8} = \frac{3}{4}\varepsilon_r \le \varepsilon_r.$$

Define

$$\widetilde{r}_{i,j}(\boldsymbol{x}, t) := x_j - \widetilde{h}_j(\boldsymbol{x}, t).$$

Then

$$|\widetilde{r}_{i,j}(\boldsymbol{x}, t) - r_{i,j}(\boldsymbol{x}, t)| \le \varepsilon_r,$$

and therefore

$$\big\|\widetilde{\boldsymbol{r}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x}, t) - \boldsymbol{r}_i(\boldsymbol{x}, t)\big\|_\infty \le \varepsilon_r \le \varepsilon_{\mathrm{res}}.$$

Step 3: Approximate the normalized residual. Let $\phi_{1/\sigma}(t)$ approximate σ_t^{-1} with

$$|\phi_{1/\sigma}(t) - \sigma_t^{-1}| \le \varepsilon_\sigma, \qquad \varepsilon_\sigma := \frac{\varepsilon_{\mathrm{res}}}{8 C_{\nu,\varepsilon}}.$$

For each j, use Theorem I.3 to multiply $\phi_{1/\sigma}(t)$ and $\widetilde{r}_{i,j}(\boldsymbol{x},t)$ with multiplication accuracy $\varepsilon_{M,\nu} := \varepsilon_{\mathrm{res}}/2$. The true inputs lie in $[-C_{\nu,\varepsilon}, C_{\nu,\varepsilon}]^2$, and the two input errors are at most $\varepsilon_{\mathrm{res}}/(8C_{\nu,\varepsilon})$. Hence

$$|\widetilde{\nu}_{i,j}(\boldsymbol{x},t) - \sigma_t^{-1} r_{i,j}(\boldsymbol{x},t)| \le \varepsilon_{M,\nu} + 2C_{\nu,\varepsilon}\frac{\varepsilon_{\mathrm{res}}}{8C_{\nu,\varepsilon}} = \frac{3}{4}\varepsilon_{\mathrm{res}} \le \varepsilon_{\mathrm{res}}.$$

Thus

$$\left\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x},t) - \boldsymbol{\nu}_i(\boldsymbol{x},t)\right\|_\infty \le \varepsilon_{\mathrm{res}}.$$

Step 4: Size. The construction consists of the projection-center network, the schedule networks for m_t and σ_t^{-1}, D product subnetworks for $m_t\boldsymbol{\pi}$, D product subnetworks for $\sigma_t^{-1}\boldsymbol{r}_i$, and affine subtraction layers. The center network contributes the bound from (163) with $C_{\mathrm{proj}}D\varepsilon_{\mathrm{cen}}^{-2} \le C_{\mathrm{res}}D\varepsilon_{\mathrm{res}}^{-2}$. The remaining components contribute only logarithmic powers in $C_{\nu,\varepsilon}$, $\varepsilon_{\mathrm{res}}^{-1}$, and $\sigma_{\underline{t}}^{-1}$, all bounded by powers of $\log(C_{\mathrm{res}}D\varepsilon_{\mathrm{res}}^{-2})$. Summing the components gives (198). □

Lemma F.18 (ReLU approximation of the projection-center Gaussian factor)**.** *Fix $i \in \{1,\dots,C_{\mathcal{M}}\}$. Fix a localization parameter $0 < \varepsilon \le 1$ such that (89) holds for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$. Define*

$$\mathcal{G}_i(\boldsymbol{x},t) := \exp\Big(-\frac{\|\boldsymbol{r}_i(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\Big) = \exp\Big(-\frac{1}{2}\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2^2\Big).$$

Let $C_{\nu,\varepsilon}$ and C_{res} be as in Theorem F.17. For every $\varepsilon_G \in (0,1]$, define

$$\varepsilon_\nu := \frac{\varepsilon_G}{12D^{1/2}C_{\nu,\varepsilon}}, \qquad C_{\mathrm{gau}} := e \vee \Big(144D^2 C_{\mathrm{res}} C_{\nu,\varepsilon}^2\Big).$$

Then there exists a ReLU network $\widetilde{\mathcal{G}}_{i,\varepsilon_G} : \mathbb{R}^D \times [\underline{t},\overline{t}] \to \mathbb{R}$ such that

$$\sup_{\substack{t\in\mathcal{I}_{\mathrm{sm}}(\varepsilon)\\(\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon)}} \left|\widetilde{\mathcal{G}}_{i,\varepsilon_G}(\boldsymbol{x},t) - \mathcal{G}_i(\boldsymbol{x},t)\right| \le \varepsilon_G.$$

Moreover,

$$\begin{aligned} L &\lesssim \log(e(d\vee\beta))\log^3(C_{\mathrm{gau}}\varepsilon_G^{-2}),\\ \|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}}dD(d\vee\beta)\big(C_{\mathrm{gau}}\varepsilon_G^{-2}\big)^{d/(2(\beta\vee1))} + A_{\mathcal{M}}D\log^3(C_{\mathrm{gau}}\varepsilon_G^{-2}),\\ S &\lesssim A_{\mathcal{M}}dD(d+2(\beta\vee1))^{d+3}\big(C_{\mathrm{gau}}\varepsilon_G^{-2}\big)^{d/(2(\beta\vee1))}\log^2(C_{\mathrm{gau}}\varepsilon_G^{-2}) + A_{\mathcal{M}}dD\log^5(C_{\mathrm{gau}}\varepsilon_G^{-2}),\\ \log B &\lesssim \log^2(C_{\mathrm{gau}}\varepsilon_G^{-2}). \end{aligned} \tag{200}$$

Proof. **Step 1: Approximate $\boldsymbol{\nu}_i$.** Set

$$s_i(\boldsymbol{x},t) := \frac{1}{2}\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2^2, \qquad \mathcal{G}_i(\boldsymbol{x},t) = e^{-s_i(\boldsymbol{x},t)}.$$

By Theorem F.17, applied with accuracy ε_ν, there exists $\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu}$ such that

$$\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu}(\boldsymbol{x},t) - \boldsymbol{\nu}_i(\boldsymbol{x},t)\|_\infty \le \varepsilon_\nu.$$

By (199),

$$\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le c_\star\sqrt{\log(e/\varepsilon)} \le C_{\nu,\varepsilon}.$$

Step 2: Approximate the exponent. Let

$$\varepsilon_{\mathrm{sq}} := \frac{\varepsilon_G}{3D}.$$

For each $\ell = 1,\dots,D$, apply Theorem I.3 to square $\widetilde{\nu}_{i,\varepsilon_\nu,\ell}$. Thus

$$|\widetilde{q}_\ell - \widetilde{\nu}_{i,\varepsilon_\nu,\ell}^2| \le \varepsilon_{\mathrm{sq}}.$$

Define

$$\widetilde{s}_i(\boldsymbol{x},t) := \frac{1}{2}\sum_{\ell=1}^{D} \widetilde{q}_\ell(\boldsymbol{x},t).$$

The active residual bound is in ℓ_2, not merely coordinatewise. Hence

$$\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu}(\boldsymbol{x},t) - \boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le D^{1/2}\varepsilon_\nu \le \frac{\varepsilon_G}{12C_{\nu,\varepsilon}}.$$

Using

$$\left|\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu}\|_2^2 - \|\boldsymbol{\nu}_i\|_2^2\right| \le \|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu} - \boldsymbol{\nu}_i\|_2 \left(2\|\boldsymbol{\nu}_i\|_2 + \|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu} - \boldsymbol{\nu}_i\|_2\right),$$

and $C_{\nu,\varepsilon} \ge 1$, the perturbation error in the exact squared norm is at most $\varepsilon_G/4$. Then

$$|\widetilde{s}_i(\boldsymbol{x},t) - s_i(\boldsymbol{x},t)| \le \frac{D}{2}\varepsilon_{\mathrm{sq}} + \frac{\varepsilon_G}{4} \le \frac{5\varepsilon_G}{12}.$$

Step 3: Approximate the exponential. Set

$$\widetilde{s}_i^{+}(\boldsymbol{x},t) := \mathsf{ReLU}(\widetilde{s}_i(\boldsymbol{x},t)).$$

Since $s_i \ge 0$ and ReLU is 1-Lipschitz,

$$|\widetilde{s}_i^{+}(\boldsymbol{x},t) - s_i(\boldsymbol{x},t)| \le \frac{5\varepsilon_G}{12}.$$

Apply Theorem I.5 with accuracy $\varepsilon_{\exp} := \varepsilon_G/3$ and define

$$\widetilde{\mathcal{G}}_{i,\varepsilon_G}(\boldsymbol{x},t) := \phi_{\exp}(\widetilde{s}_i^{+}(\boldsymbol{x},t)).$$

Then

$$|\widetilde{\mathcal{G}}_{i,\varepsilon_G}(\boldsymbol{x},t) - \mathcal{G}_i(\boldsymbol{x},t)| \le \varepsilon_{\exp} + |\widetilde{s}_i^{+}(\boldsymbol{x},t) - s_i(\boldsymbol{x},t)| \le \varepsilon_G.$$

Step 4: Network size. Since

$$C_{\mathrm{res}} D \varepsilon_\nu^{-2} = 144 D^2 C_{\mathrm{res}} C_{\nu,\varepsilon}^2 \varepsilon_G^{-2} \le C_{\mathrm{gau}} \varepsilon_G^{-2},$$

the normal-residual network has the size from (198) with $C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}$ replaced by $C_{\mathrm{gau}}\varepsilon_G^{-2}$. The square subnetworks and the exponential network add only logarithmic powers in $C_{\mathrm{gau}}\varepsilon_G^{-2}$, and the summation and ReLU gate are exact. Therefore (200) follows. □

F.4.2. Laplace coefficient networks on the active branch

The Laplace coefficients are the only factors in the small-noise expansion that depend on the density regularity. We first factor each coefficient into schedule powers and Hölder functions of the projection coordinate, extending the latter to a buffered box while preserving the chart-support property. Then we approximate these coefficient functions and compose them with the coordinate projection network. The weighted corollary includes the powers of σ_t and $\boldsymbol{\nu}_i$ that multiply each coefficient in the Laplace sum.

Lemma F.19 (Factorization and extension of the Laplace coefficient functions)**.** *Fix $i \in \{1,\dots,C_{\mathcal{M}}\}$, $R \in \mathcal{R}$, $0 \le q \le \lfloor \beta \rfloor$, and $\boldsymbol{\lambda} \in \mathbb{N}^D$ with $\|\boldsymbol{\lambda}\|_1 \le q$. Assume $\beta - q > 0$. Let K_i° and $\square_i$ be the buffered coordinate boxes fixed in (20) and (21). For every intermediate box $\square_i^A$ with $K_i^\circ \Subset \square_i^A \Subset \square_i$, there exist a nonempty finite set $\mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)}$, integers $r_e \in \mathbb{Z}$, and functions*

$$B_{i,q,\boldsymbol{\lambda},e}^{(R)} \in \mathcal{H}^{\beta-q}(\square_i^A, C_{A,q}), \qquad e \in \mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)},$$

such that the function

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u},t) := \sum_{e \in \mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)}} m_t^{r_e} B_{i,q,\boldsymbol{\lambda},e}^{(R)}(\boldsymbol{u}), \qquad (\boldsymbol{u},t) \in \square_i^A \times [\underline{t},\overline{t}], \tag{201}$$

is an extension of $A_{i,q,\boldsymbol{\lambda}}^{(R)}$:

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u},t) = A_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u},t), \qquad (\boldsymbol{u},t) \in K_i^\circ \times [\underline{t},\overline{t}]. \tag{202}$$

With $S_i^{(R)} := \mathsf{supp} a_i^{(R)}$*, the extension can be chosen so that*

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u},t) = 0, \qquad (\boldsymbol{u},t) \in \big(\square_i^A \setminus S_i^{(R)}\big) \times [\underline{t},\overline{t}]. \tag{203}$$

Moreover,

$$\sup_{t\in[\underline{t},\overline{t}]} \left\| \overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\cdot,t) \right\|_{\mathcal{H}^{\beta-q}(\square_i^A)} \le C_{A,q}. \tag{204}$$

The constants $|\mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)}|$*,* $\max_e |r_e|$*, and* $C_{A,q}$ *are finite and depend only on* $i, q, \boldsymbol{\lambda}, R, \beta, B_a, \underline{m}, \overline{m}$*, and the geometric constants.*

Proof. By (178), every $\Psi_{i,\ell}$ is the sum of a term linear in $\boldsymbol{\nu}$ with coefficient m_t and a term independent of $\boldsymbol{\nu}$ with coefficient m_t^2. Hence each coefficient of $\boldsymbol{\nu}^{\boldsymbol{\lambda}}$ in $\mathcal{E}_{i,\ell}$ and $\mathcal{U}_{i,q}^{(R)}$ is a finite sum of terms of the form

$$m_t^a P(\boldsymbol{w}) H(\boldsymbol{u}) \partial^\alpha a_i^{(R)}(\boldsymbol{u}), \qquad |\alpha| \le q,$$

where a is an integer, P is a polynomial in $\boldsymbol{w}$, and H is a smooth geometric factor on $\square_i$. Therefore $A_{i,q,\boldsymbol{\lambda}}^{(R)}$ is a finite sum of integrals

$$m_t^a H(\boldsymbol{u}) \partial^\alpha a_i^{(R)}(\boldsymbol{u}) \int_{\mathbb{R}^d} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u})\boldsymbol{w}} \boldsymbol{w}^\rho \mathrm{d}\boldsymbol{w}, \tag{205}$$

with $|\alpha| \le q$ and finite multi-indices ρ.

Since $\boldsymbol{G}_i(\boldsymbol{u})$ is smooth and uniformly positive definite on $\square_i$, the Gaussian moment formula gives

$$\int_{\mathbb{R}^d} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u})\boldsymbol{w}} \boldsymbol{w}^\rho \mathrm{d}\boldsymbol{w} = m_t^{-d-|\rho|}(2\pi)^{d/2} \det(\boldsymbol{G}_i(\boldsymbol{u}))^{-1/2} P_\rho(\boldsymbol{G}_i(\boldsymbol{u})^{-1}),$$

with the usual convention that odd moments vanish. Substituting this identity into (205) gives an explicit finite formula for the coefficient on K_i°. Fix the stated intermediate box $\square_i^A$, and use the same explicit Gaussian-moment formula to define the extension on all of $\square_i^A$. Equivalently, after collecting the finitely many powers of m_t, define $B_{i,q,\boldsymbol{\lambda},e}^{(R)}$ directly on $\square_i^A$ as the corresponding finite sums of products of smooth geometric factors and derivatives $\partial^\alpha a_i^{(R)}$, $|\alpha| \le q$. If a coefficient is identically zero, take $\mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)}$ to be a singleton with $r_e = 0$ and $B_{i,q,\boldsymbol{\lambda},e}^{(R)} \equiv 0$.

Since $a_i^{(R)} \in \mathcal{H}^\beta$ on the coordinate domain and the geometric factors are smooth on $\square_i$, each directly defined $B_{i,q,\boldsymbol{\lambda},e}^{(R)}$ belongs to $\mathcal{H}^{\beta-q}(\square_i^A)$. The formula agrees with the original coefficient on K_i°, giving (202). Moreover, if $\boldsymbol{u} \notin S_i^{(R)} = \mathsf{supp} a_i^{(R)}$, then $a_i^{(R)}$ vanishes in a neighborhood of $\boldsymbol{u}$, and all derivatives $\partial^\alpha a_i^{(R)}(\boldsymbol{u})$ with $|\alpha| \le q$ vanish. The directly defined Gaussian-moment formula is therefore zero at $\boldsymbol{u}$, which proves (203). Taking $C_{A,q}$ to dominate the finitely many $\mathcal{H}^{\beta-q}$-norms of the directly defined factors proves (204) and the stated dependence of the constants. □

Lemma F.20 (ReLU approximation of the Laplace coefficient functions)**.** *Fix* $i, R, q, \boldsymbol{\lambda}$ *as in Theorem F.19, and assume* $\beta - q > 0$*. Let* $\square_i^A$*,* $\mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)}$*,* r_e*,* $B_{i,q,\boldsymbol{\lambda},e}^{(R)}$*, and* $\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}$ *be the objects supplied by Theorem F.19. For every* $0 < \varepsilon_A \le 1/2$*, there is a ReLU network*

$$\widetilde{A}_{i,q,\boldsymbol{\lambda},\varepsilon_A}^{(R)} : \mathbb{R}^d \times [\underline{t},\overline{t}] \to \mathbb{R}$$

such that

$$\sup_{(\boldsymbol{u},t)\in\square_i^A\times[\underline{t},\overline{t}]} \left| \widetilde{A}_{i,q,\boldsymbol{\lambda},\varepsilon_A}^{(R)}(\boldsymbol{u},t) - \overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u},t) \right| \le \varepsilon_A. \tag{206}$$

Moreover, for a finite constant $C_{A,\mathrm{nn}} \ge 1$*,*

$$\begin{aligned} L, \log B &\lesssim \log(e(d \vee (\beta-q))) \log^2(C_{A,\mathrm{nn}}\varepsilon_A^{-1}), \\ \|\boldsymbol{W}\|_\infty &\lesssim (d \vee (\beta-q))(C_{A,\mathrm{nn}}\varepsilon_A^{-1})^{d/(\beta-q)} + \log^3(C_{A,\mathrm{nn}}\varepsilon_A^{-1}), \\ S &\lesssim (d+\beta-q+2)^{d+4}(C_{A,\mathrm{nn}}\varepsilon_A^{-1})^{d/(\beta-q)} \log(C_{A,\mathrm{nn}}\varepsilon_A^{-1}) + \log^4(C_{A,\mathrm{nn}}\varepsilon_A^{-1}). \end{aligned} \tag{207}$$

Thus the non-logarithmic coefficient cost is $(\varepsilon_A^{-1})^{d/(\beta-q)}$*.*

Proof. Let

$$N_A := |\mathcal{E}^{(R)}_{i,q,\boldsymbol{\lambda}}|, \qquad r_A := \max_{e\in\mathcal{E}^{(R)}_{i,q,\boldsymbol{\lambda}}} |r_e|, \qquad M_A := 1 \vee \underline{m}^{-r_A} \vee \overline{m}^{r_A}.$$

For each e, apply Theorem I.1, after affine rescaling of $\square_i^A$ to $[0,1]^d$, to approximate $B^{(R)}_{i,q,\boldsymbol{\lambda},e}$ to accuracy

$$\varepsilon_B := \frac{\varepsilon_A}{8N_A M_A}.$$

This gives the intrinsic cost $(C\varepsilon_A^{-1})^{d/(\beta-q)}$, up to logarithmic factors.

For the one-dimensional schedule factor $m_t^{r_e}$, use the following three cases. If $r_e = 0$, represent $m_t^{r_e}$ by the constant-one network. If $r_e > 0$, approximate m_t by Theorem I.2 and form the fixed positive power by Theorem I.3. If $r_e < 0$, approximate $1/m_t$ by Theorem I.6 and form the fixed positive power $(1/m_t)^{-r_e}$ by Theorem I.3. Because r_A is fixed and $m_t \in [\underline{m}, \overline{m}]$, this schedule subnetwork has only logarithmic dependence on ε_A^{-1}, and can be made accurate enough that the error in each product

$$m_t^{r_e} B^{(R)}_{i,q,\boldsymbol{\lambda},e}(\boldsymbol{u})$$

is at most $\varepsilon_A/(2N_A)$ uniformly on $\square_i^A \times [\underline{t}, \overline{t}]$. Summing the N_A products in an affine output layer yields (206). Adding the intrinsic Hölder-approximation cost, the logarithmic schedule cost, and the fixed product/summation overhead gives (207). □

Corollary F.21 (Weighted coefficient approximation on the active branch)**.** *Fix $i, R, q, \boldsymbol{\lambda}$ as in Theorem F.19, and assume $\beta - q > 0$. Let $0 < \varepsilon \le 1/2$ be a localization parameter for which (89) holds on $\mathcal{I}_{\mathrm{sm}}(\varepsilon)$. Assume the small-noise branch satisfies*

$$\sigma_t \le \sigma_\star(\varepsilon) \qquad \text{for every } t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon),$$

For every $0 < \varepsilon_{\mathrm{term}} \le 1/2$, define

$$\varepsilon_{A,q} := \frac{1}{2} \wedge c_{A,q}\, \varepsilon_{\mathrm{term}}\, \sigma_\star(\varepsilon)^{-q} \left(\log(e/\varepsilon)\right)^{-q/2}, \tag{208}$$

with $c_{A,q} > 0$ sufficiently small. Then there is a ReLU network

$$\widetilde{\mathcal{A}}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}} : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to \mathbb{R}$$

with

$$\widetilde{\mathcal{A}}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}} \in \mathcal{NN}(L_{\mathcal{A}}, \boldsymbol{W}_{\mathcal{A}}, S_{\mathcal{A}}, B_{\mathcal{A}})$$

such that

$$\sup_{\substack{t\in\mathcal{I}_{\mathrm{sm}}(\varepsilon)\\ (\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon)}} \left|\widetilde{\mathcal{A}}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}}(\boldsymbol{x},t) - A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big)\right| \le \varepsilon_{A,q}. \tag{209}$$

Consequently,

$$\sup_{\substack{t\in\mathcal{I}_{\mathrm{sm}}(\varepsilon)\\ (\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon)}} \sigma_t^q |\boldsymbol{\nu}_i(\boldsymbol{x},t)^{\boldsymbol{\lambda}}| \left|\widetilde{\mathcal{A}}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}}(\boldsymbol{x},t) - A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big)\right| \le \varepsilon_{\mathrm{term}}. \tag{210}$$

Let $\alpha_q := (\beta - q) \wedge 1$. There is a finite constant $C_{\mathcal{A},q} \ge 1$, depending only on the fixed chart, geometric, schedule, and coefficient constants, such that

$$\begin{aligned}
L_{\mathcal{A}} &\lesssim \log(e(d\vee\beta))\log^3\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big) + \log(e(d\vee(\beta-q)))\log^2\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big),\\
\|\boldsymbol{W}_{\mathcal{A}}\|_\infty &\lesssim (d\vee\beta)\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big)^{d/(2(\beta\vee1))} + D\log^3\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big)\\
&\quad + (d\vee(\beta-q))\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big)^{d/(\beta-q)} + \log^3\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big),\\
S_{\mathcal{A}} &\lesssim A_{\mathcal{M}} dD(d+2(\beta\vee1))^{d+3}\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big)^{d/(2(\beta\vee1))}\log^2\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big)\\
&\quad + A_{\mathcal{M}} dD\log^5\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big)\\
&\quad + (d+\beta-q+2)^{d+4}\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big)^{d/(\beta-q)}\log\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big) + \log^4\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big),\\
\log B_{\mathcal{A}} &\lesssim \log^2\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-2/\alpha_q}\big) + \log^2\!\big(C_{\mathcal{A},q}\varepsilon_{A,q}^{-1}\big).
\end{aligned} \tag{211}$$

In particular, the total non-logarithmic size is bounded, up to logarithmic factors and fixed constants, by

$$\varepsilon_{A,q}^{-d/(\beta-q)} + \varepsilon_{A,q}^{-d/(\alpha_q(\beta\vee 1))}. \tag{212}$$

If $\sigma_\star(\varepsilon) \asymp \varepsilon^{1/\beta}$ and $\varepsilon_{\mathrm{term}} \asymp \varepsilon$, then both terms in (212) are bounded by

$$\lesssim \varepsilon^{-d/\beta}\big(\log(e/\varepsilon)\big)^{C_{\beta,q}},$$

for a finite exponent $C_{\beta,q}$. Hence the composed coefficient approximation does not worsen the intrinsic $\varepsilon^{-d/\beta}$ rate, up to logarithmic factors.

Proof. Use the fixed buffered box K_i° and the intermediate box $\square_i^A$ from Theorem F.19. Let

$$\alpha_q := (\beta - q) \wedge 1.$$

Apply Theorem F.11 with localization parameter ε and coordinate accuracy

$$\varepsilon_u := c\varepsilon_{A,q}^{1/\alpha_q},$$

where the constant $c > 0$ is chosen small enough that the coordinate output remains in $\square_i^A$ whenever the true coordinate $\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)$ lies in K_i°. This uses the fixed positive margin $\mathrm{dist}(K_i^\circ, \partial\square_i^A)$. Denote the resulting coordinate network by $\widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}$. Apply Theorem F.20 with accuracy $\varepsilon_{A,q}/2$, and compose the resulting coefficient network with this coordinate network. Since

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\cdot, t) \in \mathcal{H}^{\beta-q}(\square_i^A, C_{A,q})$$

uniformly in t, it is α_q-Hölder uniformly in t. Thus the coordinate error contributes at most $C_{A,q}\varepsilon_u^{\alpha_q}$, which is at most $\varepsilon_{A,q}/2$ by the choice of c. The intrinsic coefficient-network error contributes the other $\varepsilon_{A,q}/2$. Since $\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t) \in K_i^\circ$, the extension identity (202) gives

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), t\big) = A_{i,q,\boldsymbol{\lambda}}^{(R)}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), t\big),$$

which proves (209).

On $\mathcal{K}_i(\varepsilon)$, (199) gives

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \le C\sqrt{\log(e/\varepsilon)}.$$

Since $\|\boldsymbol{\lambda}\|_1 \le q$, this implies

$$|\boldsymbol{\nu}_i(\boldsymbol{x}, t)^{\boldsymbol{\lambda}}| \le C_q\big(\log(e/\varepsilon)\big)^{q/2}.$$

Together with $\sigma_t \le \sigma_\star(\varepsilon)$ and $\varepsilon_{A,q} \le c_{A,q}\varepsilon_{\mathrm{term}}\sigma_\star(\varepsilon)^{-q}\big(\log(e/\varepsilon)\big)^{-q/2}$, this yields

$$\sigma_t^q|\boldsymbol{\nu}_i^{\boldsymbol{\lambda}}|\varepsilon_{A,q} \le C_q\sigma_\star(\varepsilon)^q\big(\log(e/\varepsilon)\big)^{q/2}c_{A,q}\varepsilon_{\mathrm{term}}\sigma_\star(\varepsilon)^{-q}\big(\log(e/\varepsilon)\big)^{-q/2} \le \varepsilon_{\mathrm{term}}$$

after choosing $c_{A,q}$ small enough. This proves (210).

The coefficient network contributes the terms in (207) with accuracy $\varepsilon_{A,q}/2$. The coordinate network is obtained from Theorem F.11 with accuracy $\varepsilon_u = c\varepsilon_{A,q}^{1/\alpha_q}$; substituting this value into (154) gives the terms with scale $\varepsilon_{A,q}^{-2/\alpha_q}$. The final composition and the finite summations change the displayed quantities only by fixed constants. After increasing $C_{\mathcal{A},q}$, this proves (211). In particular, the coordinate network's non-logarithmic contribution is

$$(\varepsilon_u^{-2})^{d/(2(\beta\vee 1))} \lesssim \varepsilon_{A,q}^{-d/(\alpha_q(\beta\vee 1))}.$$

Together with the coefficient contribution $\varepsilon_{A,q}^{-d/(\beta-q)}$, this proves (212).

Finally, since $\sigma_\star(\varepsilon) \asymp \varepsilon^{1/\beta}$ and $\varepsilon_{\mathrm{term}} \asymp \varepsilon$, for sufficiently small ε, the minimum in (208) is inactive and

$$\varepsilon_{A,q} \asymp \varepsilon^{1-q/\beta}\big(\log(e/\varepsilon)\big)^{-q/2} = \varepsilon^{(\beta-q)/\beta}\big(\log(e/\varepsilon)\big)^{-q/2}.$$

Substituting this relation into the first term of (212) gives

$$\varepsilon_{A,q}^{-d/(\beta-q)} \lesssim \varepsilon^{-d/\beta}\big(\log(e/\varepsilon)\big)^{qd/(2(\beta-q))}.$$

For the second term, if $\beta - q \geq 1$, then $\alpha_q = 1$ and

$$\varepsilon_{A,q}^{-d/(\alpha_q(\beta\vee 1))} \lesssim \varepsilon^{-d(\beta-q)/(\beta(\beta\vee 1))}\big(\log(e/\varepsilon)\big)^{qd/(2(\beta\vee 1))} \leq \varepsilon^{-d/\beta}\big(\log(e/\varepsilon)\big)^{qd/(2(\beta\vee 1))},$$

because $\beta - q \leq \beta \vee 1$. If $\beta - q < 1$, then $\alpha_q = \beta - q$, and

$$\varepsilon_{A,q}^{-d/(\alpha_q(\beta\vee 1))} \lesssim \varepsilon^{-d/(\beta(\beta\vee 1))}\big(\log(e/\varepsilon)\big)^{qd/(2(\beta-q)(\beta\vee 1))} \leq \varepsilon^{-d/\beta}\big(\log(e/\varepsilon)\big)^{qd/(2(\beta-q)(\beta\vee 1))},$$

because $\beta \vee 1 \geq 1$. The displayed polylogarithmic rate follows. □

Corollary F.22 (Active-branch component approximation on an enlarged chart radius)**.** *Fix $c_a > 0$. Assume the c_a-active localization condition*

$$\delta_{\mathrm{act}}^{(c_a)}(\varepsilon, t) < \eta_\Pi/2, \qquad t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon),$$

holds. Then, for every chart i, on $\mathcal{K}_i^{(c_a)}(\varepsilon)$, the residual bound

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \leq c_a\sqrt{\log(e/\varepsilon)}$$

holds, and the following fixed-c_a approximation statements are valid. Their error bounds and network-size orders are the same as in the cited default statements, with constants allowed to depend on the fixed number c_a, and with the replacement

$$\mathcal{K}_i(\varepsilon) \mapsto \mathcal{K}_i^{(c_a)}(\varepsilon), \qquad \mathcal{K}_{\mathrm{act}}(\varepsilon) \mapsto \mathcal{K}_{\mathrm{act}}^{(c_a)}(\varepsilon), \qquad C_{\nu,\varepsilon} \mapsto 2\sigma_{\underline{t}}^{-1} + c_a\sqrt{\log(e/\varepsilon)}.$$

(i) The buffered-chart and reach-tube conclusions of Theorems F.4 and F.12 hold on $\mathcal{K}_i^{(c_a)}(\varepsilon)$. In particular $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \in K_i^\circ$ there.

(ii) For every coordinate accuracy ε_u, the active coordinate network in Theorem F.11 can be chosen so that its error bound holds uniformly on $\mathcal{K}_i^{(c_a)}(\varepsilon)$.

(iii) Theorem F.14 has the same statement on the enlarged active union $\mathcal{K}_{\mathrm{act}}^{(c_a)}(\varepsilon)$.

(iv) The residual, normal, Gaussian-factor, and weighted-coefficient approximation statements in Theorems F.17, F.18 and F.21 hold uniformly on $\mathcal{K}_i^{(c_a)}(\varepsilon)$.

For fixed c_a, the polynomial and logarithmic orders in all displayed network sizes are unchanged.

Proof. For the buffered-chart and reach-tube statements, the original proofs use the active-set definition only through the distance estimate

$$\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2 \leq c_a\sigma_t\sqrt{\log(e/\varepsilon)}, \qquad \boldsymbol{y} \in \mathcal{S}_i,$$

and the resulting reach/local-coordinate condition $\delta_{\mathrm{act}}^{(c_a)}(\varepsilon, t) < \eta_\Pi/2$. The assumed c_a-active localization condition therefore gives the same projected coordinate, buffered-domain inclusion, and uniqueness conclusions on $\mathcal{K}_i^{(c_a)}(\varepsilon)$. The residual estimate becomes

$$\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2 \leq c_a\sigma_t\sqrt{\log(e/\varepsilon)}, \qquad \|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \leq c_a\sqrt{\log(e/\varepsilon)}.$$

Thus every place where $C_{\nu,\varepsilon}$ appears may use $2\sigma_{\underline{t}}^{-1} + c_a\sqrt{\log(e/\varepsilon)}$. For the active chart-coordinate lemma, the one-step Gauss–Newton and anchored iteration estimates use only the reach condition and the radius parameter $C_{\nu,\varepsilon}$. With the replacement above, the contraction, truncation, and network-realization bounds are identical up to constants depending on the fixed c_a. For the global projection theorem, the original argument uses only the following two facts: an active pair $(\boldsymbol{x}, t)$ supplies a chart $i_\star$ for which the exact projected coordinate lies in $K_{i_\star}^\circ$, and all other candidates are separated by the same objective gap outside the accepted near-minimizer set. Both facts hold with $\mathcal{K}_{\mathrm{act}}^{(c_a)}(\varepsilon)$ in place of $\mathcal{K}_{\mathrm{act}}(\varepsilon)$, because

$(\boldsymbol{x}, t) \in \mathcal{K}_{i_\star}^{(c_a)}(\varepsilon)$ and the enlarged buffered-chart conclusion gives the good coordinate in $K_{i_\star}^{\circ}$. Thus the minimizer-selection and error estimates of Theorem F.14 hold on $\mathcal{K}_{\mathrm{act}}^{(c_a)}(\varepsilon)$ with only fixed-c_a changes in constants. This proves the fixed-c_a version of Theorem F.14.

The proofs of the residual and Gaussian-factor lemmas then use this fixed-c_a global projection theorem in place of the default one, together with the enlarged residual bound displayed above. The weighted coefficient corollary uses the fixed-c_a active coordinate network and the same coefficient extension. For fixed c_a, the replacement only changes constants and powers of the same logarithmic factor. The constructions, error allocations, and network sizes are otherwise identical. □

Corollary F.23 (H-active component approximation)**.** *Fix $c_a > 0$, $H \geq 1$, and $0 < \varepsilon \leq 1/2$. Assume $[t_0, t_1] \subset \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and*

$$c_a \frac{\sigma_t}{m_t}\sqrt{H} \leq \eta_\Pi/2, \qquad t \in [t_0, t_1].$$

Set $\sigma_H^\star := \sup_{t\in[t_0,t_1]} \sigma_t$, so $\sigma_H^\star \leq C_{\mathrm{sm}}\varepsilon^{1/\beta}$. For

$$A_t(H) := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \leq C_\varrho \sigma_t \sqrt{H}\},$$

define

$$\mathcal{K}_i^{(c_a,H)} := \left\{(\boldsymbol{x}, t) : \boldsymbol{x} \in A_t(H),\ \exists \boldsymbol{y} \in \mathcal{S}_i \text{ with } \|\boldsymbol{x} - m_t\boldsymbol{y}\|_2 \leq c_a\sigma_t\sqrt{H}\right\}.$$

Then, for every chart i, on $\mathcal{K}_i^{(c_a,H)}$ the projection $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ is unique, the projected coordinate satisfies $\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t) \in K_i^\circ$, and

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \leq c_a\sqrt{H}.$$

Moreover, the component approximation statements in Theorems F.11, F.14, F.17, F.18 and F.21 have H-active versions on the sets $\mathcal{K}_i^{(c_a,H)}$ and $\bigcup_i \mathcal{K}_i^{(c_a,H)}$. The corresponding size bounds are obtained from the cited bounds by replacing $C_{\nu,\varepsilon}$ with $2\sigma_{\underline{t}}^{-1} + c_a\sqrt{H}$ and by replacing each logarithmic scale $\log(C\varepsilon^{-1})$ with $H + \log(C\varepsilon^{-1})$. More explicitly, for the weighted-coefficient statement, fix the same indices $i, R, q, \boldsymbol{\lambda}$ and the same term accuracy $0 < \varepsilon_{\mathrm{term}} \leq 1/2$ as in Theorem F.21. In the H-active version of that corollary, the coefficient tolerance is

$$\varepsilon_{A,q}^H := \frac{1}{2} \wedge c_{A,q}\varepsilon_{\mathrm{term}}(\sigma_H^\star)^{-q}H^{-q/2}.$$

With this replacement, the weighted error bound in (210) holds on $\mathcal{K}_i^{(c_a,H)}$ with $\varepsilon_{A,q}$ replaced by $\varepsilon_{A,q}^H$. For fixed c_a, these replacements preserve the same non-logarithmic intrinsic costs and introduce only additional fixed polynomial powers of H.

Proof. The proof of Theorem F.22 depends on the active set only through the reach condition and the residual bound. The assumed H-active reach condition gives uniqueness of the metric projection, buffered-coordinate inclusion, and the Gauss–Newton contraction exactly as before, with $\sqrt{\log(e/\varepsilon)}$ replaced by $\sqrt{H}$. The residual bound follows from the active-distance inequality in the same way as in (199). The active coordinate network, global projection network, residual network, and Gaussian-factor network use this radius only through the scalar ranges for products, reciprocals, ramps, and squared norms. Thus their proofs carry over with $C_{\nu,\varepsilon}$ replaced by $2\sigma_{\underline{t}}^{-1} + c_a\sqrt{H}$ and with logarithmic factors enlarged to $H + \log(C\varepsilon^{-1})$. For the coefficient composition, the only changed analytic estimate is $|\boldsymbol{\nu}_i^{\boldsymbol{\lambda}}| \leq C_q H^{q/2}$ when $\|\boldsymbol{\lambda}\|_1 \leq q$. The definition of $\varepsilon_{A,q}^H$ and the bound $\sigma_t \leq \sigma_H^\star$ give the same weighted error proof as Theorem F.21. The intrinsic Hölder approximation exponents are unchanged, and the new powers of H are absorbed into the stated logarithmic-scale replacement. □

Lemma F.24 (Certified H-active chart gate from objective weights)**.** *Fix a chart index i, a radius multiplier $c_a \geq c_\star$, $H \geq 1$, $0 < \varepsilon \leq 1/2$, and a coordinate tolerance $0 < \eta \leq 1$. Let $[t_0, t_1] \subset \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and assume*

$$c_a \frac{\sigma_t}{m_t}\sqrt{H} \leq \eta_\Pi/2, \qquad t \in [t_0, t_1].$$

Set $A_t(H) := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \leq C_\varrho\sigma_t\sqrt{H}\}$. Define $\mathcal{K}_i^{(c_a,H)}$ from this $A_t(H)$ as in Theorem F.23. Let $\square_i^A$ be any fixed coordinate box with $K_i^\circ \Subset \square_i^A \Subset \square_i$. Then there exist ReLU networks

$$\Gamma_i^H : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to [0, 1], \qquad \breve{\boldsymbol{u}}_i : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to \square_i, \qquad \breve{\boldsymbol{u}}_i^A : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to \square_i^A,$$

with the following properties, uniformly for $t \in [t_0, t_1]$ and $\boldsymbol{x} \in A_t(H)$. If $(\boldsymbol{x}, t) \in \mathcal{K}_i^{(c_a,H)}$, then

$$\Gamma_i^H(\boldsymbol{x}, t) = 1, \qquad \|\breve{\boldsymbol{u}}_i(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_\infty \le \eta, \qquad \|\breve{\boldsymbol{u}}_i^A(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_\infty \le \eta. \tag{213}$$

If $\Gamma_i^H(\boldsymbol{x}, t) > 0$, then $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t) \in U_i$ and, with $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) := \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t))$,

$$\|\breve{\boldsymbol{u}}_i(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_\infty \le \eta. \tag{214}$$

The network $\breve{\boldsymbol{u}}_i^A$ is the exact coordinatewise clipping of $\breve{\boldsymbol{u}}_i$ to $\square_i^A$. The sizes are bounded by the H-active global projection-objective construction with projection accuracy $c\eta$ and with every logarithmic scale enlarged to $H + \log(C\varepsilon^{-1})$; in particular, this gate adds no non-logarithmic factor beyond the intrinsic projection-objective cost $\eta^{-d/(\beta\vee 1)}$ and fixed-atlas arithmetic.

Proof. We repeat the objective-weight part of the proof of Theorem F.14, using the H-active extension from Theorem F.23, but we keep the chart identity of each candidate. Let

$$\mathcal{J} := \{(j, \boldsymbol{a}) : 1 \le j \le C_{\mathcal{M}},\ \boldsymbol{a} \in \mathcal{A}_j^{\mathrm{anc}}\}, \qquad N_{\mathrm{cand}} := |\mathcal{J}|.$$

Choose an ambient candidate accuracy $\varepsilon_{\mathrm{cert}} = c_{\mathrm{cert}}\eta$, where the fixed geometric constant $c_{\mathrm{cert}} > 0$ will be decreased several times below. For each $(j, \boldsymbol{a}) \in \mathcal{J}$, form the anchored coordinate candidate $\widetilde{\boldsymbol{u}}_{j,\boldsymbol{a}}$, the manifold candidate $\boldsymbol{y}_{j,\boldsymbol{a}} := \boldsymbol{z}_j(\widetilde{\boldsymbol{u}}_{j,\boldsymbol{a}})$, and an objective network $\widetilde{\mathcal{D}}_{j,\boldsymbol{a}}$ approximating $\|\boldsymbol{x}/m_t - \boldsymbol{y}_{j,\boldsymbol{a}}\|_2^2$. The coordinate candidates are box-clipped by construction, so all $\boldsymbol{y}_{j,\boldsymbol{a}}$ and all objective networks are well-defined on the whole set $A_t(H)$, even when chart j is not the active chart. The objective and arithmetic tolerances are chosen so that

$$|\widetilde{\mathcal{D}}_{j,\boldsymbol{a}} - \|\boldsymbol{x}/m_t - \boldsymbol{y}_{j,\boldsymbol{a}}\|_2^2| \le \gamma/64, \qquad \gamma := \varepsilon_{\mathrm{cert}}^2, \tag{215}$$

uniformly on $A_t(H)$. Because $c_a \ge c_\star \ge C_\varrho$ and the chart supports cover $\mathcal{M}$, every $(\boldsymbol{x}, t) \in A_t(H)$ belongs to at least one H-active chart set $\mathcal{K}_j^{(c_a,H)}$. For such a chart, Theorem F.23 gives $\boldsymbol{u}_j^{\Pi} \in K_j^\circ$, and Theorem F.7 supplies an anchor whose local Gauss–Newton candidate is within $\mathcal{O}(\varepsilon_{\mathrm{cert}})$ of the nearest projection. After decreasing c_{cert}, the corresponding exact objective is at most $\gamma/64$ above the true minimum $\mathrm{dist}(\boldsymbol{x}/m_t, \mathcal{M})^2$.

Let $\widetilde{\mathcal{D}}_{\min} := \min_{(j,\boldsymbol{a})\in\mathcal{J}} \widetilde{\mathcal{D}}_{j,\boldsymbol{a}}$ and define

$$\omega_{j,\boldsymbol{a}} := \mathsf{ReLU}\Big(\gamma - \big(\widetilde{\mathcal{D}}_{j,\boldsymbol{a}} - \widetilde{\mathcal{D}}_{\min}\big)\Big), \qquad W_{\mathrm{glob}} := \sum_{(j,\boldsymbol{a})\in\mathcal{J}} \omega_{j,\boldsymbol{a}}, \qquad \Omega_i := \sum_{\boldsymbol{a}\in\mathcal{A}_i^{\mathrm{anc}}} \omega_{i,\boldsymbol{a}}.$$

At least one candidate attains $\widetilde{\mathcal{D}}_{\min}$, hence

$$\gamma \le W_{\mathrm{glob}} \le N_{\mathrm{cand}}\gamma. \tag{216}$$

If $(\boldsymbol{x}, t) \in \mathcal{K}_i^{(c_a,H)}$, the good chart-$i$ candidate described above satisfies

$$\widetilde{\mathcal{D}}_{i,\boldsymbol{a}_\star} - \widetilde{\mathcal{D}}_{\min} \le \frac{3\gamma}{64},$$

because $\widetilde{\mathcal{D}}_{\min}$ is not below the true minimum by more than $\gamma/64$. Therefore $\omega_{i,\boldsymbol{a}_\star} \ge 61\gamma/64$, and (216) gives

$$\frac{\Omega_i}{W_{\mathrm{glob}}} \ge \tau_0, \qquad \tau_0 := \frac{1}{2N_{\mathrm{cand}}}. \tag{217}$$

Conversely, suppose $\Omega_i > 0$. Then some chart-i candidate has positive weight. The same calculation as (174) and (175), with the tolerances in (215), implies

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)\|_2 \le C\varepsilon_{\mathrm{cert}} \tag{218}$$

for every positive-weight chart-i candidate. Let

$$\delta_i^U := \mathrm{dist}_{\mathbb{R}^D}\big(\boldsymbol{z}_i(\square_i), \mathcal{M} \setminus U_i\big) > 0,$$

with the convention $\operatorname{dist}(A, \emptyset) = +\infty$. After decreasing c_{cert} so that $C\varepsilon_{\text{cert}} < \delta_i^U/2$, (218) forces $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t) \in U_i$ whenever $\Omega_i > 0$. The inverse chart is Lipschitz on the compact set $\phi_i^{-1}(\square_i)$, so every positive-weight chart-i coordinate candidate satisfies

$$\|\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_\infty \le C_i \varepsilon_{\text{cert}}. \tag{219}$$

Define the ideal chart coordinate average

$$\boldsymbol{u}_{i,\text{id}}(\boldsymbol{x}, t) := \frac{\sum_{\boldsymbol{a} \in \mathcal{A}_i^{\text{anc}}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t) \widetilde{\boldsymbol{u}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t)}{\Omega_i(\boldsymbol{x}, t)}$$

whenever $\Omega_i > 0$. If $\Omega_i / W_{\text{glob}} \ge \tau_0/32$, then (216) gives $\Omega_i \ge \tau_0\gamma/32$, and (219) gives $\|\boldsymbol{u}_{i,\text{id}} - \boldsymbol{u}_i^{\Pi}\|_\infty \le C_i \varepsilon_{\text{cert}}$. Implement the denominator as $\Omega_i \vee (\tau_0\gamma/64)$, approximate the reciprocal and products to error at most $c\eta$, and apply the exact clipping map $\Pi_i^{(\text{box})}$ as the final layer. After decreasing c_{cert}, the implemented coordinate $\breve{\boldsymbol{u}}_i$ maps into $\square_i$ and satisfies (214) whenever the gate below is nonzero, because the clipping is 1-Lipschitz and fixes $\boldsymbol{u}_i^{\Pi}$.

It remains only to construct the binary-safe chart gate. Approximate Ω_i/W_{glob} with error at most $\tau_0/32$ on the interval in (216). Let $\theta_{\text{cert}} : \mathbb{R} \to [0, 1]$ be the ReLU ramp that is zero on $(-\infty, \tau_0/16]$ and one on $[\tau_0/2, \infty)$, and define Γ_i^H by applying θ_{cert} to this implemented ratio. If $(\boldsymbol{x}, t) \in \mathcal{K}_i^{(c_a, H)}$, then (217) and the ratio error imply that the ramp input is at least $\tau_0/2$, so $\Gamma_i^H = 1$. If $\Gamma_i^H > 0$, then the exact ratio is positive and in fact $\Omega_i/W_{\text{glob}} \ge \tau_0/32$, so the clipped denominator in the coordinate average is inactive and the coordinate certification above applies. Finally, define $\breve{\boldsymbol{u}}_i^A$ by exact coordinatewise clipping of $\breve{\boldsymbol{u}}_i$ to $\square_i^A$. On $\mathcal{K}_i^{(c_a, H)}$, the true coordinate lies in $K_i^\circ \Subset \square_i^A$, and the coordinatewise clipping is 1-Lipschitz and fixes this true coordinate. All added operations are finite sums, minima, ReLU ramps, clipped reciprocals, products, and coordinatewise clipping, so their costs are logarithmic or fixed-atlas costs on top of the H-active projection-objective networks. This proves the lemma. □

F.4.3. Projection–Laplace chart approximation for small noise

We first state the chart-level projection–Laplace approximation used by the density-lower-bound small-noise construction. The proposition below is deliberately local: it certifies the neural Laplace expansion only on the enlarged chart-active set where the chart coordinate and residual subnetworks have proved error bounds. Outside the default chart-active set, the true chart integral is already small by the inactive-tail lemma; the final de-Gaussianized construction below adds a separate certified chart gate before using inactive-chart neural outputs. This separation keeps the proof auditable because no local chart-coordinate network is used outside the region where its approximation guarantee is available.

Proposition F.25 (Projection–Laplace chart approximation for small noise)**.** *Assume Theorems 2.2 and 2.4. Let $0 < \varepsilon_{\text{sm}} \le 1/2$, and assume the localization condition (89), as well as the same condition with $3c_\star$ in place of $c_\star$, holds for every $t \in \mathcal{I}_{\text{sm}}(\varepsilon_{\text{sm}})$. For every chart i and every $R \in \mathcal{R}$, there exists a ReLU network $\widetilde{R}_i^{\text{sm}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ such that*

$$\sup_{\substack{t \in \mathcal{I}_{\text{sm}}(\varepsilon_{\text{sm}}) \\ (\boldsymbol{x}, t) \in \mathcal{K}_i^{(3c_\star)}(\varepsilon_{\text{sm}})}} \left| R_i(\boldsymbol{x}, t) - \widetilde{R}_i^{\text{sm}}(\boldsymbol{x}, t) \right| \le \varepsilon_{\text{sm}}. \tag{220}$$

In addition, the true inactive tail satisfies

$$\sup_{\substack{t \in \mathcal{I}_{\text{sm}}(\varepsilon_{\text{sm}}) \\ \boldsymbol{x} \in A_t,\ (\boldsymbol{x}, t) \notin \mathcal{K}_i(\varepsilon_{\text{sm}})}} |R_i(\boldsymbol{x}, t)| \le \varepsilon_{\text{sm}}/6. \tag{221}$$

Moreover, let $C_{\text{size,sm}} \ge e$ be a fixed constant large enough to absorb the chart, schedule, coefficient, and arithmetic constants, and define

$$\Lambda_{\text{sm}} := e \vee C_{\text{size,sm}} \varepsilon_{\text{sm}}^{-1} \vee \underline{t}^{-1}. \tag{222}$$

For finite exponents $\mathfrak{a}_{\text{sm}}$ and $\mathfrak{p}_{\beta,d}$ depending only on (β, d), the same construction can be chosen so that, up to fixed geometry-dependent constants,

$$L + \log B \lesssim \log(e(d \vee \beta)) \log^3 \Lambda_{\text{sm}}, \qquad \|\boldsymbol{W}\|_\infty + S \lesssim D^{\mathfrak{a}_{\text{sm}}} A_{\mathcal{M}} \varepsilon_{\text{sm}}^{-d/\beta} \log^{\mathfrak{p}_{\beta,d}} \Lambda_{\text{sm}}. \tag{223}$$

The expanded depth, width, sparsity, and log-weight bookkeeping is recorded in Theorem F.26.

Lemma F.26 (Small-noise chart size bookkeeping)**.** *Work in the setting of Theorem F.25. For $0 \le q \le \lfloor\beta\rfloor$, write $\alpha_q := (\beta - q) \wedge 1$. There is a finite exponent $\mathfrak{a}_{\mathrm{sm}}$, depending only on (β, d), such that the network in Theorem F.25 may be chosen so that the following bounds hold. More explicitly, one may take*

$$\mathfrak{a}_{\mathrm{sm}} = \left\lceil \frac{\lfloor\beta\rfloor d}{\beta - \lfloor\beta\rfloor} + \frac{d}{\beta \vee 1} + 3 \right\rceil.$$

In the display below, every occurrence of $\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$ may equivalently be replaced by $\log \Lambda_{\mathrm{sm}}$. This only enlarges the stated bounds and makes the lower-time dependence of the schedule networks explicit.

$$\begin{aligned}
L &\lesssim \log(e(d \vee \beta)) \log^3\!\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big),\\
\|\boldsymbol{W}\|_\infty &\lesssim D^{\mathfrak{a}_{\mathrm{sm}}} \frac{(D+d+\beta)^{\lfloor\beta\rfloor+1}}{\lfloor\beta\rfloor!} \Bigg[\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-2}\big)^{d/(2(\beta\vee 1))}\\
&\qquad + \sum_{q=0}^{\lfloor\beta\rfloor} \Bigg\{ \Big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-2(1-q/\beta)/\alpha_q} \log^{q/\alpha_q}\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big)\Big)^{d/(2(\beta\vee 1))}\\
&\qquad + \Big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-(1-q/\beta)} \log^{q/2}\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big)\Big)^{d/(\beta-q)} \Bigg\} \Bigg] \log^{3+d/2}\!\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big),\\
S &\lesssim D^{\mathfrak{a}_{\mathrm{sm}}} A_{\mathcal{M}} \frac{dD(D+d+\beta)^{\lfloor\beta\rfloor}(d+2(\beta\vee 1))^{d+4}}{\lfloor\beta\rfloor!} \Bigg[\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-2}\big)^{d/(2(\beta\vee 1))}\\
&\qquad + \sum_{q=0}^{\lfloor\beta\rfloor} \Bigg\{ \Big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-2(1-q/\beta)/\alpha_q} \log^{q/\alpha_q}\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big)\Big)^{d/(2(\beta\vee 1))}\\
&\qquad + \Big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-(1-q/\beta)} \log^{q/2}\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big)\Big)^{d/(\beta-q)} \Bigg\} \Bigg] \log^{5+d/2}\!\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big),\\
\log B &\lesssim \log^2\!\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big).
\end{aligned} \tag{224}$$

The factors $\log^{3+d/2}(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$ and $\log^{5+d/2}(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$ include the logarithmic overhead of the projection-coordinate gate, the residual ramp, and the final products. No mesh of $\mathcal{S}_i$ at scale $\sigma_{\underline{t}}$ is used; therefore the small-noise branch has no polynomial $\sigma_{\underline{t}}^{-1}$ factor in its width or sparsity bound. The fixed localization constant $c_\star$ is chosen from the annular Gaussian-tail bound in Theorem F.1 and is independent of $\sigma_{\underline{t}}$. For fixed d, β and fixed geometry, the bracketed non-logarithmic factor in (224) is

$$\lesssim \varepsilon_{\mathrm{sm}}^{-d/\beta} \log^{\mathfrak{p}_{\beta,d}}\!\big(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}\big)$$

after the finite order q_{geom} in (1) is fixed, and one may take

$$\mathfrak{p}_{\beta,d} = C + \max\left\{ 8(q_{\mathrm{geom}}+2), \left\lceil \lfloor\beta\rfloor + 1 + \frac{\lfloor\beta\rfloor d}{\beta - \lfloor\beta\rfloor} + \frac{d}{\beta\vee 1} + \frac{d}{2} + 6 \right\rceil \right\}.$$

Here the C in the exponent is universal and independent of $D, n, \sigma_t, p_{\min}$, and the numerical geometry constants; the resulting polynomial ambient exponent may depend on the fixed q_{geom}. In the main $d > 2$ regime we take the canonical choice $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$ from (2); then $q_{\mathrm{geom}} = \mathcal{O}_\beta(1)$ and the displayed formula gives $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$. Thus the small-noise branch has intrinsic approximation rate $\varepsilon_{\mathrm{sm}}^{-d/\beta}$, up to logarithmic factors, while the ambient-dimension dependence is polynomial in D. The factor $D^{\mathfrak{a}_{\mathrm{sm}}}$ records the additional powers of the retained Laplace-term count that enter through the per-term coefficient accuracies, together with the squared-normal residual-gate cost; all lower-noise dependence is confined to logarithmic schedule and reciprocal accuracies.

Proof of Theorem F.25 and Theorem F.26. Fix i and $R \in \mathcal{R}$. The number of retained $(q, \boldsymbol{\lambda})$-terms is

$$\sum_{q=0}^{\lfloor\beta\rfloor} \#\{\boldsymbol{\lambda} \in \mathbb{N}^D : \|\boldsymbol{\lambda}\|_1 \le q\} = \sum_{q=0}^{\lfloor\beta\rfloor} \binom{D+q}{q} = \binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor}.$$

Also, since $d < D$ and $\lfloor\beta\rfloor < \beta$,

$$\binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor}(D+d+\beta) \le \frac{(D+d+\beta)^{\lfloor\beta\rfloor+1}}{\lfloor\beta\rfloor!},$$

which is the simpler ambient-dimension factor used in the width bound. Similarly,

$$\binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor} dD(d+2(\beta\vee 1))^{d+4} \le \frac{dD(D+d+\beta)^{\lfloor\beta\rfloor}(d+2(\beta\vee 1))^{d+4}}{\lfloor\beta\rfloor!},$$

which is the corresponding factor used in the sparsity bound. The same retained-term count also appears in the allocated accuracies $\varepsilon^{\mathrm{sm}}_{\mathrm{term}}$ and $\varepsilon^{\mathrm{sm}}_{A,q}$ below. Since $q \in \{0,\dots,\lfloor\beta\rfloor\}$ is a finite set and the coefficient network sizes contain only fixed powers of $(\varepsilon^{\mathrm{sm}}_{A,q})^{-1}$, all additional powers of $\binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor}$ are bounded by $D^{\mathfrak{a}_{\mathrm{sm}}}$ after increasing $\mathfrak{a}_{\mathrm{sm}}(\beta,d)$. Indeed,

$$\binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor} \le C_\beta D^{\lfloor\beta\rfloor}, \qquad \max_{0\le q\le\lfloor\beta\rfloor}\left\{\frac{d}{\beta-q}, \frac{d}{((\beta-q)\wedge 1)(\beta\vee 1)}\right\} \le \frac{d}{\beta-\lfloor\beta\rfloor}.$$

Thus the retained-term allocation contributes at most $D^{\lfloor\beta\rfloor d/(\beta-\lfloor\beta\rfloor)}$, up to fixed (β,d)-dependent constants, before the other displayed ambient factors are included. The residual and coefficient-coordinate constructions contribute at most $D^{d/(2(\beta\vee 1))}$ beyond the explicit leading D-factor. The largest projection-normal auxiliary cost comes from operations depending on $\|\boldsymbol{\nu}_i\|_2^2$. For the original chart expansion this is the Gaussian factor in Theorem F.18; for the de-Gaussianized construction below it is the residual-norm gate. In both cases the normal approximation must be controlled in ℓ_2, so a coordinatewise normal accuracy of order $D^{-1/2}$ is required up to logarithmic factors. Equivalently the internal scale in (198) is at most D^2, and therefore this part contributes at most $D^{d/(\beta\vee 1)}$. The remaining residual ramps, products, and exact affine sums contribute only fixed ambient powers, absorbed by the $+3$ in $\mathfrak{a}_{\mathrm{sm}}$. The constant summands in the bracketed size factor are dominated by the $q=0$ terms of the displayed sum because $C_{\mathrm{size,sm}} \ge e$, $0 < \varepsilon_{\mathrm{sm}} \le 1/2$, and $d \ge 1$. Set

$$\varepsilon^{\mathrm{sm}}_{\mathrm{term}} := c_{\mathrm{term}}\varepsilon_{\mathrm{sm}}\binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor}^{-1}, \qquad \sigma_{\mathrm{sm}} := C_{\mathrm{sm}}\varepsilon_{\mathrm{sm}}^{1/\beta}.$$

For $0 \le q \le \lfloor\beta\rfloor$, define

$$\alpha_q := (\beta-q)\wedge 1, \qquad \varepsilon^{\mathrm{sm}}_{A,q} := \frac{1}{2} \wedge c^{\mathrm{sm}}_{A,q}\varepsilon_{\mathrm{sm}}^{1-q/\beta}\big(\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})\big)^{-q/2},$$

where $c^{\mathrm{sm}}_{A,q} > 0$ is chosen no larger than the fixed quantity

$$c_{A,q}c_{\mathrm{term}}C_{\mathrm{sm}}^{-q}\binom{D+\lfloor\beta\rfloor+1}{\lfloor\beta\rfloor}^{-1},$$

and $c_{A,q}$ is the constant from Theorem F.21. By the fixed choice of $c_\star$ in the active-set definition, and after decreasing the fixed numerical budgets below if necessary, the inactive tail in Theorem F.1 is at most $\varepsilon_{\mathrm{sm}}/6$.

Step 1: Analytic small-noise target. On the active region $\mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, use Theorem F.16. On the inactive region inside A_t, use zero. Thus define

$$R^{\mathrm{sm}}_{i,\mathrm{app}}(\boldsymbol{x},t) := \begin{cases} \exp\left(-\frac{\|\boldsymbol{r}_i(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right)\sum_{q=0}^{\lfloor\beta\rfloor}\sigma_t^q\sum_{\|\boldsymbol{\lambda}\|_1\le q}A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big)\boldsymbol{\nu}_i(\boldsymbol{x},t)^{\boldsymbol{\lambda}}, & (\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon_{\mathrm{sm}}),\\ 0, & (\boldsymbol{x},t)\notin\mathcal{K}_i(\varepsilon_{\mathrm{sm}}). \end{cases}$$

By Theorems F.1 and F.16, uniformly over $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$ and $\boldsymbol{x} \in A_t$,

$$|R_i(\boldsymbol{x},t) - R^{\mathrm{sm}}_{i,\mathrm{app}}(\boldsymbol{x},t)| \le \varepsilon_{\mathrm{sm}}/3. \tag{225}$$

The active part follows because the remainder in (190) is bounded by $C^{\max}_{\mathrm{Lap}}\sigma_t^\beta$, and the threshold C_{sm} in Theorem D.1 was chosen so that $C^{\max}_{\mathrm{Lap}}C^\beta_{\mathrm{sm}} \le 1/6$. Hence $C^{\max}_{\mathrm{Lap}}\sigma_t^\beta \le \varepsilon_{\mathrm{sm}}/6$ on $\mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$. The inactive part is bounded by $\varepsilon_{\mathrm{sm}}/6$ by the fixed choice of $c_\star$. This proves the true-tail statement (221).

Step 2: Projection-coordinate gate. Let $\square_i^A$ be the coefficient-extension box from Theorem F.19, so that $K_i^\circ \Subset \square_i^A \Subset \square_i$. Define the enlarged active set

$$\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}}) := \mathcal{K}_i^{(3c_\star)}(\varepsilon_{\mathrm{sm}}).$$

By the $3c_\star$-variant of the localization condition, the proof of Theorem F.4 applies on $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}})$. In particular, $\boldsymbol{u}_i^\Pi(\boldsymbol{x},t) \in K_i^\circ$ there and the fixed-radius component approximations in Theorem F.22, with $c_a = 3c_\star$, are available.

Choose two piecewise-affine ReLU ramps. The first is a coordinate bump $\theta_i^A : \mathbb{R}^d \to [0,1]$, equal to one on K_i° and equal to zero outside $\square_i^A$. Since the chart atlas and boxes are fixed, this bump is implemented with a fixed number of one-dimensional ramps and ReLU max/min operations. The second is a residual bump $\theta_{\mathrm{res}} : \mathbb{R}_+ \to [0,1]$, equal to one on

$$\left[0, \left(\frac{3c_\star}{2}\right)^2 \log(e/\varepsilon_{\mathrm{sm}})\right]$$

and equal to zero on

$$\left[(2c_\star)^2 \log(e/\varepsilon_{\mathrm{sm}}), \infty\right].$$

Let $\widetilde{\boldsymbol{u}}_i^\Pi$ be the chart coordinate of the approximate projection produced by the fixed-radius version of Theorem F.14, and let $\widetilde{\boldsymbol{\nu}}_i$ be the normalized-residual network from the fixed-radius version of Theorem F.17. Write

$$\Lambda_{\mathrm{gate}} := 1 + \log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1}), \qquad \varepsilon_{\nu,\mathrm{gate}} := c_{\mathrm{gate}}\big(1 \wedge D^{-1/2}\Lambda_{\mathrm{gate}}^{1/2}\big).$$

Choose the projection-coordinate accuracy and normal-residual accuracy so that, on $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}})$,

$$\|\widetilde{\boldsymbol{u}}_i^\Pi - \boldsymbol{u}_i^\Pi\|_\infty \le c_{\mathrm{gate}}\Lambda_{\mathrm{gate}}^{-1}, \qquad \|\widetilde{\boldsymbol{\nu}}_i - \boldsymbol{\nu}_i\|_\infty \le \varepsilon_{\nu,\mathrm{gate}},$$

where $c_{\mathrm{gate}} > 0$ is a sufficiently small geometric constant. Form $\|\widetilde{\boldsymbol{\nu}}_i\|_2^2$ using the multiplication lemma and define

$$\Theta_i^{\mathrm{sm}}(\boldsymbol{x},t) := \theta_i^A\big(\widetilde{\boldsymbol{u}}_i^\Pi(\boldsymbol{x},t)\big)\,\theta_{\mathrm{res}}\big(\|\widetilde{\boldsymbol{\nu}}_i(\boldsymbol{x},t)\|_2^2\big), \tag{226}$$

with the final product again implemented by Theorem I.3. The normal-residual accuracy gives

$$\|\widetilde{\boldsymbol{\nu}}_i - \boldsymbol{\nu}_i\|_2 \le c_{\mathrm{gate}}\Lambda_{\mathrm{gate}}^{1/2},$$

and, on the enlarged active region, $\|\boldsymbol{\nu}_i\|_2 \le C\Lambda_{\mathrm{gate}}^{1/2}$. Hence

$$\big|\|\widetilde{\boldsymbol{\nu}}_i\|_2^2 - \|\boldsymbol{\nu}_i\|_2^2\big| \le Cc_{\mathrm{gate}}\Lambda_{\mathrm{gate}}.$$

After choosing c_{gate} and the square-subnetwork accuracy small enough, the implemented squared norm stays inside the fixed separation margin between the two residual-ramp thresholds. The corresponding residual-network size has internal scale $C_{\mathrm{res}}D\varepsilon_{\nu,\mathrm{gate}}^{-2} \le CC_{\mathrm{res}}D^2$, up to logarithmic factors, which is the $D^{d/(\beta\vee 1)}$ contribution recorded in $\mathfrak{a}_{\mathrm{sm}}$. The gate satisfies

$$\Theta_i^{\mathrm{sm}} = 1 \quad \text{on } \mathcal{K}_i(\varepsilon_{\mathrm{sm}}),$$

with the allocated gate error absorbed into the final $\mathcal{O}(\varepsilon_{\mathrm{sm}})$ approximation budget. Indeed, on $\mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, the buffered-chart lemma gives $\boldsymbol{u}_i^\Pi \in K_i^\circ$ and $\|\boldsymbol{\nu}_i\|_2 \le c_\star\sqrt{\log(e/\varepsilon_{\mathrm{sm}})}$, so both ramps are equal to one after the above perturbation margin. The proposition makes no neural-output claim outside $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}})$. This is essential: the chart coordinate and residual subnetworks above are certified on $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}})$, and the analytic tail bound (221) handles the true chart integral away from the default active set. Thus the implemented localization is obtained from projection coordinates, normalized residuals, and support-preserving coefficient extensions only on the certified enlarged active region, not from a $\sigma_{\underline{t}}$-mesh of $\mathcal{S}_i$. Its size is absorbed into the same polylogarithmic factor as the component networks and has no polynomial $\sigma_{\underline{t}}^{-1}$ contribution.

On the transition region $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}}) \setminus \mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, no point of $\mathcal{S}_i$ satisfies the active distance condition with radius $c_\star$. Therefore the true chart integral is $\mathcal{O}(\varepsilon_{\mathrm{sm}})$ by Theorem F.1 and the fixed choice of $c_\star$. The neural output on this transition region is controlled after the component construction by the support dichotomy in Step 4 below.

Step 3: Neural approximation of the active expansion. For every $q \le \lfloor\beta\rfloor$ and $\|\boldsymbol{\lambda}\|_1 \le q$, approximate

$$A_{i,q,\boldsymbol{\lambda}}^{(R)}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x},t),t\big)$$

by the fixed-radius version of Theorem F.21, supplied by Theorem F.22, with term accuracy

$$\varepsilon_{\mathrm{term}}^{\mathrm{sm}}.$$

Here and below, Theorem F.21 and the residual and Gaussian lemmas are used through Theorem F.22 with $c_a = 3c_\star$. Approximate $\boldsymbol{\nu}_i$ by Theorem F.17, form the monomial $\boldsymbol{\nu}_i^{\boldsymbol{\lambda}}$ by Theorem I.3, approximate σ_t^q by the constant-one network when $q = 0$ and by Theorem I.8 when $q \geq 1$, and approximate the Gaussian factor by Theorem F.18. In the case $q \geq 1$, the accuracy parameter passed to Theorem I.8 is chosen as the minimum of the required schedule accuracy, $\underline{t}/2$, and $1/2$. This makes the admissibility condition in Theorem I.8 automatic. The resulting factor $\log \underline{t}^{-1}$ is bounded by $\log \Lambda_{\mathrm{sm}}$, hence is included in the logarithmic factors displayed in (224). Internal accuracies are chosen so that each retained term contributes at most $\varepsilon_{\mathrm{term}}^{\mathrm{sm}}$ to the final error after multiplication by the bounded remaining factors. The bounds

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \leq C\sqrt{\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})}, \qquad \sigma_t \leq C_{\mathrm{sm}}\varepsilon_{\mathrm{sm}}^{1/\beta},$$

are exactly the bounds used in Theorem F.21, so the coefficient-composition cost remains $\varepsilon_{\mathrm{sm}}^{-d/\beta}$ up to logarithms for every $\beta > 0$.

Multiplying by the projection-coordinate gate Θ_i^{sm}, summing the finitely many $(q, \boldsymbol{\lambda})$-terms in an affine output layer, and allocating one more $\mathcal{O}(\varepsilon_{\mathrm{sm}})$ budget to the final products gives a network $\widetilde{R}_i^{\mathrm{sm}}$.

Step 4: Transition-region control. We first bound this network on $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}}) \setminus \mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, where the gate can take values between 0 and 1. Let $\boldsymbol{u}^\Pi = \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)$. If $\boldsymbol{u}^\Pi \in S_i^{(R)} := \mathrm{supp} a_i^{(R)}$, then $\boldsymbol{z}_i(\boldsymbol{u}^\Pi) = \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t)) \in \mathcal{S}_i$. Since $(\boldsymbol{x}, t) \notin \mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, this particular support point does not satisfy the active distance condition. Hence

$$\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2 = \|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u}^\Pi)\|_2 > c_\star \sigma_t \sqrt{\log \varepsilon_{\mathrm{sm}}^{-1}}.$$

The exact normal Gaussian factor is therefore at most $\varepsilon_{\mathrm{sm}}^{c_\star^2/2}$, and the approximated Gaussian factor is small up to its allocated additive error. If $\boldsymbol{u}^\Pi \notin S_i^{(R)}$, then (203) gives $\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}^\Pi, t) = 0$ for every retained coefficient. The implemented coefficient network is evaluated at the approximate coordinate $\widetilde{\boldsymbol{u}}_i^\Pi$ used in $\widetilde{\mathcal{A}}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}^{\mathrm{sm}}}^{(R)}$, not exactly at $\boldsymbol{u}^\Pi$. By the fixed-radius coordinate approximation in Theorem F.22, the choice $\varepsilon_u \asymp (\varepsilon_{A,q}^{\mathrm{sm}})^{1/((\beta-q)\wedge 1)}$, the $(\beta - q)$-Hölder bound in (204), and the intrinsic coefficient-network error,

$$\left|\widetilde{\mathcal{A}}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}^{\mathrm{sm}}}^{(R)}(\boldsymbol{x}, t)\right| \leq C\|\widetilde{\boldsymbol{u}}_i^\Pi - \boldsymbol{u}^\Pi\|_\infty^{(\beta-q)\wedge 1} + \varepsilon_{A,q}^{\mathrm{sm}} \leq C\varepsilon_{A,q}^{\mathrm{sm}}.$$

In both cases, the bounds

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \leq C\sqrt{\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})}, \qquad \sigma_t^q \leq \sigma_{\mathrm{sm}}^q,$$

and the definition of $\varepsilon_{A,q}^{\mathrm{sm}}$ imply that each retained term is bounded by its term budget. After summing the finitely many retained terms and decreasing c_{term}, if necessary, the neural output on the transition region is at most $\varepsilon_{\mathrm{sm}}/3$.

On $\mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, the gate equals 1, and the component approximations give the same bound against the active expansion. Consequently

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}}) \\ (\boldsymbol{x},t) \in \mathcal{K}_i^+(\varepsilon_{\mathrm{sm}})}} \left|\widetilde{R}_i^{\mathrm{sm}}(\boldsymbol{x}, t) - R_{i,\mathrm{app}}^{\mathrm{sm}}(\boldsymbol{x}, t)\right| \leq 2\varepsilon_{\mathrm{sm}}/3. \tag{227}$$

Combining (225) and (227) proves (220).

Step 5: Network size. The projection, residual, Gaussian, coefficient, schedule-power, monomial, gate, and final product subnetworks are all finite in number for fixed d, β. Adding their displayed size bounds, substituting the definition of $\varepsilon_{A,q}^{\mathrm{sm}}$, and enlarging $C_{\mathrm{size,sm}}$ to absorb the fixed coefficient constants gives (224); the additional powers of the retained-term count are absorbed by $D^{\mathfrak{a}_{\mathrm{sm}}}$. The projection-coordinate gate in (226) consists of a fixed chart coordinate bump, one residual-norm computation, and one scalar residual ramp. Consequently its cost is polynomial in D and polylogarithmic in $\varepsilon_{\mathrm{sm}}^{-1}$, and it introduces no $\sigma_{\underline{t}}$-mesh or polynomial lower-noise covering factor. The only remaining non-logarithmic terms are the projection/coefficient Hölder approximation terms recorded in (198), (200) and (212); under $\sigma_t \lesssim \varepsilon_{\mathrm{sm}}^{1/\beta}$, these are bounded by $\varepsilon_{\mathrm{sm}}^{-d/\beta}$ up to powers of $\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$. The geometric projection terms use the smoothness $2(\beta \vee 1)$, while the density coefficient terms use their true regularities $\beta - q$; this is the point that keeps the displayed rate valid also for $0 < \beta < 1$. □

F.4.4. Score approximation on the small-noise regime

Theorem F.27 is the key lower-bound step in the small-noise branch. Without de-Gaussianization, the denominator contains the normal factor $\exp\{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)\}$ and a reciprocal network would have to resolve an exponentially small quantity on the tube. The lemma removes this common factor and proves a polynomial floor, which is what makes the later ratio network stable.

Lemma F.27 (Small-noise de-Gaussianized denominator lower bound)**.** *Assume Theorems 2.2 and 2.5. There are constants $c_Q > 0$ and $C_Q > 0$, depending only on the displayed finite geometry controls and the uniform schedule bounds in (11), such that*

$$c_Q^{-1} \vee C_Q \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}.$$

The following holds. Fix $t \in [\underline{t}, \overline{t}]$, put $H := D \vee \log n$, and let

$$A_t := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le C_\varrho \sigma_t \sqrt{H}\}.$$

For every $\boldsymbol{x} \in A_t$, choose any nearest point

$$\boldsymbol{y}_\star \in \arg\min_{\boldsymbol{y}\in\mathcal{M}} \|\boldsymbol{x} - m_t\boldsymbol{y}\|_2, \qquad \boldsymbol{r}_\star := \boldsymbol{x} - m_t\boldsymbol{y}_\star.$$

Then

$$\exp\left(\frac{\|\boldsymbol{r}_\star\|_2^2}{2\sigma_t^2}\right)\mathcal{Q}(\boldsymbol{x},t) \ge c_Q p_{\min}\big(1 + C_\varrho\sigma_t\sqrt{H}\big)^{-d/2} \ge C_Q^{-1} p_{\min} H^{-d/4}. \tag{228}$$

In particular, if $\boldsymbol{x}/m_t$ lies in the reach tube and $\boldsymbol{y}_\star = \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$, the same bound applies to the de-Gaussianized denominator defined with the projection residual $\boldsymbol{r}(\boldsymbol{x},t) = \boldsymbol{x} - m_t\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$.

Proof. Step 1: choose a chart with uniformly positive amplitude. Because $\mathcal{M}$ is compact, the nearest point $\boldsymbol{y}_\star$ exists. Since the partition of unity is nonnegative and sums to one, there is an index $i_\star$ such that $\rho_{i_\star}(\boldsymbol{y}_\star) \ge C_{\mathcal{M}}^{-1}$. Let $\boldsymbol{u}_\star := \phi_{i_\star}(\boldsymbol{y}_\star)$. The support of $\rho_{i_\star}$ is contained in $U_{i_\star}$, and $\phi_{i_\star}(\mathcal{S}_{i_\star}) \Subset K_{i_\star}^\circ$. Using the finite atlas, the C^1-bounds on $\bar{\rho}_i$, the positive Jacobian lower bound on each compact coordinate box, and the lower bound $\rho_{i_\star}(\boldsymbol{y}_\star) \ge C_{\mathcal{M}}^{-1}$, there are constants $r_Q > 0$ and $c_a > 0$, independent of $(\boldsymbol{x}, t, D, n)$, such that

$$B_{\mathbb{R}^d}(\boldsymbol{u}_\star, r_Q) \subset K_{i_\star}^\circ, \qquad a_{i_\star}^{(\mathcal{Q})}(\boldsymbol{u}) = \bar{\rho}_{i_\star}(\boldsymbol{u}) p_0(\boldsymbol{z}_{i_\star}(\boldsymbol{u}))\sqrt{\det \boldsymbol{G}_{i_\star}(\boldsymbol{u})} \ge c_a p_{\min}$$

for every $\boldsymbol{u} \in B_{\mathbb{R}^d}(\boldsymbol{u}_\star, r_Q)$. By Theorems B.5 and F.3, these constants may be chosen with $r_Q^{-1} \vee c_a^{-1} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}$.

Step 2: compare the phase to the nearest-point phase. Set $i = i_\star$ and $\boldsymbol{z} = \boldsymbol{z}_i$. Since $\boldsymbol{y}_\star$ minimizes $\boldsymbol{y} \mapsto \|\boldsymbol{x} - m_t\boldsymbol{y}\|_2^2$ over the smooth manifold, the first variation vanishes:

$$\boldsymbol{J}_i(\boldsymbol{u}_\star)^\top \boldsymbol{r}_\star = 0.$$

For $\boldsymbol{u} = \boldsymbol{u}_\star + \boldsymbol{\delta} \in B_{\mathbb{R}^d}(\boldsymbol{u}_\star, r_Q)$, Taylor's theorem gives

$$\boldsymbol{z}(\boldsymbol{u}) - \boldsymbol{z}(\boldsymbol{u}_\star) = \boldsymbol{J}_i(\boldsymbol{u}_\star)\boldsymbol{\delta} + \boldsymbol{E}(\boldsymbol{\delta}), \qquad \|\boldsymbol{E}(\boldsymbol{\delta})\|_2 \le C\|\boldsymbol{\delta}\|_2^2.$$

Therefore, using $0 < m_t \le 1$,

$$\begin{aligned} 0 &\le \|\boldsymbol{x} - m_t\boldsymbol{z}(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}_\star\|_2^2 \\ &= 2m_t\langle \boldsymbol{r}_\star, \boldsymbol{y}_\star - \boldsymbol{z}(\boldsymbol{u})\rangle + m_t^2\|\boldsymbol{z}(\boldsymbol{u}) - \boldsymbol{y}_\star\|_2^2 \\ &\le C(1 + \|\boldsymbol{r}_\star\|_2)\|\boldsymbol{\delta}\|_2^2. \end{aligned}$$

The first inequality uses only the nearest-point property of $\boldsymbol{y}_\star$. The last inequality uses the vanished linear term $\langle \boldsymbol{r}_\star, \boldsymbol{J}_i(\boldsymbol{u}_\star)\boldsymbol{\delta}\rangle = 0$ and the uniform C^2-bound on the chart map.

Step 3: integrate over a fixed rescaled coordinate ball. Let $R_Q \in (0, r_Q]$ be fixed and write

$$a_\star := 1 + \|\boldsymbol{r}_\star\|_2, \qquad \boldsymbol{u} = \boldsymbol{u}_\star + \sigma_t a_\star^{-1/2}\boldsymbol{w}, \qquad \|\boldsymbol{w}\|_2 \le R_Q.$$

Since $\sigma_t \le 1$, this coordinate ball is contained in $B_{\mathbb{R}^d}(\boldsymbol{u}_\star, r_Q)$. By Steps 1 and 2,

$$\begin{aligned}\exp\left(\frac{\|\boldsymbol{r}_\star\|_2^2}{2\sigma_t^2}\right)\mathcal{Q}(\boldsymbol{x},t) &\ge \sigma_t^{-d}\int_{B(\boldsymbol{u}_\star,r_Q)} \exp\left(-\frac{\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2-\|\boldsymbol{r}_\star\|_2^2}{2\sigma_t^2}\right) a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\\ &\ge c_a p_{\min} a_\star^{-d/2}\int_{\|\boldsymbol{w}\|_2\le R_Q} e^{-C\|\boldsymbol{w}\|_2^2}\mathrm{d}\boldsymbol{w}\\ &\ge c_Q p_{\min}(1+\|\boldsymbol{r}_\star\|_2)^{-d/2}.\end{aligned}$$

Step 4: specialize to A_t. If $\boldsymbol{x}\in A_t$, then $\|\boldsymbol{r}_\star\|_2 = \mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le C_\varrho\sigma_t\sqrt{H}$, which gives the first bound in (228). Because $H\ge 1$ and $\sigma_t\le 1$, $(1+C_\varrho\sigma_t\sqrt{H})^{d/2}\le C_Q H^{d/4}$ after increasing C_Q. This proves the second bound. □

Lemma F.28 (H-localized de-Gaussianized small-noise chart approximation)**.** *Assume Theorems 2.2 and 2.4. Let $H\ge 1$, $0<\varepsilon_{\mathrm{sm}}\le 1/2$, and suppose*

$$H \ge \log(e/\varepsilon_{\mathrm{sm}}).$$

By the global choice above, $c_\star \ge 2C_\varrho$. Assume that $[t_0,t_1]\subset\mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$ and that, for every $t\in[t_0,t_1]$,

$$3c_\star\frac{\sigma_t}{m_t}\sqrt{H}\le \eta_\Pi/2. \tag{229}$$

For

$$A_t(H) := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M})\le C_\varrho\sigma_t\sqrt{H}\},$$

define the projection residual

$$\boldsymbol{r}(\boldsymbol{x},t) := \boldsymbol{x} - m_t\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t), \qquad (\boldsymbol{x},t)\in A_t(H), \quad t\in[t_0,t_1],$$

which is well-defined by (229). For $R\in\{\mathcal{Q},\mathcal{P}_1,\dots,\mathcal{P}_D\}$, set

$$\bar{R}_i(\boldsymbol{x},t) := \exp\left(\frac{\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right)R_i(\boldsymbol{x},t).$$

In particular, write $\bar{\mathcal{Q}}_i$ and $\bar{\mathcal{P}}_{i,k}$ for the choices $R=\mathcal{Q}$ and $R=\mathcal{P}_k$, and define

$$\bar{\mathcal{Q}} := \sum_i \bar{\mathcal{Q}}_i, \qquad \bar{\mathcal{P}}_k := \sum_i \bar{\mathcal{P}}_{i,k}, \qquad \bar{\mathcal{U}}_{i,k} := \frac{m_t\bar{\mathcal{P}}_{i,k} - x_k\bar{\mathcal{Q}}_i}{\sigma_t}, \qquad \bar{\mathcal{U}}_k := \sum_i \bar{\mathcal{U}}_{i,k}.$$

Then, for every chart i and every such R, there is a ReLU network $\widehat{\bar{R}}^{\mathrm{sm}}_{i,H}$ such that

$$\sup_{\substack{t\in[t_0,t_1]\\ \boldsymbol{x}\in A_t(H)}} |\widehat{\bar{R}}^{\mathrm{sm}}_{i,H}(\boldsymbol{x},t) - \bar{R}_i(\boldsymbol{x},t)| \le C\varepsilon_{\mathrm{sm}}H^C. \tag{230}$$

Moreover, after summing the fixed atlas, there are networks $\widehat{\bar{\mathcal{Q}}}^{\mathrm{sm}}_H$ and $\widehat{\bar{\mathcal{U}}}^{\mathrm{sm}}_H$ satisfying

$$|\widehat{\bar{\mathcal{Q}}}^{\mathrm{sm}}_H(\boldsymbol{x},t) - \bar{\mathcal{Q}}(\boldsymbol{x},t)| \vee \|\widehat{\bar{\mathcal{U}}}^{\mathrm{sm}}_H(\boldsymbol{x},t) - \bar{\mathcal{U}}(\boldsymbol{x},t)\|_\infty \le C\varepsilon_{\mathrm{sm}}H^C \tag{231}$$

uniformly over $t\in[t_0,t_1]$ and $\boldsymbol{x}\in A_t(H)$. All networks in this lemma may be chosen so that their depth, width, sparsity, and log-weight obey the bounds in (224), with the logarithmic factors $\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$ replaced by $H+\log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$, and with the fixed-atlas sum and the D-coordinate parallelization absorbed by the displayed ambient polynomial powers. Hence, when $H = D\vee\log n$, the extra H^C factors are absorbed by the displayed $(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$ powers below.

Proof. Step 1: H-active localization and reach control. For $c_a>0$, define

$$\mathcal{K}_i^{(c_a,H)} := \left\{(\boldsymbol{x},t) : \boldsymbol{x}\in A_t(H),\ \exists \boldsymbol{y}\in\mathcal{S}_i \text{ with } \|\boldsymbol{x}-m_t\boldsymbol{y}\|_2\le c_a\sigma_t\sqrt{H}\right\}.$$

Because $c_\star \geq 2C_\varrho$ and the sets $\mathcal{S}_i$ cover $\mathcal{M}$, every $(\boldsymbol{x}, t) \in A_t(H)$ belongs to $\mathcal{K}_i^{(c_\star, H)}$ for at least one chart. If $(\boldsymbol{x}, t) \in \mathcal{K}_i^{(3c_\star, H)}$, then

$$\operatorname{dist}(\boldsymbol{x}/m_t, \mathcal{M}) \leq 3c_\star \frac{\sigma_t}{m_t}\sqrt{H} \leq \eta_\Pi/2.$$

Thus the nearest projection $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ is unique, its chart coordinate in chart i lies in K_i°, and the Gauss–Newton and buffered-coordinate estimates used earlier hold with $\sqrt{\log(e/\varepsilon_{\mathrm{sm}})}$ replaced by $\sqrt{H}$. Throughout the active part of the proof write

$$\boldsymbol{\xi} := \boldsymbol{x}/m_t, \qquad \boldsymbol{\pi} := \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \qquad \boldsymbol{u}^\Pi := \phi_i(\boldsymbol{\pi}), \qquad \boldsymbol{r} := \boldsymbol{x} - m_t\boldsymbol{\pi}, \qquad \boldsymbol{\nu} := \sigma_t^{-1}\boldsymbol{r}.$$

The active condition gives $\boldsymbol{u}^\Pi \in K_i^\circ$ and $\|\boldsymbol{\nu}\|_2 \leq 3c_\star\sqrt{H}$.

Step 2: de-Gaussianized coercivity. Let $\Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\|_2^2$, and write $S_i^{(R)} := \mathsf{supp} a_i^{(R)}$. The projection orthogonality gives $\nabla_{\boldsymbol{u}}\Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) = 0$. The proof of Theorem F.5, with the H-reach condition above, gives

$$\nabla_{\boldsymbol{u}}^2 \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) \succeq \frac{1}{2}\lambda_i^{(-)}\boldsymbol{I}_d.$$

Hence there are constants $r_i^0 > 0$ and $c_0 > 0$ such that

$$\Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) - \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) \geq c_0\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2^2$$

whenever $\boldsymbol{u} \in \square_i$ and $\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \leq r_i^0$. On the compact set of pairs with $\boldsymbol{u} \in S_i^{(R)}$ and $\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \geq r_i^0$, uniqueness of the reach projection gives a strictly positive phase gap. Combining the local quadratic gap with the compact-away gap and using the bounded diameter of $\square_i$, we obtain the uniform de-Gaussianized lower bound

$$\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u}^\Pi + \sigma_t\boldsymbol{w})\|_2^2 - \|\boldsymbol{r}\|_2^2}{2\sigma_t^2} \geq c\|\boldsymbol{w}\|_2^2$$

for all $\boldsymbol{u}^\Pi + \sigma_t\boldsymbol{w} \in S_i^{(R)}$, after decreasing the fixed small-noise threshold in (197) if necessary. Consequently, for $R_H := A\sqrt{H}$ with A fixed sufficiently large,

$$\int_{\|\boldsymbol{w}\|_2 > R_H} \exp\left(-\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u}^\Pi + \sigma_t\boldsymbol{w})\|_2^2 - \|\boldsymbol{r}\|_2^2}{2\sigma_t^2}\right)(1 + \|\boldsymbol{w}\|_2)^C \mathrm{d}\boldsymbol{w} \leq C\varepsilon_{\mathrm{sm}}H^{-C},$$

where we used $H \geq \log(e/\varepsilon_{\mathrm{sm}})$ and increased A to dominate the finitely many polynomial powers used below.

Step 3: scalar Taylor expansion after removing the normal Gaussian. For $s \in [0, \sigma_t]$, define the smooth difference quotient

$$\boldsymbol{Z}_s(\boldsymbol{u}^\Pi, \boldsymbol{w}) := \int_0^1 \boldsymbol{J}_i(\boldsymbol{u}^\Pi + \theta s\boldsymbol{w})\boldsymbol{w}\mathrm{d}\theta,$$

so that $\boldsymbol{z}_i(\boldsymbol{u}^\Pi + s\boldsymbol{w}) - \boldsymbol{z}_i(\boldsymbol{u}^\Pi) = s\boldsymbol{Z}_s(\boldsymbol{u}^\Pi, \boldsymbol{w})$ and $\boldsymbol{Z}_0(\boldsymbol{u}^\Pi, \boldsymbol{w}) = \boldsymbol{J}_i(\boldsymbol{u}^\Pi)\boldsymbol{w}$. Define the de-Gaussianized phase along this scalar path by

$$\Theta_s(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) := \frac{1}{2}\|\boldsymbol{\nu} - m_t\boldsymbol{Z}_s(\boldsymbol{u}^\Pi, \boldsymbol{w})\|_2^2 - \frac{1}{2}\|\boldsymbol{\nu}\|_2^2.$$

At $s = \sigma_t$, this is exactly the exponent in the de-Gaussianized chart integral, and at $s = 0$ it equals $\frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{G}_i(\boldsymbol{u}^\Pi)\boldsymbol{w}$ because $\boldsymbol{J}_i(\boldsymbol{u}^\Pi)^\top\boldsymbol{\nu} = 0$. We now verify that the same coercivity holds uniformly along this interpolation. For $s = 0$, ellipticity of $\boldsymbol{G}_i(\boldsymbol{u}^\Pi)$ gives

$$\Theta_0(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) \geq c\|\boldsymbol{w}\|_2^2.$$

For $0 < s \leq \sigma_t$, set $\boldsymbol{x}_s := m_t\boldsymbol{\pi} + s\boldsymbol{\nu}$. Since $\|\boldsymbol{\nu}\|_2 \leq 3c_\star\sqrt{H}$ and (229) holds,

$$\operatorname{dist}(\boldsymbol{x}_s/m_t, \mathcal{M}) \leq s\|\boldsymbol{\nu}\|_2/m_t \leq 3c_\star\sigma_t m_t^{-1}\sqrt{H} \leq \eta_\Pi/2.$$

Moreover $\boldsymbol{\pi} = \Pi_{\mathcal{M}}(\boldsymbol{x}_s/m_t)$, because $\boldsymbol{x}_s/m_t$ lies on the normal segment issued from $\boldsymbol{\pi}$ and remains inside the reach tube. Applying the reach-tube objective gap Theorem F.13 to the pair $\boldsymbol{x}_s/m_t$ and $\boldsymbol{z}_i(\boldsymbol{u}^\Pi + s\boldsymbol{w})$ gives

$$\|\boldsymbol{x}_s - m_t\boldsymbol{z}_i(\boldsymbol{u}^\Pi + s\boldsymbol{w})\|_2^2 - \|\boldsymbol{x}_s - m_t\boldsymbol{\pi}\|_2^2 \geq c\,m_t^2\|\boldsymbol{z}_i(\boldsymbol{u}^\Pi + s\boldsymbol{w}) - \boldsymbol{\pi}\|_2^2.$$

On the compact coordinate box, the inverse chart is uniformly bi-Lipschitz, hence

$$\|\boldsymbol{z}_i(\boldsymbol{u}^\Pi + s\boldsymbol{w}) - \boldsymbol{\pi}\|_2 \geq c\, s\|\boldsymbol{w}\|_2 \qquad \text{whenever } \boldsymbol{u}^\Pi + s\boldsymbol{w} \in \square_i.$$

Dividing the previous objective gap by $2s^2$ gives

$$\Theta_s(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) = \frac{\|\boldsymbol{x}_s - m_t\boldsymbol{z}_i(\boldsymbol{u}^\Pi + s\boldsymbol{w})\|_2^2 - \|\boldsymbol{x}_s - m_t\boldsymbol{\pi}\|_2^2}{2s^2} \geq c\|\boldsymbol{w}\|_2^2.$$

Thus $\Theta_s(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) \geq c\|\boldsymbol{w}\|_2^2$ on the chart support for every $s \in [0, \sigma_t]$. For $R \in \{\mathcal{Q}, \mathcal{P}_1, \ldots, \mathcal{P}_D\}$, set

$$G_R(s, \boldsymbol{w}) := \exp\{-\Theta_s(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)\}a_i^{(R)}(\boldsymbol{u}^\Pi + s\boldsymbol{w}).$$

The coercivity from Step 2 and the smoothness of $\boldsymbol{z}_i$ imply that all derivatives in s falling on the geometric and exponential factors are bounded by

$$CH^C(1 + \|\boldsymbol{w}\|_2)^C e^{-c\|\boldsymbol{w}\|_2^2}$$

for $0 \leq s \leq \sigma_t$ and $\|\boldsymbol{w}\|_2 \leq R_H$. The density-dependent amplitude is expanded only to order $\lfloor\beta\rfloor$, and the remainder is controlled by the $\mathcal{H}^\beta$ modulus of $a_i^{(R)}$ along the line $\boldsymbol{u}^\Pi + s\boldsymbol{w}$. Taylor's theorem in the scalar variable s therefore gives

$$G_R(\sigma_t, \boldsymbol{w}) = \sum_{q=0}^{\lfloor\beta\rfloor} \sigma_t^q G_{R,q}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \mathcal{E}_R(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t),$$

with

$$|\mathcal{E}_R(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)| \leq CH^C(1 + \|\boldsymbol{w}\|_2)^C e^{-c\|\boldsymbol{w}\|_2^2}.$$

Each coefficient $G_{R,q}$ is a finite sum of a polynomial in $(\boldsymbol{\nu}, \boldsymbol{w})$, a smooth geometric factor in $\boldsymbol{u}^\Pi$, a bounded schedule power of m_t, and a derivative $\partial^\alpha a_i^{(R)}(\boldsymbol{u}^\Pi)$ with $|\alpha| \leq q$. After integration in $\boldsymbol{w}$, the coefficients agree with the $A_{i,q,\boldsymbol{\lambda}}^{(R)}$ defined in Theorem F.16, because both constructions differentiate the same recentered integrand after the common normal Gaussian factor has been removed. Integrating the Taylor expansion on $\|\boldsymbol{w}\|_2 \leq R_H$ and adding the tail from Step 2 yields

$$\bar{R}_i(\boldsymbol{x}, t) = \sum_{q=0}^{\lfloor\beta\rfloor} \sigma_t^q \sum_{\|\boldsymbol{\lambda}\|_1 \leq q} A_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}^\Pi, t)\boldsymbol{\nu}^{\boldsymbol{\lambda}} + \mathrm{Rem}_{i,H}^{(R)}(\boldsymbol{x}, t), \qquad |\mathrm{Rem}_{i,H}^{(R)}(\boldsymbol{x}, t)| \leq C\sigma_t^\beta H^C.$$

Since $t \in [t_0, t_1] \subset \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$, this remainder is bounded by $C\varepsilon_{\mathrm{sm}}H^C$.

Step 4: centered numerator expansion without a σ_t^{-1} loss. For each coordinate k, define

$$N_k(s, \boldsymbol{w}) := -\nu_k + m_t(\boldsymbol{Z}_s(\boldsymbol{u}^\Pi, \boldsymbol{w}))_k.$$

At $s = \sigma_t$, this satisfies

$$N_k(\sigma_t, \boldsymbol{w}) = \frac{m_t z_{i,k}(\boldsymbol{u}^\Pi + \sigma_t\boldsymbol{w}) - x_k}{\sigma_t}.$$

Thus $\bar{\mathcal{U}}_{i,k}$ is the integral of $N_k(\sigma_t, \boldsymbol{w})G_{\mathcal{Q}}(\sigma_t, \boldsymbol{w})$. Applying the same scalar Taylor argument to the product $N_k(s, \boldsymbol{w})G_{\mathcal{Q}}(s, \boldsymbol{w})$ gives

$$\bar{\mathcal{U}}_{i,k}(\boldsymbol{x}, t) = \sum_{q=0}^{\lfloor\beta\rfloor} \sigma_t^q \sum_{\|\boldsymbol{\lambda}\|_1 \leq q+1} A_{i,k,q,\boldsymbol{\lambda}}^{(\bar{U})}(\boldsymbol{u}^\Pi, t)\boldsymbol{\nu}^{\boldsymbol{\lambda}} + \mathrm{Rem}_{i,k,H}^{(\bar{U})}(\boldsymbol{x}, t),$$

where $|\mathrm{Rem}_{i,k,H}^{(\bar{U})}(\boldsymbol{x}, t)| \leq C\varepsilon_{\mathrm{sm}}H^C$. For clarity, we verify the centered-numerator coefficient analogue used by the network construction. Differentiating $N_k(s, \boldsymbol{w})G_{\mathcal{Q}}(s, \boldsymbol{w})$ at $s = 0$ at most q times produces finite sums of terms of the form

$$m_t^a P(\boldsymbol{\nu}, \boldsymbol{w})H(\boldsymbol{u}^\Pi)\partial^\alpha a_i^{(\mathcal{Q})}(\boldsymbol{u}^\Pi)\exp\left(-\frac{m_t^2}{2}\boldsymbol{w}^\top\boldsymbol{G}_i(\boldsymbol{u}^\Pi)\boldsymbol{w}\right), \qquad |\alpha| \leq q,$$

where P is a polynomial whose $\boldsymbol{\nu}$-degree is at most $q+1$; the extra degree is exactly the leading centered factor $-\nu_k$ in N_k. After integrating in $\boldsymbol{w}$, the Gaussian moment formula used in Theorem F.19 expresses each coefficient $A^{(\bar{U})}_{i,k,q,\boldsymbol{\lambda}}$ as a finite sum of powers of m_t times products of smooth geometric factors and derivatives $\partial^\alpha a_i^{(\mathcal{Q})}$ with $|\alpha| \le q$. Therefore, on the same intermediate box $K_i^\circ \Subset \square_i^A \Subset \square_i$, define the extension by this identical finite formula. If $\boldsymbol{u} \notin S_i^{(\mathcal{Q})} = \mathsf{supp} a_i^{(\mathcal{Q})}$, then $a_i^{(\mathcal{Q})}$ vanishes in a neighborhood of $\boldsymbol{u}$, so all derivatives $\partial^\alpha a_i^{(\mathcal{Q})}(\boldsymbol{u})$ with $|\alpha| \le q$ vanish; the centered extension is therefore support-preserving. Since $a_i^{(\mathcal{Q})} \in \mathcal{H}^\beta$ and only derivatives up to order q are used, the extended coefficient belongs to $\mathcal{H}^{\beta-q}(\square_i^A)$ with norm bounded by the same finite geometry and density factors as in Theorem F.19. Applying the coefficient network construction of Theorem F.20 to this finite family gives the same depth, width, sparsity, and log-weight bounds, up to one additional fixed power of D for the coordinate index k and the extra normal monomial. This power is absorbed into $\mathfrak{p}_{\beta,d}$. This direct centered expansion is the reason the proof never forms $(m_t\bar{\mathcal{P}} - x\bar{\mathcal{Q}})/\sigma_t$ from two separately approximated networks.

Step 5: certified chart gate and network realization. Fix one intermediate coefficient box $\square_i^A$ with $K_i^\circ \Subset \square_i^A \Subset \square_i$, and use the coefficient-extension construction on this common box for the finite family of retained terms. Choose the coordinate tolerance $\varepsilon_u^{\mathrm{cert}}$ as a fixed constant multiple of the smallest coordinate tolerance required by these retained coefficient networks. Apply Theorem F.24 with $c_a = 3c_\star$, $\varepsilon = \varepsilon_{\mathrm{sm}}$, $\eta = \varepsilon_u^{\mathrm{cert}}$, and this box $\square_i^A$. It gives networks Γ_i^H, $\breve{\boldsymbol{u}}_i$, and $\breve{\boldsymbol{u}}_i^A$ such that $\Gamma_i^H = 1$ and $\breve{\boldsymbol{u}}_i^A$ is an accurate coefficient coordinate on $\mathcal{K}_i^{(3c_\star,H)}$, while $\Gamma_i^H > 0$ anywhere in $A_t(H)$ implies that $\breve{\boldsymbol{u}}_i$ is a certified chart-i coordinate for $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$. For the coefficient functions in Steps 3 and 4, use the extensions from Theorem F.19 and the centered-numerator analogues verified in Step 4, evaluate them at the clipped certified coordinate $\breve{\boldsymbol{u}}_i^A$, and multiply the retained expansion by Γ_i^H. Use the global normal-residual network supplied by Theorem F.23 on the active union $\bigcup_j \mathcal{K}_j^{(3c_\star,H)}$, which contains $A_t(H)$ by Step 1. Allocate accuracy $c\varepsilon_{\mathrm{sm}}H^{-C}$ to each coefficient, monomial, schedule power, product, normal-residual operation, certified chart gate, and final affine summation. On $\mathcal{K}_i^{(3c_\star,H)}$ the gate equals one, the coordinate and normal-residual errors are certified, and the Taylor expansions from Steps 3 and 4 are approximated with total error $C\varepsilon_{\mathrm{sm}}H^C$. The resulting size bounds are the same as those in (224), with every logarithmic scale enlarged to $H + \log(C_{\mathrm{size,sm}}\varepsilon_{\mathrm{sm}}^{-1})$, because the certified objective gate reuses the global projection objective networks and adds only fixed-atlas reciprocals, ramps, products, and sums.

Step 6: inactive and transition regions after de-Gaussianization. If $(\boldsymbol{x},t) \notin \mathcal{K}_i^{(3c_\star,H)}$ and $\boldsymbol{u} \in S_i^{(R)}$, then

$$\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2 > 3c_\star\sigma_t\sqrt{H}.$$

Since $\boldsymbol{x} \in A_t(H)$ gives $\|\boldsymbol{r}(\boldsymbol{x},t)\|_2 \le C_\varrho\sigma_t\sqrt{H}$ and $c_\star \ge 2C_\varrho$, we have

$$\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2 \ge cc_\star^2\sigma_t^2 H.$$

The de-Gaussianized inactive integrand is therefore bounded by $e^{-cc_\star^2 H}$ times a fixed polynomial in H. The global choice of $c_\star$ makes the true inactive contribution bounded by $C\varepsilon_{\mathrm{sm}}H^C$. It remains to control the implemented chart output on the same inactive region. If $\Gamma_i^H = 0$, the implemented chart output is zero. If $\Gamma_i^H > 0$, the certification in Step 5 gives a valid chart coordinate $\boldsymbol{u}_i^\Pi = \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t))$ and $\|\breve{\boldsymbol{u}}_i - \boldsymbol{u}_i^\Pi\|_\infty \le \varepsilon_u^{\mathrm{cert}}$. Because $\boldsymbol{x} \in A_t(H)$ and $c_\star \ge 2C_\varrho$, the condition $\boldsymbol{u}_i^\Pi \in S_i^{(R)}$ would imply $(\boldsymbol{x},t) \in \mathcal{K}_i^{(c_\star,H)} \subset \mathcal{K}_i^{(3c_\star,H)}$, contradicting the inactive assumption. Hence $\boldsymbol{u}_i^\Pi \notin S_i^{(R)}$ whenever the inactive-region gate is nonzero. If $\boldsymbol{u}_i^\Pi \in \square_i^A$, the support-preserving extension gives $\overline{A}^{(R)}_{i,q,\boldsymbol{\lambda}}(\boldsymbol{u}_i^\Pi,t) = 0$, and the Hölder estimate plus the 1-Lipschitz clipping to $\square_i^A$ gives $|\overline{A}^{(R)}_{i,q,\boldsymbol{\lambda}}(\breve{\boldsymbol{u}}_i^A,t)| \le C(\varepsilon_u^{\mathrm{cert}})^{(\beta-q)\wedge 1}$. If $\boldsymbol{u}_i^\Pi \notin \square_i^A$, the fixed positive separation of $S_i^{(R)} \subset K_i^\circ \Subset \square_i^A$ from $\partial\square_i^A$ and the choice of $\varepsilon_u^{\mathrm{cert}}$ ensure that the clipped coefficient input remains in $\square_i^A \setminus S_i^{(R)}$, so the same bound holds with the intrinsic coefficient-network accuracy alone. Choosing $\varepsilon_u^{\mathrm{cert}}$ according to the coefficient tolerances in Step 5 makes every retained inactive term at most its allocated $c\varepsilon_{\mathrm{sm}}H^{-C}$ budget after multiplication by σ_t^q and the $H^{q/2}$ normal-monomial bound. Combining the active approximation, the true inactive tail, and the certified off-support network control proves (230). Summing over the fixed atlas and using the direct centered-numerator networks from Step 4 proves (231). The displayed network-size statement follows from the size bounds in Step 5, the fixed atlas summation, and the coordinatewise parallelization for $\bar{\mathcal{U}}$. □

Corollary F.29 (Score approximation on small noise regime $[t_0,t_1]$)**.** *Assume Theorems 2.2, 2.4 and 2.5, and suppose $D \vee p_{\min}^{-1} \le n^{a_0}$. Put $H := D \vee \log n$, let q_{geom} be as in (1), and, after this finite order is fixed, take*

$$\mathfrak{p}_{\beta,d} = C + \max\left\{8(q_{\mathrm{geom}}+2), \left\lceil \lfloor\beta\rfloor + 1 + \frac{\lfloor\beta\rfloor d}{\beta - \lfloor\beta\rfloor} + \frac{d}{\beta\vee 1} + \frac{d}{2} + 6\right\rceil\right\},$$

where the C in the exponent is universal and independent of $D, n, \sigma_t, p_{\min}$, and the numerical geometry constants. Thus $\mathfrak{p}_{\beta,d}$ may depend on the fixed finite order q_{geom}; in the main $d > 2$ regime, choosing $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$ yields $q_{\mathrm{geom}} = \mathcal{O}_\beta(1)$ and hence $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$. Let $C_{\mathrm{den,sm}} \geq 1$ be a fixed constant large enough to dominate the reciprocal and product constants in the small-noise ratio step. Assume the small-noise reach condition

$$3c_\star \underline{m}^{-1}\sigma_{t_1}\sqrt{H} \leq \eta_\Pi/2 \tag{232}$$

and the small-noise denominator-resolution condition

$$C_{\mathrm{den,sm}} p_{\min}^{-1}\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C n^{-\beta/(d+2\beta)} H^{\mathfrak{p}_{\beta,d}/2-1/2} \leq 1. \tag{233}$$

For the VP/OU schedules used here, $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$, so (232) is implied by

$$D \vee \log n \leq c\,\underline{m}^2 \eta_\Pi^2\, n^{2/(d+2\beta)}.$$

There exists a switched ReLU score network $s^{\mathrm{lb}}_{\mathrm{sm}}$ on $[t_0, t_1]$ such that for

$$A_t := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \leq C_\varrho \sigma_t\sqrt{H}\},$$

$$\|\sigma_t s^{\mathrm{lb}}_{\mathrm{sm}}(\boldsymbol{x},t) - \sigma_t s^*(\boldsymbol{x},t)\|_2 \leq C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}/2}, \quad \boldsymbol{x}\in A_t,\ \ t\in[t_0,t_1], \tag{234}$$

or equivalently,

$$\|s^{\mathrm{lb}}_{\mathrm{sm}}(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_2^2 \leq C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}\sigma_t^{-2}, \quad \boldsymbol{x}\in A_t, \quad t\in[t_0,t_1]. \tag{235}$$

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|s^{\mathrm{lb}}_{\mathrm{sm}}(\boldsymbol{x},t)\|_\infty \leq C\sigma_t^{-1}(D\vee\log n)^{1/2}, \qquad t\in[t_0,t_1], \tag{236}$$

and

$$L,\ \log B \leq C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}, \tag{237}$$

and

$$\|\boldsymbol{W}\|_\infty,\ S \leq C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}. \tag{238}$$

Proof. Step 1: de-Gaussianize the chart integrals. On the tube

$$A_t := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \leq C_\varrho \sigma_t\sqrt{H}\},$$

the reach condition (232) gives

$$\mathrm{dist}(\boldsymbol{x}/m_t, \mathcal{M}) \leq C_\varrho \frac{\sigma_t}{m_t}\sqrt{H} \leq \eta_\Pi/2.$$

Thus $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ is the unique metric projection, and we may define

$$\boldsymbol{r}(\boldsymbol{x},t) := \boldsymbol{x} - m_t\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t).$$

Define the de-Gaussianized denominator and centered numerator by

$$\bar{\mathcal{Q}}(\boldsymbol{x},t) := \exp\left(\frac{\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right)\mathcal{Q}(\boldsymbol{x},t), \qquad \bar{\mathcal{U}}(\boldsymbol{x},t) := \exp\left(\frac{\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right)\mathcal{U}(\boldsymbol{x},t). \tag{239}$$

Equivalently, for $R \in \{\mathcal{Q}, \mathcal{P}_1, \ldots, \mathcal{P}_D\}$, define

$$\bar{R}_i(\boldsymbol{x},t) := \exp\left(\frac{\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right) R_i(\boldsymbol{x},t). \tag{240}$$

Using (48), this removes exactly the common normal Gaussian factor:

$$\bar{R}_i(\boldsymbol{x},t) = \sigma_t^{-d}\int_{K_i^{\mathrm{cell}}} \exp\left(-\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right) a_i^{(R)}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}. \tag{241}$$

Thus all exponential factors in the lower-bound small-noise approximation are displayed explicitly: the original chart Gaussian

$$\exp\left(-\frac{\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\right),$$

the common de-Gaussianizing multiplier

$$\exp\left(\frac{\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right),$$

and their product

$$\exp\left(-\frac{\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2-\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\right).$$

Consequently,

$$\bar{\mathcal{Q}}=\sum_i\bar{\mathcal{Q}}_i,\qquad \bar{\mathcal{P}}_k=\sum_i\bar{\mathcal{P}}_{i,k},\qquad \bar{\mathcal{U}}_k=\frac{m_t\bar{\mathcal{P}}_k-x_k\bar{\mathcal{Q}}}{\sigma_t}. \tag{242}$$

The common exponential factor cancels, so

$$\frac{\bar{\mathcal{U}}(\boldsymbol{x},t)}{\bar{\mathcal{Q}}(\boldsymbol{x},t)}=\frac{\mathcal{U}(\boldsymbol{x},t)}{\mathcal{Q}(\boldsymbol{x},t)}=\sigma_t s^*(\boldsymbol{x},t)\qquad(\boldsymbol{x}\in A_t).$$

Step 2: lower-bound the de-Gaussianized denominator. By Theorem F.27, the density lower bound gives the following denominator floor after de-Gaussianization:

$$\bar{\mathcal{Q}}(\boldsymbol{x},t)\ge C^{-1}\Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}}p_{\min}(D\vee\log n)^{-d/4},\qquad(\boldsymbol{x},t)\in A_t,\quad t\in[t_0,t_1], \tag{243}$$

The factor $(D\vee\log n)^{d/4}$ in the reciprocal is absorbed into the $(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$ part of the explicit ambient polynomial below, and the displayed $\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}}$ factor is carried with the other geometry terms. The centered-ratio bound from (64), used with $\underline{t}=t_0$ and the assumption $p_{\min}^{-1}\le n^{a_0}$, gives

$$\|\bar{\mathcal{U}}(\boldsymbol{x},t)\|_2\le CH^{1/2}\bar{\mathcal{Q}}(\boldsymbol{x},t)\qquad((\boldsymbol{x},t)\in A_t). \tag{244}$$

Step 3: approximate the de-Gaussianized numerator and denominator. Set $a_\beta:=\beta/(d+2\beta)$ and $\varepsilon_{\mathrm{sm}}:=\frac{1}{2}\wedge e\,n^{-a_\beta}$. If $\varepsilon_{\mathrm{sm}}=e\,n^{-a_\beta}$, then $\log(e/\varepsilon_{\mathrm{sm}})=a_\beta\log n\le H$. If $\varepsilon_{\mathrm{sm}}=1/2$, then $\log(e/\varepsilon_{\mathrm{sm}})=\log(2e)\le H$ because $d\ge1$, $d<D$, and hence $D\ge2$. Use Theorem F.28 with $\varepsilon_{\mathrm{sm}}$ and $H=D\vee\log n$. The inclusion $[t_0,t_1]\subset\mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$ follows from the choice of the small-noise endpoint t_1. The reach condition needed by that lemma is exactly (232). This is the point where the de-Gaussianized construction uses H-active charts rather than the smaller $\sqrt{\log(e/\varepsilon_{\mathrm{sm}})}$-active charts: for every $\boldsymbol{x}\in A_t$, a nearest chart is active, while charts outside the $3c_\star\sqrt{H}$-enlargement have

$$\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2-\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2\ge c\sigma_t^2H$$

on their supports and are exponentially negligible after de-Gaussianization. After summing over charts, this gives networks $\widehat{\bar{\mathcal{Q}}}$ and $\widehat{\bar{\mathcal{U}}}$ satisfying

$$|\widehat{\bar{\mathcal{Q}}}(\boldsymbol{x},t)-\bar{\mathcal{Q}}(\boldsymbol{x},t)|\vee\|\widehat{\bar{\mathcal{U}}}(\boldsymbol{x},t)-\bar{\mathcal{U}}(\boldsymbol{x},t)\|_\infty\le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}}B_0^{C}n^{-\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}/2-d/4-1/2} \tag{245}$$

on A_t, uniformly in $t\in[t_0,t_1]$. Here the universal constant in the definition of $\mathfrak{p}_{\beta,d}$ is chosen large enough that the H^C loss in (231) is dominated by $(D\vee\log n)^{\mathfrak{p}_{\beta,d}/2-d/4-1/2}$. The size is bounded by the right-hand sides of (237) and (238): in (224) we use $\varepsilon_{\mathrm{sm}}=\frac{1}{2}\wedge e\,n^{-\beta/(d+2\beta)}$, so $\varepsilon_{\mathrm{sm}}^{-d/\beta}\lesssim n^{d/(d+2\beta)}$. The extra $\log\underline{t}^{-1}$ from (222) is absorbed into $(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$ because $\underline{t}=t_0\asymp n^{-2(\beta+1)/(d+2\beta)}$.

Step 4: form the stable ratio. By (243),

$$\bar{\mathcal{Q}}(\boldsymbol{x},t)^{-1}\le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}}p_{\min}^{-1}(D\vee\log n)^{d/4}\qquad((\boldsymbol{x},t)\in A_t).$$

The denominator-resolution condition (233) and (245) imply

$$C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C n^{-\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}/2-d/4-1/2} \le c\Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min} (D \vee \log n)^{-d/4}/4,$$

by the fixed choice of $C_{\mathrm{den,sm}}$. Hence $\widehat{\widehat{\mathcal{Q}}} \vee c\Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min}(D \vee \log n)^{-d/4}/2 \ge c\Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min}(D \vee \log n)^{-d/4}/2$, and the standard reciprocal and product networks applied to $\widehat{\widehat{\mathcal{U}}}/(\widehat{\widehat{\mathcal{Q}}} \vee c\Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min}(D \vee \log n)^{-d/4}/2)$, together with (244) and (245), give a network $\widehat{v}^{\mathrm{lb}}_{\mathrm{sm}}$ such that

$$\|\widehat{v}^{\mathrm{lb}}_{\mathrm{sm}}(\boldsymbol{x},t) - \sigma_t s^*(\boldsymbol{x},t)\|_2 \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}/2}, \qquad (\boldsymbol{x},t) \in A_t.$$

By (64), with $\underline{t} = t_0$, the polynomial assumption $D \vee p_{\min}^{-1} \le n^{a_0}$, and $\sigma_{t_0}^{-1} \le n^C$, the true score obeys

$$\sigma_t \|s^*(\boldsymbol{x},t)\|_\infty \le \sigma_t \|s^*(\boldsymbol{x},t)\|_2 \le C(D \vee \log n)^{1/2}, \qquad (\boldsymbol{x},t) \in A_t.$$

Set $s^{\mathrm{lb}}_{\mathrm{sm}} := \sigma_t^{-1} \widehat{v}^{\mathrm{lb}}_{\mathrm{sm}}$, with coordinatewise clipping at $C\sigma_t^{-1}(D \vee \log n)^{1/2}$. On A_t, (234) and (235) follow. Coordinatewise clipping is 1-Lipschitz and fixes the target coordinatewise on A_t, so it does not increase the tube approximation error. The clipping construction gives (236).

Step 5: record size. The reciprocal, products, clipping, and time switch add only polynomial factors in D, $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$, B_0, $p_{\min}^{-1}$, and polylogarithmic factors in n. The scalar reciprocal and clipping layers only affect $L + \log B$ by polylogarithmic factors and do not change the leading intrinsic width/sparsity term. Hence (237) and (238) follow. □

G. Score Estimation on Compact Smooth Manifolds

This appendix proves the learned-score bounds used by the distribution estimation results in Sec. H. The clipped neural network class $\widetilde{\mathcal{NN}}$, the clipped score-estimation oracle inequality, and the covering-number evaluation are stated in the main text in Sec. 4. Here we apply those tools to the density-lower-bound small-noise and large-noise branches. On a large-noise slab $I_k = [t_{k-1}, t_k]$, the tangent-cell class uses cells of radius $r_k = c_0 \sigma_{t_{k-1}} \wedge r_{\star,n}$. The cap $r_{\star,n}$, defined below, is the accuracy-limited radius needed to make the fixed-order tangent-cell expansion uniform also on order-one-noise slabs. The dyadic sum of the $\sigma_{t_{k-1}}^{-d}$ terms is dominated by the first large-noise slab; the additional $r_{\star,n}^{-d}$ contribution is tracked explicitly in the oracle and W_1 bounds. The small-noise branch uses the projection–Laplace class under Theorem 2.5. In all size bounds in this section, the factor $\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}$ may be read as

$$C(1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}} \mathcal{K}^{\mathfrak{a}_{\mathrm{geom}}}_{\mathcal{M},q_{\mathrm{geom}}}$$

using the reach-explicit finite-geometry convention.

The section has one generic tool and two applications. The oracle inequality Theorem G.1 is the only statistical argument; the large-noise and small-noise results then insert the deterministic comparison classes from Theorems E.5 and F.29 and evaluate their entropy sizes. Readers interested only in the final rates can read Theorems G.2 and G.3 first and return to the oracle proof for concentration details.

G.1. Oracle Inequality for Score Estimation

Theorem G.1 (Oracle inequality for clipped score estimation)**.** *Let $n^{-\mathcal{O}(1)} \le \underline{t} < \overline{t} \le n^{\mathcal{O}(1)}$, $D \le n^{\mathcal{O}(1)}$, and fix an arbitrary $A > 0$. Let $\widehat{s}$ be an empirical risk minimizer over the neural network class $\mathcal{NN}_{\underline{t}} \equiv \widetilde{\mathcal{NN}}(L_{\underline{t}}, \boldsymbol{W}_{\underline{t}}, S_{\underline{t}}, B_{\underline{t}})$ for the interval-restricted denoising loss $\ell_{[\underline{t},\overline{t}]}$, namely*

$$\widehat{s} \in \operatorname*{arg\,min}_{s \in \mathcal{NN}_{\underline{t}}} \frac{1}{n} \sum_{i=1}^{n} \ell_{[\underline{t},\overline{t}]}(s, \boldsymbol{x}^{(i)}).$$

The same conclusion holds for any measurable approximate empirical minimizer with empirical excess at most n^{-A}, after increasing the negligible remainder. Then the expected excess risk satisfies

$$\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n}\Big[\int_{\underline{t}}^{\overline{t}} \mathbb{E}_{\boldsymbol{X}_t}[\|\widehat{s}(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2]\mathrm{d}t\Big] \\
&\lesssim \inf_{s \in \mathcal{NN}_{\underline{t}}} \int_{\underline{t}}^{\overline{t}} \mathbb{E}_{\boldsymbol{X}_t}\big[\|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\big]\mathrm{d}t + \frac{(D \vee \log n)^2 \int_{\underline{t}}^{\overline{t}} \sigma_t^{-2}\mathrm{d}t \log\big(\mathcal{N}(n^{-\mathcal{O}(1)}, \mathcal{NN}_{\underline{t}}, \|\cdot\|_\infty) \vee n\big)}{n} + n^{-A}.
\end{aligned}$$

Proof. Step 1: reduce the denoising risk to the marginal score risk. Write

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}^{(i)})$$

for the empirical average over the training sample. For $s \in \mathcal{NN}_{\underline{t}}$, set

$$\mathcal{L}_{[\underline{t},\overline{t}]}(s) := \int_{\underline{t}}^{\overline{t}} \mathbb{E}_{\boldsymbol{X}_t}\|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2 \mathrm{d}t.$$

The denoising-score identity, applied on the interval $[\underline{t},\overline{t}]$, gives

$$\mathcal{L}_{[\underline{t},\overline{t}]}(s) = \mathbb{E}_{\boldsymbol{X}_0}\{\ell_{[\underline{t},\overline{t}]}(s,\boldsymbol{X}_0) - \ell_{[\underline{t},\overline{t}]}(s^*,\boldsymbol{X}_0)\}.$$

Since the last term is independent of s, the empirical minimizer of $\mathbb{P}_n \ell_{[\underline{t},\overline{t}]}(s)$ is also an empirical minimizer of $\mathbb{P}_n\{\ell_{[\underline{t},\overline{t}]}(s) - \ell_{[\underline{t},\overline{t}]}(s^*)\}$.

Step 2: record the envelope and variance bounds. The clipped definition of $\widetilde{\mathcal{NN}}$ and Theorem I.12 imply

$$\sup_{s,s'\in\mathcal{NN}_{\underline{t}}}\sup_{\boldsymbol{x}\in\mathbb{R}^D}|\ell_{[\underline{t},\overline{t}]}(s,\boldsymbol{x})-\ell_{[\underline{t},\overline{t}]}(s',\boldsymbol{x})|\le C(D\vee\log n)^2\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t.$$

Moreover, applying Theorem I.11 to $s,s'\in\mathcal{NN}_{\underline{t}}$, using the same clipped envelope, and then averaging over $\boldsymbol{X}_0$, gives

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}_0}\left[\{\ell_{[\underline{t},\overline{t}]}(s,\boldsymbol{X}_0)-\ell_{[\underline{t},\overline{t}]}(s',\boldsymbol{X}_0)\}^2\right]\\&\le C(D\vee\log n)^2\left(\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t\right)\int_{\underline{t}}^{\overline{t}}\mathbb{E}_{\boldsymbol{X}_t}\|s(\boldsymbol{X}_t,t)-s'(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t\\&\le C(D\vee\log n)^2\left(\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t\right)\{\mathcal{L}_{[\underline{t},\overline{t}]}(s)+\mathcal{L}_{[\underline{t},\overline{t}]}(s')\}.\end{aligned}\tag{246}$$

Here the last inequality is the elementary bound $\|s-s'\|_2^2\le 2\|s-s^*\|_2^2+2\|s'-s^*\|_2^2$.

Step 3: apply Bernstein on a finite cover. Let $\eta=n^{-A_0}$, with A_0 a sufficiently large fixed constant, and let $\mathcal{N}_\eta$ be a minimal η-net of $\mathcal{NN}_{\underline{t}}$ in the supremum norm after the input truncation and affine rescaling specified in Sec. 4. Because all functions in $\mathcal{NN}_{\underline{t}}$ are realized after this input truncation, this is also a global η-net for the clipped functions on $\mathbb{R}^D\times[\underline{t},\overline{t}]$. For fixed $u,v\in\mathcal{N}_\eta$, Bernstein's inequality and (246) give, with probability at least $1-e^{-r}$,

$$\begin{aligned}&|(\mathbb{P}_n-\mathbb{E}_{\boldsymbol{X}_0})\{\ell_{[\underline{t},\overline{t}]}(u)-\ell_{[\underline{t},\overline{t}]}(v)\}|\\&\le C\left[\sqrt{\frac{(D\vee\log n)^2\left(\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t\right)\{\mathcal{L}_{[\underline{t},\overline{t}]}(u)+\mathcal{L}_{[\underline{t},\overline{t}]}(v)\}r}{n}}+\frac{(D\vee\log n)^2\left(\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t\right)r}{n}\right].\end{aligned}$$

Taking a union bound over $\mathcal{N}_\eta^2$ and choosing

$$r=C\log\big(\mathcal{N}(\eta,\mathcal{NN}_{\underline{t}},\|\cdot\|_\infty)\vee n\big)$$

shows that the same display holds simultaneously for all $u,v\in\mathcal{N}_\eta$ with probability at least $1-n^{-A_1}$, after increasing C.

Step 4: transfer from the cover to the clipped class. If u is an η-net approximation of s, then the pointwise bound in Theorem I.11, together with the clipped envelope, yields

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D}|\ell_{[\underline{t},\overline{t}]}(s,\boldsymbol{x})-\ell_{[\underline{t},\overline{t}]}(u,\boldsymbol{x})|\le C\sqrt{D(\overline{t}-\underline{t})(D\vee\log n)^2\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t}\,\eta.$$

Thus

$$\mathbb{E}_{\boldsymbol{X}_0}|\ell_{[\underline{t},\overline{t}]}(s)-\ell_{[\underline{t},\overline{t}]}(u)|+\mathbb{P}_n|\ell_{[\underline{t},\overline{t}]}(s)-\ell_{[\underline{t},\overline{t}]}(u)|\le n^{-A_2}$$

deterministically after taking A_0 large enough. This uses $D\le n^{\mathcal{O}(1)}$ and $n^{-\mathcal{O}(1)}\le\underline{t}<\overline{t}\le n^{\mathcal{O}(1)}$, so that $(D\vee\log n)^2\int_{\underline{t}}^{\overline{t}}\sigma_t^{-2}\mathrm{d}t\le n^{\mathcal{O}(1)}$. Since

$$\mathcal{L}_{[\underline{t},\overline{t}]}(s)-\mathcal{L}_{[\underline{t},\overline{t}]}(u)=\mathbb{E}_{\boldsymbol{X}_0}\{\ell_{[\underline{t},\overline{t}]}(s,\boldsymbol{X}_0)-\ell_{[\underline{t},\overline{t}]}(u,\boldsymbol{X}_0)\},$$

the same estimate also gives $|\mathcal{L}_{[\underline{t},\overline{t}]}(s)-\mathcal{L}_{[\underline{t},\overline{t}]}(u)|\le n^{-A_2}$. Hence the finite-cover Bernstein display applies to arbitrary $s,s'\in\mathcal{NN}_{\underline{t}}$, with $\mathcal{L}_{[\underline{t},\overline{t}]}(u)+\mathcal{L}_{[\underline{t},\overline{t}]}(v)$ replaced by $\mathcal{L}_{[\underline{t},\overline{t}]}(s)+\mathcal{L}_{[\underline{t},\overline{t}]}(s')+n^{-A_2}$, up to another additive n^{-A_2}. Enlarging A_2, this is the same display with the covering number evaluated at $n^{-\mathcal{O}(1)}$.

Step 5: use the ERM basic inequality. Fix any $s_0\in\mathcal{NN}_{\underline{t}}$. On the high-probability event from Steps 3–4,

$$\mathbb{E}_{\boldsymbol{X}_0}\ell_{[\underline{t},\overline{t}]}(\widehat{s},\boldsymbol{X}_0)-\mathbb{E}_{\boldsymbol{X}_0}\ell_{[\underline{t},\overline{t}]}(s_0,\boldsymbol{X}_0)\le(\mathbb{E}_{\boldsymbol{X}_0}-\mathbb{P}_n)\{\ell_{[\underline{t},\overline{t}]}(\widehat{s})-\ell_{[\underline{t},\overline{t}]}(s_0)\}+n^{-A_2},$$

because $\mathbb{P}_n \ell_{[\underline{t},\bar{t}]}(\widehat{s}) \le \mathbb{P}_n \ell_{[\underline{t},\bar{t}]}(s_0)$. Therefore,

$$\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le \mathcal{L}_{[\underline{t},\bar{t}]}(s_0) + C\left[\sqrt{\frac{(D\vee\log n)^2\left(\int_{\underline{t}}^{\bar{t}}\sigma_t^{-2}\mathrm{d}t\right)\{\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s})+\mathcal{L}_{[\underline{t},\bar{t}]}(s_0)\}r}{n}} + \frac{(D\vee\log n)^2\left(\int_{\underline{t}}^{\bar{t}}\sigma_t^{-2}\mathrm{d}t\right)r}{n}\right] + n^{-A_2}.$$

Using $\sqrt{ab} \le a/(4C) + Cb$ first with $a = \mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s})$ and then with $a = \mathcal{L}_{[\underline{t},\bar{t}]}(s_0)$, we absorb the square-root term and obtain

$$\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le C\mathcal{L}_{[\underline{t},\bar{t}]}(s_0) + C\frac{(D\vee\log n)^2\left(\int_{\underline{t}}^{\bar{t}}\sigma_t^{-2}\mathrm{d}t\right)r}{n} + n^{-A_2}.$$

Taking the infimum over $s_0 \in \mathcal{NN}_{\underline{t}}$ gives the high-probability oracle bound.

Step 6: pass to expectation. On the complement of the high-probability event, the clipped envelope and Theorem I.10 give

$$\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le 2\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t + 2\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t}\|s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \le C(D\vee\log n)^2\int_{\underline{t}}^{\bar{t}}\sigma_t^{-2}\mathrm{d}t.$$

Choosing A_1 large makes this exceptional contribution $\mathcal{O}(n^{-A})$. We obtain

$$\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n}\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le C\inf_{s\in\mathcal{NN}_{\underline{t}}}\mathcal{L}_{[\underline{t},\bar{t}]}(s) + C\frac{(D\vee\log n)^2\int_{\underline{t}}^{\bar{t}}\sigma_t^{-2}\mathrm{d}t\log\big(\mathcal{N}(n^{-\mathcal{O}(1)},\mathcal{NN}_{\underline{t}},\|\cdot\|_\infty)\vee n\big)}{n} + n^{-A},$$

which is the claimed oracle inequality. □

G.2. Time-Grid and Integral Conventions

For the VP/OU schedules used in the rate theorem, $\sigma_t^2 \asymp t$ for $0 < t \le 1$ and $\sigma_t \asymp 1$ for $t \ge 1$ on the truncated interval. The large-noise grid is chosen so that

$$[t_1, T] = \bigcup_{k=2}^{K} I_k, \qquad I_k = [t_{k-1}, t_k], \qquad t_k = 2t_{k-1} \text{ while } t_k \le 1, \qquad K \le C\log n, \tag{247}$$

where the dyadic portion is followed by $\mathcal{O}(\log n)$ order-one slabs of uniformly bounded length from time 1 to $T \asymp \log n$. On every large-noise slab,

$$c\,\sigma_{t_{k-1}} \le \sigma_t \le C\,\sigma_{t_{k-1}}, \qquad \int_{I_k}\sigma_t^{-2}\mathrm{d}t \le C.$$

For the small-noise interval and $t_0 \asymp n^{-2(\beta+1)/(d+2\beta)}$, $t_1 \asymp n^{-2/(d+2\beta)}$,

$$\int_{t_0}^{t_1}\sigma_t^{-2}\mathrm{d}t \le C\log(t_1/t_0) \le C\log n. \tag{248}$$

G.3. Score Estimation on Large-Noise Regime

We construct the deterministic comparison network inside the oracle proof below, so the approximation loss, slab-local entropy scale, and stochastic term are checked for the same large-noise class.

Proposition G.2 (Score estimation on large-noise regime)**.** *Let $I_k = [t_{k-1}, t_k] \subset [t_1, T]$ be a large-noise slab satisfying (247). Assume Theorems 2.2, 2.4 and 2.5 and $D \vee p_{\min}^{-1} \le n^{a_0}$. On the global grid used in the main theorem, $\sigma_{t_{k-1}}^{-1} \le C\sigma_{t_1}^{-1} \lesssim n^{1/(d+2\beta)}$, so every logarithmic range depending on $\sigma_{t_{k-1}}^{-1}$ contributes only polylogarithmic factors in n, which are hidden in $\tilde{\mathcal{O}}$ after enlarging constants depending only on (β, d, a_0). Let $\mathcal{F}_k^{\lg}$ be the clipped large-noise network class with the parameter bounds displayed in (250) below and coordinatewise clipping at $C\sigma_t^{-1}(D\vee\log n)^{1/2}$. Recall $\mathfrak{r}_n$ and $q_\star$ from (3), and define the accuracy-limited tangent-cell radius*

$$r_{\star,n} := c_\star\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C_\star}B_0^{-C_\star}p_{\min}^{C_\star}(D\vee\log n)^{-C_\star}\mathfrak{r}_n^{1/q_\star},$$

where $C_\star$ is sufficiently large and $c_\star$ sufficiently small, depending only on (β, d) and the fixed numerical constants in Theorem E.5. In particular,

$$r_{\star,n}^{-d} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} (D \vee \log n)^C \mathfrak{r}_n^{-d/q_\star}.$$

Let $\widehat{s}_k^{\mathrm{lg}}$ be an empirical risk minimizer over this class on I_k. Then

$$\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n} \int_{I_k} \mathbb{E}_{\boldsymbol{X}_t} \|\widehat{s}_k^{\mathrm{lg}}(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t \\
&\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \left[n^{-2(\beta+1)/(d+2\beta)} + \frac{\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}}{n} \right] (D \vee \log n)^{\mathfrak{p}_{\beta,d}} + n^{-A}.
\end{aligned} \tag{249}$$

Writing $L_k, \boldsymbol{W}_k, S_k, B_k$ for the depth, width vector, sparsity, and weight bound of the learned large-noise class, its parameters satisfy

$$\begin{aligned}
L_k &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C (D \vee \log n)^{\mathfrak{p}_{\beta,d}}, \\
\log B_k &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C (D \vee \log n)^{\mathfrak{p}_{\beta,d}}, \\
\|\boldsymbol{W}_k\|_\infty &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \left(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\right) (D \vee \log n)^{\mathfrak{p}_{\beta,d}}, \\
S_k &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \left(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\right) (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.
\end{aligned} \tag{250}$$

Proof. Step 1: construct a comparison network in the slab-local class. Apply Theorem E.5 on I_k, with the theorem's local time-window notation set to $\underline{t} = t_{k-1}$ and $\bar{t} = t_k$. Set

$$r_k := c_0 \sigma_{t_{k-1}} \wedge r_{\star,n},$$

where $c_0 > 0$ is a sufficiently small numerical constant. Use theorem parameters $\varepsilon_{\mathrm{cell}} = r_k^\beta$ and $\varepsilon_{\mathrm{sc}} = n^{-(\beta+1)/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}/2}/\sqrt{D}$. The constant c_0 is chosen so that $\sigma_t \ge c_{\mathrm{lg}} r_k$ throughout I_k, using the slab comparability in (247). The polynomial bounds on D, $p_{\min}^{-1}$, and the automatic grid bound on $\sigma_{t_{k-1}}^{-1}$ imply

$$1 + R_{\mathrm{lb}} \le (D \vee \log n)^{2(q_{\mathrm{geom}}+2)+C}, \qquad H_{\mathrm{lg}}^{\mathrm{lb}} \le \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} (D \vee \log n)^{2(q_{\mathrm{geom}}+2)+C}, \qquad \sqrt{D} \le D \vee \log n.$$

By the displayed bound on $H_{\mathrm{lg}}^{\mathrm{lb}}$ and the definition of $r_{\star,n}$, the phase-smallness part of (65) holds:

$$\Gamma_{\mathcal{M},q_{\mathrm{geom}}} r_k (H_{\mathrm{lg}}^{\mathrm{lb}})^{1/2} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} r_{\star,n} (D \vee \log n)^C \le c_{\mathrm{ph,cell}}.$$

The same choice of $C_\star$ and $c_\star$ gives the high-order geometry remainder bound

$$\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} r_k^{q_{\mathrm{geom}}-2} (H_{\mathrm{lg}}^{\mathrm{lb}})^C \le c\, \frac{n^{-(\beta+1)/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}/2}}{\sqrt{D}(1 + R_{\mathrm{lb}})}.$$

Here we used the polynomial bound on R_{lb} and the fact that $r_k \le r_{\star,n}$; this is the step that prevents the order-one-noise slabs from violating the fixed-order Taylor remainder condition. Bounded n are absorbed by increasing constants. Thus the geometry condition in (65) is satisfied, and (66) gives a network $s_{k,0}^{\mathrm{lg}}$ such that, on the tube A_t defined in (7),

$$\|s_{k,0}^{\mathrm{lg}}(\boldsymbol{x}, t) - s^*(\boldsymbol{x}, t)\|_2^2 \le C\sigma_t^{-2} n^{-2(\beta+1)/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

After adding coordinatewise clipping at $C\sigma_t^{-1}(D \vee \log n)^{1/2}$, and enlarging constants if needed, we obtain $s_k^{\mathrm{lg}} \in \mathcal{F}_k^{\mathrm{lg}}$ with the same tube error and with

$$\sup_{\boldsymbol{x} \in \mathbb{R}^D} \|s_k^{\mathrm{lg}}(\boldsymbol{x}, t)\|_\infty \le C\sigma_t^{-1} (D \vee \log n)^{1/2}.$$

Indeed, on A_t, (64) and the preceding polynomial assumptions on $D, p_{\min}^{-1}$, and $\sigma_{t_{k-1}}^{-1}$ imply

$$\sigma_t \|s^*(\boldsymbol{x}, t)\|_\infty \le \sigma_t \|s^*(\boldsymbol{x}, t)\|_2 \le C(D \vee \log n)^{1/2}.$$

Coordinatewise clipping at this radius is 1-Lipschitz and fixes the target value coordinatewise, so it cannot increase the tube approximation error. Its size is bounded by (67). The identity

$$(r_k^{\beta})^{-d/\beta} = r_k^{-d} \le C\left(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\right),$$

and the absorption of $H_{\mathrm{lg}}^{\mathrm{lb}}$ and R_{lb} into $(D \vee \log n)^{\mathfrak{p}_{\beta,d}}$ shows that the resulting class may be chosen with the parameter bounds displayed in (250).

Step 2: integrate the comparison loss. On A_t^c, apply Theorem C.3 with $\ell_{\mathrm{off}} = D \vee \log n$; the clipping bound for s_k^{lg} is then within the required coordinatewise envelope. Thus, for any prescribed fixed $A' > 0$, after increasing C_ϱ,

$$\int_{A_t^c} \|s_k^{\mathrm{lg}}(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_2^2 p_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \le C\sigma_t^{-2}(D \vee \log n)^{\mathfrak{p}_{\beta,d}} n^{-A'}.$$

Taking A' large enough and using the polynomial lower bound on σ_t, the off-tube contribution is absorbed into the tube rate. Therefore

$$\int_{I_k} \mathbb{E}_{\boldsymbol{X}_t}\|s_k^{\mathrm{lg}}(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2(\beta+1)/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

where we used $\int_{I_k} \sigma_t^{-2}\mathrm{d}t \le C_{\mathrm{grid}}$. Hence

$$\inf_{s \in \mathcal{F}_k^{\mathrm{lg}}} \int_{I_k} \mathbb{E}_{\boldsymbol{X}_t}\|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2(\beta+1)/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

Step 3: evaluate the slab-local entropy. Apply the clipped score-estimation oracle Theorem G.1 on the slab I_k. By (250), the large-noise class satisfies the entropy evaluation (8) with

$$M_k := \sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}.$$

Indeed, $L_k + \log B_k$ is polylogarithmic up to the displayed polynomial prefactors, while $S_k + \|\boldsymbol{W}_k\|_\infty$ is bounded by $M_k(D \vee \log n)^{\mathfrak{p}_{\beta,d}}$ times those same fixed polynomial prefactors. Therefore

$$\log\big(\mathcal{N}(n^{-A}, \mathcal{F}_k^{\mathrm{lg}}, \|\cdot\|_\infty) \vee n\big) \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C}\left(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\right)(D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

Substituting this covering bound into the oracle inequality, using $\int_{I_k} \sigma_t^{-2}\mathrm{d}t \le C_{\mathrm{grid}}$, and absorbing the oracle envelope factor $(D \vee \log n)^2$ into $\mathfrak{p}_{\beta,d}$, gives the stochastic contribution

$$C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C}\frac{\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}}{n}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

with all manifold and smoothness prefactors displayed. Combining this with the comparison loss from Step 2 proves (249). □

G.4. Score Estimation on Small-Noise Regime

We also fold the deterministic small-noise comparison loss into the oracle proof, so the comparison function and the learned class are checked in the same place.

Theorem G.3 (Score estimation on small-noise regime)**.** *Assume Theorems 2.2, 2.4 and 2.5 and $D \vee p_{\min}^{-1} \le n^{a_0}$. Assume also the small-noise reach and denominator-resolution conditions (232) and (233). Let $\mathcal{F}_{\mathrm{sm}}^{\mathrm{lb}}$ be the clipped small-noise network class with clipping radius $C\sigma_t^{-1}(D \vee \log n)^{1/2}$ and with the size bounds in (237) and (238), and let $\widehat{s}_{\mathrm{sm}}^{\mathrm{lb}}$ be an empirical risk minimizer over $\mathcal{F}_{\mathrm{sm}}^{\mathrm{lb}}$ on $[t_0, t_1]$. Then*

$$\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n}\int_{t_0}^{t_1} \mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}_{\mathrm{sm}}^{\mathrm{lb}}(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \\
&\quad \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}} \\
&\qquad + C\frac{(D \vee \log n)^C \int_{t_0}^{t_1} \sigma_t^{-2}\mathrm{d}t \log\big(\mathcal{N}(n^{-A}, \mathcal{F}_{\mathrm{sm}}^{\mathrm{lb}}, \|\cdot\|_\infty) \vee n\big)}{n} + n^{-A}.
\end{aligned} \tag{251}$$

In particular, using (8) *with* $M_n = n^{d/(d+2\beta)}$,

$$\begin{aligned}&\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n} \int_{t_0}^{t_1} \mathbb{E}_{\boldsymbol{X}_t} \|\widehat{s}_{\mathrm{sm}}^{\mathrm{lb}}(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t \\ &\le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} + n^{-A}.\end{aligned} \tag{252}$$

Proof. The comparison network $s_{\mathrm{sm}}^{\mathrm{lb}}$ belongs to $\mathcal{F}_{\mathrm{sm}}^{\mathrm{lb}}$, after the same coordinatewise clipping at radius $C\sigma_t^{-1}(D \vee \log n)^{1/2}$. Let A_t be the tube defined in (7). On A_t, (235) gives the following bound after inserting the harmless factors $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C \ge 1$:

$$\|s_{\mathrm{sm}}^{\mathrm{lb}}(\boldsymbol{x}, t) - s^*(\boldsymbol{x}, t)\|_2^2 \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} \sigma_t^{-2}.$$

The single-slab integral bound (248) gives $\int_{t_0}^{t_1} \sigma_t^{-2} \mathrm{d}t \le C \log n \le C(D \vee \log n)^C$, so the tube contribution is bounded by

$$C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

On A_t^c, apply Theorem C.3 with $\ell_{\mathrm{off}} = D \vee \log n$; the clipping bound (236) satisfies the required envelope. After increasing C_ϱ, this gives n^{-A}. Thus

$$\inf_{s \in \mathcal{F}_{\mathrm{sm}}^{\mathrm{lb}}} \int_{t_0}^{t_1} \mathbb{E}_{\boldsymbol{X}_t} \|s(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} + n^{-A}.$$

Applying the clipped oracle inequality Theorem G.1 to the class $\mathcal{F}_{\mathrm{sm}}^{\mathrm{lb}}$ gives (251). The entropy bound follows from (8), (237) and (238); it is bounded by

$$C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C p_{\min}^{-C} n^{d/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

with all manifold and smoothness prefactors displayed. Since

$$\frac{n^{d/(d+2\beta)}}{n} = n^{-1+d/(d+2\beta)} = n^{-2\beta/(d+2\beta)}$$

and $\int_{t_0}^{t_1} \sigma_t^{-2} \mathrm{d}t \le C(D \vee \log n)^C$, the stochastic term is absorbed into the right side of (252). □

H. W_1 Distribution Estimation

This appendix section contains the formal W_1 distribution-estimation results. The main-paper summaries are stated in Sec. 5; the generic score-estimation tools are stated in Sec. 4, and the formal learned-score inputs are proved in Sec. G. All displayed W_1 bounds below use the branchwise density-lower-bound learned-score inputs and the off-tube localization bounds.

Whenever an appendix estimate invokes the regular VP schedule, it refers to the standard VP/OU schedule facts that, for universal schedule constants independent of (n, D),

$$c_\sigma t \le \sigma_t^2 \le C_\sigma t, \quad 1 - m_t \le C_m t \qquad (0 \le t \le 1), \qquad m_t \le C_{\mathrm{mix}} e^{-c_{\mathrm{mix}} t} \quad (t \ge 0). \tag{253}$$

In particular, on every early interval $[0, t_1]$ used below, m_t is bounded below by a positive schedule constant $\underline{m}$ once $t_1 \le 1$.

The main statement is Theorem H.1. Its proof is a bookkeeping argument with four terms: early stopping, terminal Gaussian initialization, the small-noise score error, and the dyadic sum of large-noise score errors. The only distribution-level input is the slabwise perturbation lemma Theorem I.17; all score-estimation inputs come from Theorems G.2 and G.3.

H.1. Controlled-regime checklist

This subsection expands the non-asymptotic regime conditions invoked in Theorem 5.1. Let $H_n := D \vee \log n$ and let q_{geom} be as in (1). The conditions split into the following visually distinct groups.

Condition group	Meaning
Schedule	The regular VP schedule convention equation 253, so that $\sigma_t^2 \asymp t$ on $[0,1]$, $m_t \asymp 1$ on the early-stopping regime, and m_T decays exponentially on the terminal regime
Geometry	P_0 is supported on a compact smooth d-manifold $\mathcal{M} \subset [0,1]^D$ with reach $\kappa > 0$, volume $S_{\mathcal{M}}$, and finite envelope $\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}$
Density	$p_0 \in \mathcal{H}^\beta(\mathcal{M}, B_0)$ and $p_0 \ge p_{\min} > 0$
Polynomial control	$D \vee p_{\min}^{-1} \le n^{a_0}$ and the displayed geometry/density factors are bounded uniformly in n, or by fixed powers of D in growing-ambient families
Time grid	$t_0 \asymp n^{-2(\beta+1)/(d+2\beta)}$, $t_1 \asymp n^{-2/(d+2\beta)}$, and $T \asymp \log n$
Small-noise reach	$3c_\star \underline{m}^{-1} \sigma_{t_1} \sqrt{H_n} \le \eta_\Pi/2$, so the forward tube is inside the region where nearest projection is unique
Denominator resolution	$C_{\mathrm{den,sm}} \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-1} n^{-\beta/(d+2\beta)} H_n^{\mathfrak{p}_{\beta,d}/2-1/2} \le 1$, so the de-Gaussianized denominator can be approximated before taking a reciprocal
Estimator	Clipped ERM over the slab-dependent ReLU classes from Sec. 4 and G

The small-noise reach and denominator-resolution conditions are sample-size conditions, not additional smoothness assumptions. For fixed D, fixed positive reach, fixed $p_{\min}$, and fixed finite-order geometry, both hold for all sufficiently large n. For growing D or deteriorating geometry, they state explicitly how much ambient growth, small reach, or small density lower bound the proof can tolerate.

H.2. W_1 distribution estimation

We use the same slabwise partition throughout this subsection: after choosing t_0, t_1, T, write

$$[t_0, T] = [t_0, t_1] \cup \bigcup_{k=2}^{K} I_k, \qquad I_k = [t_{k-1}, t_k],$$

where the large-noise slabs satisfy (247); in particular, this is the dyadic grid $t_k = 2t_{k-1}$ while $t_k \le 1$, followed by order-one slabs up to T.

Audit checkpoint: rate assembly. Theorem H.1 is where the two score-estimation regimes become one distribution rate. The small-noise score risk is larger, but the interval length contributes the factor $\sqrt{t_1} \asymp n^{-1/(d+2\beta)}$. The large-noise stochastic terms are summed over the dyadic grid, where $\sum_k \sigma_{t_{k-1}}^{-d}$ is dominated by the first large-noise slab, while the

accuracy-cap contribution $\mathfrak{r}_n^{-d/q_\star}$, with $\mathfrak{r}_n$ and $q_\star$ from (3), is carried explicitly through the W_1 perturbation sum. These checks are the only rate-assembly points after the branchwise learned-score bounds are available.

Rate Decomposition. Table 3 summarizes the early stopping, score-estimation, and terminal terms. Throughout this section we use the global notation $\mathfrak{r}_n$ and $q_\star$ from (3).

Table 3. Rate decomposition for the W_1 generative bound.

Temporal Regime	**Deterministic Error**	**Intrinsic Capacity (M)**	**Stochastic Penalty**	**Final W_1 Contribution**
Early Stopping $[0, t_0]$	—	—	—	$\tilde{\mathcal{O}}(n^{-(\beta+1)/(d+2\beta)})$
Small Noise $[t_0, t_1]$	$\tilde{\mathcal{O}}(n^{-2\beta/(d+2\beta)})$	$n^{d/(d+2\beta)}$	$n^{-2\beta/(d+2\beta)}$	$\sqrt{t_1}\, n^{-\beta/(d+2\beta)} \asymp n^{-(\beta+1)/(d+2\beta)}$
Large Noise $[t_1, T]$ (evaluated per slab I_k)	$\tilde{\mathcal{O}}(\mathfrak{r}_n^2)$	$\sigma_{t_{k-1}}^{-d} + \mathfrak{r}_n^{-d/q_\star}$	$(\sigma_{t_{k-1}}^{-d} + \mathfrak{r}_n^{-d/q_\star})/n$	$\mathfrak{r}_n + n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}$ (following dyadic summation)
Terminal Init $[T, \infty]$	—	—	—	Negligible for $T \asymp \log n$

Corollary H.1 (W_1 distribution estimation)**.** *Assume Theorems 2.2, 2.4 and 2.5, $D \vee p_{\min}^{-1} \le n^{a_0}$, and the two small-noise conditions (232) and (233). Let $t_0 \asymp n^{-2(\beta+1)/(d+2\beta)}$, $t_1 \asymp n^{-2/(d+2\beta)}$, $T \asymp \log n$, and use $\mathfrak{r}_n, q_\star$ from (3). Let $\widehat{s}^{\mathrm{lb}}$ be the switched clipped ERM obtained by using Theorem G.3 on $[t_0, t_1]$ and Theorem G.2 on the large-noise slabs I_k, $k = 2, \dots, K$. Let $\widehat{P}_{t_0}^{\mathrm{lb}}$ be the generated law driven by $\widehat{s}^{\mathrm{lb}}$. Then there is a constant $C < \infty$, independent of (n, D), such that*

$$\begin{aligned}\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}_{t_0}^{\mathrm{lb}}) &\le C\sqrt{Dt_0} + C\sqrt{D}\,\exp(-c_{\mathrm{mix}}T) \\ &\quad + C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C}\Big(\mathfrak{r}_n + n^{-1/2} + n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}\Big)(D \vee \log n)^{\mathfrak{p}_{\beta,d}}.\end{aligned} \tag{254}$$

If $T = c_T \log n$ with c_T large enough, depending only on the fixed constants of the setup, the terminal term is absorbed by the third term in equation 254. Since $t_0 \asymp \mathfrak{r}_n^2$ and $\mathfrak{p}_{\beta,d} \ge 1/2$, the early-stopping term is also absorbed by that term. If $d > 2$ and $q_\star \ge d(\beta+1)/(d-2)$, then $n^{-1/2} + n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)} \le C\mathfrak{r}_n$, and hence

$$\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}_{t_0}^{\mathrm{lb}}) \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C}\mathfrak{r}_n (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

Writing $L, \boldsymbol{W}, S, B$ for the depth, width vector, sparsity, and weight bound of the switched learned score, the assembled class can be chosen so that

$$L \vee \log B \le C\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$
$$\|\boldsymbol{W}\|_\infty \vee S \le C\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

where

$$\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} := (1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}^{\mathfrak{a}_{\mathrm{geom}}}.$$

One may take

$$\mathfrak{p}_{\beta,d} = C + \max\left\{8(q_{\mathrm{geom}}+2), \left\lceil \lfloor\beta\rfloor + 1 + \frac{\lfloor\beta\rfloor d}{\beta - \lfloor\beta\rfloor} + \frac{d}{\beta \vee 1} + \frac{d}{2} + 6\right\rceil\right\}.$$

The universal C in this exponent is independent of $\beta, d, D, n, p_{\min}$, and the numerical geometry constants; under the above $d > 2$ finite-order choice, $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$.

Proof. Step 1: decompose the distribution error. Use the interpolation laws from Theorem I.17 on the partition $[t_0, T] = [t_0, t_1] \cup \bigcup_{k=2}^K I_k$. Define $\mathsf{W}_1^{\mathrm{sm,lb}} := \mathsf{W}_1(\overline{P}_{t_0}^{(0)}, \overline{P}_{t_0}^{(1)})$ for the single small-noise replacement and $\mathsf{W}_1^{(k),\mathrm{lg,lb}} := \mathsf{W}_1(\overline{P}_{t_0}^{(k-1)}, \overline{P}_{t_0}^{(k)})$ for the large-noise slab I_k. With this notation, $\overline{P}_{t_0}^{(K)}$ is the exact-terminal reverse law driven by the switched learned score $\widehat{s}^{\mathrm{lb}}$. By the triangle inequality along the reverse-time telescoping construction,

$$\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}_{t_0}^{\mathrm{lb}}) \le \mathsf{W}_1(P_0, P_{t_0}) + \mathbb{E}\mathsf{W}_1(\overline{P}_{t_0}, \widehat{P}_{t_0}) + \mathbb{E}\,\mathsf{W}_1^{\mathrm{sm,lb}} + \sum_{k:\, I_k \subset [t_1, T]} \mathbb{E}\,\mathsf{W}_1^{(k),\mathrm{lg,lb}}.$$

The early-stopping and initialization bounds follow from Theorems I.15 and I.16:

$$\mathsf{W}_1(P_0, P_{t_0}) \le C\sqrt{Dt_0}, \qquad \mathbb{E}\mathsf{W}_1(\overline{P}_{t_0}, \widehat{P}_{t_0}) \le C\sqrt{D}\,\exp(-c_{\mathrm{mix}}T).$$

Step 2: bound the small-noise contribution. By the clipping radii in Theorems G.2 and G.3, and by the bounded ReLU time switches used to assemble the branch estimators,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|\widehat{s}^{\mathrm{lb}}(\boldsymbol{x},t)\|_\infty \le C\sigma_t^{-1}(D\vee\log n)^{1/2}, \qquad t\in[t_0,T].$$

This verifies the clipped-score hypothesis of Theorem I.17 with $a_{\mathrm{clip}}=1/2$. For the small-noise slab, apply Theorem I.17. Together with (252), $t_1 \asymp n^{-2/(d+2\beta)}$, and choosing the negligible exponent A large, this gives

$$\begin{aligned}\mathbb{E}\,\mathsf{W}_1^{\mathrm{sm,lb}} &\le C(D\vee\log n)^{a_{\mathrm{W}}} n^{-1/(d+2\beta)}\sqrt{\log n}\Big(\mathbb{E}\int_{t_0}^{t_1}\mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}^{\mathrm{lb}}(\boldsymbol{X}_t,t)-s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t\Big)^{1/2}\\ &\quad + C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C}(D\vee\log n)^{a_{\mathrm{W}}} n^{-(\beta+1)/(d+2\beta)}\\ &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C} n^{-(\beta+1)/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.\end{aligned}$$

Here we used

$$n^{-1/(d+2\beta)}\,n^{-\beta/(d+2\beta)} = n^{-(\beta+1)/(d+2\beta)}.$$

Step 3: bound the large-noise contributions. On the large-noise slabs, Theorem I.17 and (249), again with the negligible exponent chosen large, give

$$\begin{aligned}\mathbb{E}\,\mathsf{W}_1^{(k),\mathrm{lg,lb}} &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C}(D\vee\log n)^{a_{\mathrm{W}}}\sqrt{t_k\log n}\,\Big[n^{-(\beta+1)/(d+2\beta)}\\ &\qquad + n^{-1/2}\sigma_{t_{k-1}}^{-d/2} + n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}\Big](D\vee\log n)^{\mathfrak{p}_{\beta,d}}\\ &\quad + C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C}(D\vee\log n)^{a_{\mathrm{W}}} n^{-(\beta+1)/(d+2\beta)}\end{aligned}$$

for each large-noise slab I_k. The first term sums to $n^{-(\beta+1)/(d+2\beta)}$ up to the displayed polynomial ambient factor because $t_k \le T \lesssim \log n$ and $K \lesssim \log n$. For the stochastic slab-size term, the dyadic grid $t_k = 2t_{k-1}$ while $t_k \le 1$, the relation $\sigma_t^2 \asymp t$ near zero, and $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$ give

$$n^{-1/2}\sum_{k:\,t_k\le 1} t_k^{1/2}\sigma_{t_{k-1}}^{-d/2} \le Cn^{-1/2}\big(1+\sigma_{t_1}^{1-d/2}\big)(D\vee\log n)^{\mathfrak{p}_{\beta,d}} \le C\left(n^{-(\beta+1)/(d+2\beta)}+n^{-1/2}\right)(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.$$

The remaining order-one-time slabs have $\sigma_{t_{k-1}} \asymp 1$, and contribute only $n^{-1/2}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$. The accuracy-cap term is independent of the slab index except for logarithmic factors; since $K \lesssim \log n$ and $t_k \le T \lesssim \log n$, its large-noise contribution is bounded by

$$Cn^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.$$

Hence

$$\sum_{k:\,I_k\subset[t_1,T]}\mathbb{E}\,\mathsf{W}_1^{(k),\mathrm{lg,lb}} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C}\left(n^{-(\beta+1)/(d+2\beta)}+n^{-1/2}+n^{-1/2}\mathfrak{r}_n^{-d/(2q_\star)}\right)(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.$$

Combining these terms proves (254).

Step 4: record the network sizes. The small-noise learned class from Theorem G.3 has separate bounds

$$L_{\mathrm{sm}}\vee\log B_{\mathrm{sm}} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C}(D\vee\log n)^{\mathfrak{p}_{\beta,d}},$$

and

$$\|\boldsymbol{W}_{\mathrm{sm}}\|_\infty\vee S_{\mathrm{sm}} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C} n^{d/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.$$

For each large-noise slab, (250) gives

$$\begin{aligned}L_k\vee\log B_k &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C(D\vee\log n)^{\mathfrak{p}_{\beta,d}},\\ \|\boldsymbol{W}_k\|_\infty\vee S_k &\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C p_{\min}^{-C}\left(\sigma_{t_{k-1}}^{-d}+\mathfrak{r}_n^{-d/q_\star}\right)(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.\end{aligned}$$

There are $K \lesssim \log n$ large-noise slabs, and the ReLU time switches add only $\mathcal{O}(K)$ depth/sparsity and a logarithmic number of weights. The dyadic grid gives

$$\sum_{k=2}^{K} \sigma_{t_{k-1}}^{-d} \le C\sigma_{t_1}^{-d} \asymp n^{d/(d+2\beta)}.$$

The additional accuracy-cap contribution satisfies

$$K\mathfrak{r}_n^{-d/q_\star} \le Cn^{d/(d+2\beta)}(D \vee \log n)^C,$$

because $q_\star > \beta + 1$ by (1). The subnetworks are assembled in parallel behind those switches. Thus the depth satisfies

$$L \le \max\{L_{\mathrm{sm}}, L_2, \dots, L_K\} + CK \le C(1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}^{\mathfrak{a}_{\mathrm{geom}}} B_0^C p_{\min}^{-C}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

which is the stated depth bound. The width satisfies

$$\|\boldsymbol{W}\|_\infty \le \|\boldsymbol{W}_{\mathrm{sm}}\|_\infty + \sum_{k=2}^{K} \|\boldsymbol{W}_k\|_\infty + CK \le C(1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}^{\mathfrak{a}_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

after absorbing $K \lesssim \log n$; this proves the stated width bound. The sparsity satisfies

$$S \le S_{\mathrm{sm}} + \sum_{k=2}^{K} S_k + CK \le C(1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}^{\mathfrak{a}_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

which is the stated sparsity bound. Finally,

$$\begin{aligned}\log B &\le \max\{\log B_{\mathrm{sm}}, \log B_2, \dots, \log B_K\} + C\log K \\ &\le C(1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}^{\mathfrak{a}_{\mathrm{geom}}} B_0^C p_{\min}^{-C}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},\end{aligned}$$

which is the stated log-weight bound. □

Corollary H.2 (Fixed-geometry sufficient ambient-growth regime)**.** *Fix $\beta > 0$ and $d \ge 1$. Work under the regular VP schedule convention equation 253. Suppose Theorems 2.2, 2.4 and 2.5 holds and that, for the fixed order q_{geom}, the quantities $\underline{m}$, η_Π, $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$, B_0, and $p_{\min}$ are bounded above or below by positive constants independent of (D, n). Let*

$$a_{\beta,d} := \min\left\{\frac{2}{d+2\beta}, \frac{2\beta}{(d+2\beta)(\mathfrak{p}_{\beta,d}-1)}\right\}.$$

Then there exists a constant $c > 0$, depending only on the preceding fixed quantities, such that

$$D \vee \log n \le c\, n^{a_{\beta,d}}$$

implies both (232) and (233). Consequently, under this fixed-geometry growth regime, the bound in Theorem H.1 applies.

Proof. Write $H_n = D \vee \log n$. By equation 253 and $t_1 \asymp n^{-2/(d+2\beta)}$, we have $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$. Hence the small-noise reach condition (232) follows whenever

$$H_n \le c_1 n^{2/(d+2\beta)}$$

for a constant c_1 depending only on $\underline{m}$, η_Π, and the schedule constants. Likewise, after absorbing the fixed quantities $p_{\min}^{-1}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C B_0^C$ into a constant, the denominator condition (233) is implied by

$$H_n^{(\mathfrak{p}_{\beta,d}-1)/2} \le c_2 n^{\beta/(d+2\beta)},$$

equivalently,

$$H_n \le c_3 n^{2\beta/((d+2\beta)(\mathfrak{p}_{\beta,d}-1))}.$$

Taking $c := \min\{c_1, c_3\}$ proves the claim. □

I. Technical Toolkit

This appendix collects reusable tools that are not specific to the manifold geometry. They can be treated as black boxes when reading the approximation and estimation proofs. The most frequently used ingredients are:

Tool class	Representative labels
Basic ReLU approximation and arithmetic	Theorems I.1 to I.5
Schedule, reciprocal, power, and clipping networks	Theorems I.6 to I.9
Concentration, Bernstein, and covering arguments	Theorems I.11 to I.14
Distribution-estimation support	Theorems I.15 to I.17

I.1. Neural network approximation toolkit

I.1.1. Imported neural network approximation lemmas

Theorem I.1 (ReLU approximation of Hölder functions, after (Schmidt-Hieber, 2020, Theorem 5))**.** *For any function $f \in \mathcal{H}^\beta([0,1]^D, B)$ and any integers $m \geq 1$ and $N \geq (\beta+1)^D \vee (B+1)e^D$, there exists a network*

$$\widetilde{f} \in \mathcal{NN}(L, \boldsymbol{W}, S, 1)$$

with

$$\begin{aligned} L &= 8 + (m+5)(1 + \lceil \log_2(D \vee \beta) \rceil), \\ \|\boldsymbol{W}\|_\infty &\leq 6(D \vee \lceil \beta \rceil)N, \\ S &\leq 141(D+\beta+1)^{D+3} N(m+6) \end{aligned}$$

such that

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^D)} \leq (2B+1)(1+D^2+\beta^2)6^D N 2^{-m} + B 3^\beta N^{-\frac{\beta}{D}}.$$

Lemma I.2 (ReLU approximation of VP schedules, after Lemma B.1 of (Oko et al., 2023))**.** *Let m_t and σ_t be the schedules of the VP forward process on $[\underline{t}, \overline{t}]$. Let $0 < \varepsilon < \frac{1}{2}$. Then, there exists a neural network $\phi_m(t) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ that approximates m_t uniformly on $[\underline{t}, \overline{t}]$, within the additive error of ε, where $L = \mathcal{O}(\log^2 \varepsilon^{-1})$, $\|\boldsymbol{W}\|_\infty = \mathcal{O}(\log \varepsilon^{-1})$, $S = \mathcal{O}(\log^2 \varepsilon^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$.*

Also, there exists a neural network $\phi_\sigma(t) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ that approximates σ_t uniformly on $[\underline{t}, \overline{t}]$, within the additive error of ε, where $L \leq \mathcal{O}(\log^2 \varepsilon^{-1})$, $\|\boldsymbol{W}\|_\infty = \mathcal{O}(\log^3 \varepsilon^{-1})$, $S = \mathcal{O}(\log^4 \varepsilon^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \varepsilon^{-1}))$.

Lemma I.3 (ReLU multiplication network, after Lemma F.6 of (Oko et al., 2023))**.** *Let $d \geq 2, C \geq 1$, and $\varepsilon' \in (0,1]$. For any $\varepsilon > 0$, there exists a ReLU network $\phi_{\text{multi}}(x_1, \ldots, x_d) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$\begin{aligned} L &\lesssim \log d(\log \varepsilon^{-1} + d \log C), \\ \|\boldsymbol{W}\|_\infty &= 48d, \\ S &\lesssim d \log \varepsilon^{-1} + d \log C, \\ B &= C^d, \end{aligned}$$

such that

$$\left| \phi_{\text{multi}}(x_1', \ldots, x_d') - \prod_{k=1}^d x_k \right| \leq \varepsilon + dC^{d-1}\varepsilon', \quad \textit{for all } \boldsymbol{x} \in [-C, C]^d \textit{ and } \boldsymbol{x}' \in \mathbb{R}^d \textit{ with } \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \varepsilon'.$$

Moreover, $|\phi_{\text{multi}}(\boldsymbol{x})| \leq C^d$ for all $\boldsymbol{x} \in \mathbb{R}^d$, and $\phi_{\text{multi}}(x_1', \ldots, x_d') = 0$ if at least one of $x_k' = 0$. We note that some of x_i, x_j $(i \neq j)$ can be shared. For the product $\prod_{k=1}^I x_k^{\alpha_k}$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I) \in \mathbb{N}^I$ and $\sum_{k=1}^I \alpha_k = d$, there exists a neural network satisfying the same bounds as above, denoted by $\phi_{\text{multi}}(\boldsymbol{x}; \boldsymbol{\alpha})$.

Lemma I.4 (ReLU reciprocal network, after Lemma F.7 of (Oko et al., 2023))**.** *For any $\varepsilon \in (0,1)$, there exists $\phi_{\mathrm{rec}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$L \lesssim \log^2 \varepsilon^{-1}, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1}, \quad S \lesssim \log^4 \varepsilon^{-1}, \quad \log B \lesssim \log \varepsilon^{-1},$$

such that

$$\left|\phi_{\mathrm{rec}}(x') - \frac{1}{x}\right| \le \varepsilon + \frac{|x'-x|}{\varepsilon}, \quad \textit{for all } x \in [\varepsilon, \varepsilon^{-1}] \textit{ and } x' \in \mathbb{R}.$$

Lemma I.5 (Pointwise ReLU approximation of e^{-x}, after Lemma F.12 of (Oko et al., 2023))**.** *For any $\varepsilon > 0$, there exists ReLU network $\phi_{\exp} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ such that*

$$\left|e^{-x'} - \phi_{\exp}(x)\right| \le \varepsilon + |x - x'| \qquad \textit{for all } x, x' \ge 0$$

holds, where

$$L \lesssim \log^2 \varepsilon^{-1}, \quad \|\boldsymbol{W}\|_\infty \lesssim \log \varepsilon^{-1}, \quad S \lesssim \log^2 \varepsilon^{-1}, \quad \log B \lesssim \log^2 \varepsilon^{-1}.$$

Moreover, for all $x \ge \log(3/\varepsilon)$ it holds that $|\phi_{\exp}(x)| \le \varepsilon$.

I.1.2. Auxiliary schedule and arithmetic approximation lemmas

All schedule-dependent lemmas in this subsection are consequences of Theorem I.2 and the algebraic network lemmas.

Lemma I.6 (ReLU approximation of the inverse VP mean $1/m_t$)**.** *Let $\underline{m} := \inf_{t \in [\underline{t}, \bar{t}]} m_t > 0$. Fix $\varepsilon \in (0, 1/2]$, there exist ReLU networks $\widetilde{\phi}_{1/m} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$L,\ \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 \underline{m}^{-1}, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + \log^3 \underline{m}^{-1}, \quad S \lesssim \log^4 \varepsilon^{-1} + \log^4 \underline{m}^{-1}$$

such that

$$\|\widetilde{\phi}_{1/m}(t) - m_t^{-1}\|_{L^\infty([\underline{t}, \bar{t}])} \le \varepsilon.$$

Proof. Choose an auxiliary accuracy

$$\varepsilon_m := \frac{\underline{m}\varepsilon^2}{4}.$$

By Theorem I.2 for m_t, there exists a ReLU network $\widetilde{\phi}_m \in \mathcal{NN}(L_1, \boldsymbol{W}_1, S_1, B_1)$ with

$$L_1 = \mathcal{O}(\log^2 \varepsilon_m^{-1}), \quad \|\boldsymbol{W}_1\|_\infty = \mathcal{O}(\log \varepsilon_m^{-1}), \quad S_1 = \mathcal{O}(\log^2 \varepsilon_m^{-1}), \quad \log B_1 = \mathcal{O}(\log^2 \varepsilon_m^{-1})$$

such that

$$\|\widetilde{\phi}_m - m_t\|_{L^\infty([\underline{t}, \bar{t}])} \le \varepsilon_m.$$

Now apply the reciprocal lemma (Theorem I.4) with parameter

$$\varepsilon_{\mathrm{rec}} := \frac{\underline{m}\varepsilon}{2}.$$

Since $m_t \in [\underline{m}, 1] \subset [\varepsilon_{\mathrm{rec}}, \varepsilon_{\mathrm{rec}}^{-1}]$, the reciprocal network ϕ_{rec} satisfies

$$\left|\phi_{\mathrm{rec}}(\widetilde{\phi}_m(t)) - \frac{1}{m_t}\right| \le \varepsilon_{\mathrm{rec}} + \frac{|\widetilde{\phi}_m(t) - m_t|}{\varepsilon_{\mathrm{rec}}} \le \frac{\underline{m}\varepsilon}{2} + \frac{\varepsilon_m}{\underline{m}\varepsilon/2} \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Hence we may define

$$\widetilde{\phi}_{1/m}(t) := \phi_{\mathrm{rec}}(\widetilde{\phi}_m(t)).$$

and we have $\widetilde{\phi}_{1/m}(t) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with

$$L,\ \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 \underline{m}^{-1}, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + \log^3 \underline{m}^{-1}, \quad S \lesssim \log^4 \varepsilon^{-1} + \log^4 \underline{m}^{-1}.$$

□

Lemma I.7 (ReLU approximation of the rescaled input $\boldsymbol{x}/m_t$)**.** *Let $M_x \geq 1$ and assume $\|\boldsymbol{x}\|_\infty \leq M_x$. For every $\varepsilon \in (0, 1/2]$ there exists a vector-valued ReLU network*

$$\widetilde{\phi}(\boldsymbol{x}, t) = \big(\widetilde{\phi}_1(\boldsymbol{x}, t), \ldots, \widetilde{\phi}_D(\boldsymbol{x}, t)\big) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$$

with

$$\begin{aligned} L,\, \log B &\lesssim \log^2 \varepsilon^{-1} + \log^2(M_x \vee \underline{m}^{-1}), \\ \|\boldsymbol{W}\|_\infty &\lesssim D\big(\log^3 \varepsilon^{-1} + \log^3(M_x \vee \underline{m}^{-1})\big), \\ S &\lesssim D\big(\log^4 \varepsilon^{-1} + \log^4(M_x \vee \underline{m}^{-1})\big), \end{aligned}$$

such that

$$\Big\|\widetilde{\phi}(\boldsymbol{x}, t) - \frac{\boldsymbol{x}}{m_t}\Big\|_{L^\infty(\{\|\boldsymbol{x}\|_\infty \leq M_x\} \times [\underline{t}, \bar{t}])} \leq \varepsilon.$$

Proof. Let $\widetilde{\phi}_{1/m}$ be the network from Theorem I.6. Choose an auxiliary accuracy

$$\varepsilon_0 := \frac{\varepsilon}{4(1 \vee M_x \vee \underline{m}^{-1})}.$$

Refine $\widetilde{\phi}_{1/m}$ so that

$$\|\widetilde{\phi}_{1/m} - m_t^{-1}\|_{L^\infty([\underline{t}, \bar{t}])} \leq \varepsilon_0.$$

For each coordinate $j = 1, \ldots, D$, apply the multiplication lemma (Theorem I.3) on $[-M_x, M_x] \times [-\underline{m}^{-1}, \underline{m}^{-1}]$ to obtain a ReLU network $\phi_{\mathrm{mult},j} \in \mathcal{NN}(L_1, \boldsymbol{W}_1, S_1, B_1)$ satisfying

$$\begin{aligned} L_1 &\lesssim \log \varepsilon_0^{-1} + \log(M_x \vee \underline{m}^{-1}), \\ \|\boldsymbol{W}_1\|_\infty &= 96, \\ S_1 &\lesssim \log \varepsilon_0^{-1} + \log(M_x \vee \underline{m}^{-1}), \\ \log B_1 &= 2\log(M_x \vee \underline{m}^{-1}), \end{aligned}$$

such that

$$\Big|\phi_{\mathrm{mult},j}\big(x_j, \widetilde{\phi}_{1/m}(t)\big) - x_j m_t^{-1}\Big| \leq \frac{\varepsilon}{2} + 2(1 \vee M_x \vee \underline{m}^{-1})\, \varepsilon_0 \leq \varepsilon.$$

Define

$$\widetilde{\phi}_j(\boldsymbol{x}, t) := \phi_{\mathrm{mult},j}\big(x_j, \widetilde{\phi}_{1/m}(t)\big), \qquad j = 1, \ldots, D.$$

Parallel stacking of the D coordinate subnetworks gives the stated vector-valued approximant. □

Lemma I.8 (ReLU approximation of positive powers of σ_t)**.** *For any $k \in \mathbb{N}_+$ and any $\varepsilon \in (0, 1/2]$ such that $\varepsilon \leq \underline{t} < \bar{t} < \infty$, there exists a ReLU network $\phi_{\sigma^k} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$\begin{aligned} &L \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + k, \\ &S \lesssim \log^4 \varepsilon^{-1} + \log^4 k + k \log \varepsilon^{-1}, \quad \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \end{aligned}$$

such that

$$\big\|\phi_{\sigma^k}(t) - \sigma_t^k\big\|_{L^\infty([\underline{t}, \bar{t}])} \leq 3\varepsilon.$$

Proof. By Theorem I.2, applied with accuracy parameter ε/k, there exists a neural network $\phi_\sigma(t) \in \mathcal{NN}(L_\sigma, \boldsymbol{W}_\sigma, S_\sigma, B_\sigma)$ with

$$\begin{aligned} &L \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + \log^3 k, \\ &S \lesssim \log^4 \varepsilon^{-1} + \log^4 k, \quad \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \end{aligned} \tag{255}$$

such that

$$|\phi_\sigma(t) - \sigma_t| \le \varepsilon/k, \quad \text{for all } t \in [\underline{t}, \overline{t}].$$

Notice that $\sigma_t \in [0,1]$ for all $t \in [\underline{t}, \overline{t}]$. If $k = 1$, take $\phi_{\sigma^k} := \phi_\sigma$, and the claim follows. Assume now that $k \ge 2$. Apply Theorem I.3 with $d = k$, $C = 1$, and with all k true inputs equal to σ_t and all k perturbed inputs equal to $\phi_\sigma(t)$. The shared-input form is allowed by the final sentence of Theorem I.3. Thus there exists a ReLU network $\phi_{\text{multi}} \in \mathcal{NN}(L_1, \boldsymbol{W}_1, S_1, B_1)$ with

$$L_1 \lesssim \log k \, \log \varepsilon^{-1}, \quad \|\boldsymbol{W}_1\|_\infty = 48k, \quad S_1 \lesssim k \log \varepsilon^{-1}, \quad B_1 = 1 \tag{256}$$

such that

$$\Big|\phi_{\text{multi}}\big(\underbrace{\phi_\sigma(t), \dots, \phi_\sigma(t)}_{k \text{ copies}}\big) - \sigma_t^k\Big| \le \varepsilon + k\frac{\varepsilon}{k} \le 2\varepsilon. \tag{257}$$

For all $t \in [\underline{t}, \overline{t}]$, we define

$$\phi_{\sigma^k}(t) := \phi_{\text{multi}}\big(\underbrace{\phi_\sigma(t), \dots, \phi_\sigma(t)}_{k \text{ copies}}\big).$$

Then by the configurations given in (255) and (256) and the bound in (257), we obtain that there exists a ReLU network $\phi_{\sigma^k} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with

$$\begin{aligned} &L \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + k, \\ &S \lesssim \log^4 \varepsilon^{-1} + \log^4 k + k \log \varepsilon^{-1}, \quad \log B \lesssim \log \varepsilon^{-1} + \log k, \end{aligned}$$

such that

$$\big\|\phi_{\sigma^k}(t) - \sigma_t^k\big\|_{L^\infty([\underline{t}, \overline{t}])} \le 3\varepsilon.$$

□

Lemma I.9 (ReLU approximation of inverse powers of σ_t)**.** *For any $k \in \mathbb{N}_+$ and any $\varepsilon \in (0,1]$, there exists a ReLU network $\phi_{\sigma^{-k}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$\begin{aligned} L &\lesssim \log^2 \varepsilon^{-1} + k^2 \log^2 \sigma_{\underline{t}}^{-2}, \\ \|\boldsymbol{W}\|_\infty &\lesssim \log^3 \varepsilon^{-1} + k^3 \log^3 \sigma_{\underline{t}}^{-2}, \\ S &\lesssim \log^4 \varepsilon^{-1} + k^4 \log^4 \sigma_{\underline{t}}^{-2} + k \log(e/\varepsilon) + k^2 \log(e\sigma_{\underline{t}}^{-2}), \\ \log B &\lesssim \log^2 \varepsilon^{-1} + k \log^2 \sigma_{\underline{t}}^{-2}, \end{aligned}$$

such that

$$\big\|\phi_{\sigma^{-k}}(t) - \sigma_t^{-k}\big\|_{L^\infty([\underline{t}, \overline{t}])} \le 4\varepsilon.$$

Proof. Set

$$\varepsilon_{\sigma,k} := \min\Big\{\frac{\sigma_{\underline{t}}^k \varepsilon^2}{12}, \frac{\underline{t}}{2}\Big\}.$$

Apply Theorem I.8 with accuracy parameter $\varepsilon_{\sigma,k}$. By construction $\varepsilon_{\sigma,k} \le \underline{t}$, so this parameter lies in the admissible range of Theorem I.8. Moreover,

$$\log(\varepsilon_{\sigma,k}^{-1}) \le C\big(\log \varepsilon^{-1} + k \log \sigma_{\underline{t}}^{-1} + 1\big),$$

where the fixed compact time interval contributes only to the constant through $\log \underline{t}^{-1}$. Then, there exists a ReLU network $\phi_{\sigma^k} \in \mathcal{NN}(L_1, \boldsymbol{W}_1, S_1, B_1)$ with

$$\begin{aligned} L_1 &\lesssim \log^2 \varepsilon^{-1} + k^2 \log^2 \sigma_{\underline{t}}^{-2} + \log^2 k, \\ \|\boldsymbol{W}_1\|_\infty &\lesssim \log^3 \varepsilon^{-1} + k^3 \log^3 \sigma_{\underline{t}}^{-2} + k, \\ S_1 &\lesssim \log^4 \varepsilon^{-1} + k^4 \log^4 \sigma_{\underline{t}}^{-2} + \log^4 k + k \log(e/\varepsilon) + k^2 \log(e\sigma_{\underline{t}}^{-2}), \\ \log B_1 &\lesssim \log^2 \varepsilon^{-1} + k^2 \log^2 \sigma_{\underline{t}}^{-2} + \log^2 k, \end{aligned}$$

such that

$$\left\|\phi_{\sigma^k}(t) - \sigma_t^k\right\|_{L^\infty([\underline{t},\bar{t}])} \le 3\varepsilon_{\sigma,k} \le \frac{\sigma_{\underline{t}}^k \varepsilon^2}{4}.$$

Note that for any $t \in [\underline{t}, \bar{t}]$, we have $\sigma_{\underline{t}}^k \lesssim \sigma_t^k \le 1$.

Let $\varepsilon_{\mathrm{rec}} := \sigma_{\underline{t}}^k \varepsilon/2$ and let ϕ_{rec} be the network in Theorem I.4 with parameter $\varepsilon_{\mathrm{rec}}$. Since $\varepsilon_{\mathrm{rec}} \le \sigma_{\underline{t}}^k \le \sigma_t^k$ and $\sigma_t^k \le 1 \le \varepsilon_{\mathrm{rec}}^{-1}$, the domain condition in Theorem I.4 holds for every $t \in [\underline{t}, \bar{t}]$. Then, we have $\phi_{\mathrm{rec}} \in \mathcal{NN}(L_{\mathrm{rec}}, \boldsymbol{W}_{\mathrm{rec}}, S_{\mathrm{rec}}, B_{\mathrm{rec}})$ with

$$\begin{aligned} L_{\mathrm{rec}} &\lesssim \log^2 \varepsilon^{-1} + k^2 \log^2 \sigma_{\underline{t}}^{-2}, \\ \|\boldsymbol{W}_{\mathrm{rec}}\|_\infty &\lesssim \log^3 \varepsilon^{-1} + k^3 \log^3 \sigma_{\underline{t}}^{-2}, \\ S_{\mathrm{rec}} &\lesssim \log^4 \varepsilon^{-1} + k^4 \log^4 \sigma_{\underline{t}}^{-2}, \\ \log B_{\mathrm{rec}} &\lesssim \log \varepsilon^{-1} + k \log \sigma_{\underline{t}}^{-2}, \end{aligned}$$

such that for all $t \in [\underline{t}, \bar{t}]$, it holds that

$$\left|\phi_{\mathrm{rec}}(\phi_{\sigma^k}(t)) - \frac{1}{\sigma_t^k}\right| \le \varepsilon_{\mathrm{rec}} + \frac{|\phi_{\sigma^k}(t) - \sigma_t^k|}{\varepsilon_{\mathrm{rec}}} \le \frac{\sigma_{\underline{t}}^k \varepsilon}{2} + \frac{\sigma_{\underline{t}}^k \varepsilon^2/4}{\sigma_{\underline{t}}^k \varepsilon/2} \le \varepsilon.$$

Let $\phi_{\sigma^{-k}}(t) := \phi_{\mathrm{rec}}(\phi_{\sigma^k}(t))$. Adding the displayed size bounds for ϕ_{σ^k} and ϕ_{rec}, and absorbing the lower-order $\log k$ terms into the stated k-dependent bounds, gives the claimed network size. □

I.2. Support lemmas for generalization analysis

Lemma I.10 (Unconditional Fisher bound for the smoothed data law)**.** *For the VP process* (4)*, for each* $t > 0$*,*

$$\mathbb{E}_{\boldsymbol{X}_t}\left[\|\nabla \log p_t(\boldsymbol{X}_t)\|_2^2\right] \le D\sigma_t^{-2}.$$

Proof. Write $\boldsymbol{X}_t = m_t \boldsymbol{X}_0 + \sigma_t \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_D)$ is independent of $\boldsymbol{X}_0$. Tweedie's formula gives

$$\nabla \log p_t(\boldsymbol{y}) = \mathbb{E}\left[-\frac{\boldsymbol{Z}}{\sigma_t} \,\middle|\, \boldsymbol{X}_t = \boldsymbol{y}\right].$$

Therefore Jensen's inequality and the tower property imply

$$\mathbb{E}_{\boldsymbol{X}_t}\|\nabla \log p_t(\boldsymbol{X}_t)\|_2^2 \le \mathbb{E}\left\|\frac{\boldsymbol{Z}}{\sigma_t}\right\|_2^2 = D\sigma_t^{-2}.$$

This is an unconditional Fisher-information bound under the marginal law of $\boldsymbol{X}_t$; no conditional bound given $\boldsymbol{X}_0$ is used. □

Lemma I.11 (A variant lemma of (Fu & Lee, 2025, Lemma 25))**.** *Fix any* $0 < \underline{t} < \bar{t} < \infty$*. Let* $s : \mathbb{R}^D \times [\underline{t}, \bar{t}] \to \mathbb{R}^D$ *and* $s' : \mathbb{R}^D \times [\underline{t}, \bar{t}] \to \mathbb{R}^D$ *be any Borel functions. Then, for all* $\boldsymbol{x} \in \mathbb{R}^D$*, it holds that*

$$\begin{aligned} &\left(\ell_{\underline{t},\bar{t}}(s, \boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s', \boldsymbol{x})\right)^2 \\ &\le 3\left(\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\left[\|s(\boldsymbol{X}_t, t)\|_2^2 + \|s'(\boldsymbol{X}_t, t)\|_2^2\right]\mathrm{d}t + 4D\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t\right) \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\left[\left\|s(\boldsymbol{X}_t, t) - s'(\boldsymbol{X}_t, t)\right\|_2^2\right]\mathrm{d}t. \end{aligned}$$

Proof.

$$
\begin{aligned}
&\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x})-\ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})\\
&=\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t)+\frac{\boldsymbol{X}_t-m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2-\Big\|s'(\boldsymbol{X}_t,t)+\frac{\boldsymbol{X}_t-m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t\\
&=\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\|s(\boldsymbol{X}_t,t)\|_2^2+2\Big\langle s(\boldsymbol{X}_t,t),\frac{\boldsymbol{X}_t-m_t\boldsymbol{x}}{\sigma_t^2}\Big\rangle-\|s'(\boldsymbol{X}_t,t)\|_2^2-2\Big\langle s'(\boldsymbol{X}_t,t),\frac{\boldsymbol{X}_t-m_t\boldsymbol{x}}{\sigma_t^2}\Big\rangle\Big]\mathrm{d}t\\
&=\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big(s(\boldsymbol{X}_t,t)-s'(\boldsymbol{X}_t,t)\big)^\top\big(s(\boldsymbol{X}_t,t)+s'(\boldsymbol{X}_t,t)\big)+\big(s(\boldsymbol{X}_t,t)-s'(\boldsymbol{X}_t,t)\big)^\top\frac{2(\boldsymbol{X}_t-m_t\boldsymbol{x})}{\sigma_t^2}\Big]\mathrm{d}t\\
&=\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big(s(\boldsymbol{X}_t,t)-s'(\boldsymbol{X}_t,t)\big)^\top\Big(s(\boldsymbol{X}_t,t)+s'(\boldsymbol{X}_t,t)+\frac{2(\boldsymbol{X}_t-m_t\boldsymbol{x})}{\sigma_t^2}\Big)\Big]\mathrm{d}t.
\end{aligned}
$$

Applying the Cauchy-Schwarz inequality and Young's inequality, we obtain that

$$
\begin{aligned}
&|\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x})-\ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})|\\
&\le\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big\|s(\boldsymbol{X}_t,t)-s'(\boldsymbol{X}_t,t)\big\|\cdot\Big\|s(\boldsymbol{X}_t,t)+s'(\boldsymbol{X}_t,t)+\frac{2(\boldsymbol{X}_t-m_t\boldsymbol{x})}{\sigma_t^2}\Big\|\Big]\mathrm{d}t\\
&\hspace{20em}\text{(by Cauchy-Schwarz inequality)}\\
&\le\sqrt{\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big\|s(\boldsymbol{X}_t,t)-s'(\boldsymbol{X}_t,t)\big\|_2^2\Big]\mathrm{d}t}\cdot\sqrt{\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t)+s'(\boldsymbol{X}_t,t)+\frac{2(\boldsymbol{X}_t-m_t\boldsymbol{x})}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t}.\\
&\hspace{20em}\text{(by Young's inequality)}
\end{aligned}
$$

For the VP process, $\boldsymbol{X}_t=m_t\boldsymbol{x}+\sigma_t\boldsymbol{Z}$ under $\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}$, with $\boldsymbol{Z}\sim\mathcal{N}(0,\boldsymbol{I}_D)$. Hence

$$
\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|\frac{\boldsymbol{X}_t-m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2\Big]=\frac{D}{\sigma_t^2}.
$$

Then it holds that

$$
\begin{aligned}
&\sqrt{\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t)+s'(\boldsymbol{X}_t,t)+\frac{2(\boldsymbol{X}_t-m_t\boldsymbol{x})}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t}\\
&\le\sqrt{3\int_{\underline{t}}^{\bar{t}}\Big(\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2+\|s'(\boldsymbol{X}_t,t)\|_2^2\big]+4\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|\frac{\boldsymbol{X}_t-m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2\Big]\Big)\mathrm{d}t}\\
&\le\sqrt{3\int_{\underline{t}}^{\bar{t}}\Big(\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2+\|s'(\boldsymbol{X}_t,t)\|_2^2\big]+4D\sigma_t^{-2}\Big)\mathrm{d}t}.
\end{aligned}
\tag{258}
$$

Hence we obtain

$$
\begin{aligned}
&\sqrt{\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t)+s'(\boldsymbol{X}_t,t)+\frac{2(\boldsymbol{X}_t-m_t\boldsymbol{x})}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t}\\
&\le\sqrt{3\int_{\underline{t}}^{\bar{t}}\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2+\|s'(\boldsymbol{X}_t,t)\|_2^2\big]\mathrm{d}t+12D\int_{\underline{t}}^{\bar{t}}\sigma_t^{-2}\mathrm{d}t}.
\end{aligned}
\tag{259}
$$

Thus, again by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\big(\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})\big)^2 \\
&\le \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big\|s(\boldsymbol{X}_t,t) - s'(\boldsymbol{X}_t,t)\big\|_2^2\Big]\mathrm{d}t \cdot \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t) + s'(\boldsymbol{X}_t,t) + \frac{2(\boldsymbol{X}_t - m_t\boldsymbol{x})}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t \\
&\le 3\Big(\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2 + \|s'(\boldsymbol{X}_t,t)\|_2^2\big]\mathrm{d}t + 4D\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t\Big)\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big\|s(\boldsymbol{X}_t,t) - s'(\boldsymbol{X}_t,t)\big\|_2^2\Big]\mathrm{d}t.
\end{aligned}
$$

□

Lemma I.12 (Uniform boundedness of clipped denoising losses)**.** *Fix $0 < \underline{t} < \bar{t} < \infty$. Let $s : \mathbb{R}^D \times [\underline{t},\bar{t}] \to \mathbb{R}^D$ be Borel and assume that*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|s(\boldsymbol{x},t)\|_\infty \le C_{\mathrm{clip}}\sigma_t^{-1}(D \vee \log n)^{1/2}, \qquad t \in [\underline{t},\bar{t}].$$

Then

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} |\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x})| \le C(1 + C_{\mathrm{clip}}^2)(D \vee \log n)^2 \int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t.$$

Consequently, if s and s' both satisfy the same clipped envelope, then

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} |\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})| \le C(1 + C_{\mathrm{clip}}^2)(D \vee \log n)^2 \int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t.$$

Proof. For the VP process, $\boldsymbol{X}_t \mid \boldsymbol{X}_0 = \boldsymbol{x} \sim \mathcal{N}(m_t\boldsymbol{x}, \sigma_t^2\boldsymbol{I}_D)$, and

$$\nabla \log p_t(\boldsymbol{X}_t \mid \boldsymbol{X}_0 = \boldsymbol{x}) = -\frac{\boldsymbol{X}_t - m_t\boldsymbol{x}}{\sigma_t^2}.$$

The clipped envelope gives

$$\|s(\boldsymbol{y},t)\|_2^2 \le DC_{\mathrm{clip}}^2\sigma_t^{-2}(D \vee \log n) \le CC_{\mathrm{clip}}^2(D \vee \log n)^2\sigma_t^{-2}.$$

Also

$$\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\left\|\frac{\boldsymbol{X}_t - m_t\boldsymbol{x}}{\sigma_t^2}\right\|_2^2 = D\sigma_t^{-2} \le (D \vee \log n)^2\sigma_t^{-2}.$$

Using $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$ inside the definition of $\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x})$ and integrating over t proves the first display. The second follows from the triangle inequality applied to $\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})$. □

I.2.1. Bernstein's inequality

Theorem I.13 (Bernstein's inequality for bounded distributions (Vershynin, 2018))**.** *Let $X_1,\dots,X_n$ be independent random variables such that $|X_i| \le K$ for all $i \in [n]$. Then, for every $t \ge 0$, we have*

$$\Pr\Big[\Big|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\Big| \ge t\Big] \le 2\exp\Big(-\frac{t^2/2}{\sum_{i=1}^n \mathbb{E}[X_i^2] + Kt/3}\Big).$$

In other words, with probability at least $1-\delta$, it holds that

$$\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\Big| \le \frac{t}{n} \le \frac{2K\log(2/\delta)}{3n} + \sqrt{\frac{\frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i^2]\log(2/\delta)}{n}}.$$

I.2.2. Covering number bounds

Lemma I.14 ((Suzuki, 2019, Lemma 3)). *For any input dimension D_{in} and any $\tau > 0$, the covering number of $\mathcal{NN}(L, \boldsymbol{W}, S, B)$ on $[0,1]^{D_{\text{in}}}$ can be bounded by*

$$\log \mathcal{N}(\tau, \mathcal{NN}(L, \boldsymbol{W}, S, B), \|\cdot\|_{L^\infty([0,1]^{D_{\text{in}}})}) \lesssim SL \log\big(\tau^{-1} L(\|\boldsymbol{W}\|_\infty + 1)(B \vee 1)\big).$$

In the score-estimation oracle $D_{\text{in}} = D+1$, and the difference between D_{in} and D is absorbed into the displayed polynomial ambient factors.

I.3. Auxiliary lemmas

We use the following distribution-estimation lemmas from (Oko et al., 2023) as black boxes. For the slabwise lemma we use the standard clipped-score form with coordinatewise clipping at a fixed polynomial scale; Sec. G supplies the integrated score-error, clipping, and network-size inputs for the learned scores. The final W_1 corollary is therefore unconditional as a perturbation theorem given those score inputs.

Lemma I.15 (Early stopping for the VP forward law). *For the forward law P_t and P_0 supported on $[0,1]^D$,*

$$\mathsf{W}_1(P_0, P_{t_0}) \le \sqrt{D}\,\{1 - m_{t_0} + \sigma_{t_0}\}. \tag{260}$$

Consequently, under the regular VP schedule convention (253), $\mathsf{W}_1(P_0, P_{t_0}) \le C\sqrt{Dt_0}$ for $t_0 \le 1$.

Proof. Couple $\boldsymbol{X}_0 \sim P_0$ with $\boldsymbol{X}_{t_0} = m_{t_0}\boldsymbol{X}_0 + \sigma_{t_0}\boldsymbol{Z}$ using an independent $\boldsymbol{Z} \sim \mathcal{N}(0, I_D)$. Since $\mathcal{M} \subset [0,1]^D$, $\|\boldsymbol{X}_0\|_2 \le \sqrt{D}$ almost surely and $\mathbb{E}\|\boldsymbol{Z}\|_2 \le \sqrt{D}$. Therefore

$$\mathsf{W}_1(P_0, P_{t_0}) \le \mathbb{E}\|\boldsymbol{X}_{t_0} - \boldsymbol{X}_0\|_2 \le (1 - m_{t_0})\sqrt{D} + \sigma_{t_0}\sqrt{D}.$$

The final display follows from (253). □

Lemma I.16 (Terminal Gaussian initialization with ambient factor). *Let $\overline{P}_{t_0}$ be the reverse-process output law initialized from the exact forward terminal law P_T, and let $\widehat{P}_{t_0}$ be the corresponding output law initialized from the Gaussian proxy $\widehat{P}_T$. Under the regular VP schedule convention (253),*

$$\mathbb{E}\big[\mathsf{W}_1(\overline{P}_{t_0}, \widehat{P}_{t_0})\big] \le C\mathsf{TV}(P_T, \widehat{P}_T) \le C\sqrt{D}\,\exp(-c_{\text{mix}}T). \tag{261}$$

Proof. The first inequality is the terminal perturbation bound from (Oko et al., 2023, Lemma D.6). For the second, condition on $\boldsymbol{X}_0 = \boldsymbol{y}$. The terminal conditional law is $\mathcal{N}(m_T\boldsymbol{y}, \sigma_T^2 I_D)$ with $\sigma_T^2 = 1 - m_T^2$, while $\widehat{P}_T = \mathcal{N}(0, I_D)$. When $m_T \le 1/2$, the Gaussian KL divergence is bounded by CDm_T^2 because $\|\boldsymbol{y}\|_2 \le \sqrt{D}$ and $\sigma_T^2 = 1 - m_T^2$. Pinsker's inequality gives a $C\sqrt{D}\,m_T$ bound for each conditional law. When $m_T > 1/2$, the same bound follows after increasing C from the trivial estimate $\mathsf{TV} \le 1$. Convexity over the mixture gives $\mathsf{TV}(P_T, \widehat{P}_T) \le C\sqrt{D}\,m_T$. The regular VP terminal decay $m_T \le C_{\text{mix}}e^{-c_{\text{mix}}T}$ proves the claim. □

Lemma I.17 (Slabwise W_1 perturbation, after (Oko et al., 2023, Lemma D.7)). *Fix a partition $[t_0, T] = \bigcup_{k=1}^{K_*}[t_{k-1}, t_k]$. Let $\overline{P}_{t_0}^{(k)}$ denote the reverse-process output law initialized from the exact terminal law P_T, using the learned score on the first k reverse-time slabs and the true score on the remaining slabs. Thus $\overline{P}_{t_0}^{(0)} = \overline{P}_{t_0}$ is the exact-score reverse law and $\overline{P}_{t_0}^{(K_*)}$ is the learned-score reverse law before replacing P_T by the Gaussian proxy. The increment $\mathsf{W}_1(\overline{P}_{t_0}^{(k-1)}, \overline{P}_{t_0}^{(k)})$ is therefore precisely the perturbation caused by using the learned score on the single slab $[t_{k-1}, t_k]$. Suppose that, for a fixed exponent a_{clip}, the score used by the reverse process satisfies*

$$\|\widehat{s}(\cdot, t)\|_\infty \le C_{\text{clip}}\sigma_t^{-1}(D \vee \log n)^{a_{\text{clip}}}.$$

Then, the following holds for all $k = 1, 2, \ldots, K_*$:

$$\mathsf{W}_1(\overline{P}_{t_0}^{(k-1)}, \overline{P}_{t_0}^{(k)}) \le C(D \vee \log n)^{a_{\mathrm{W}}}\left\{\sqrt{t_k \log n}\sqrt{\int_{t=t_{k-1}}^{t_k} \mathbb{E}_{\boldsymbol{x}_t}\|\widehat{s}(\boldsymbol{x}_t, t) - \nabla \log p_t(\boldsymbol{x}_t)\|_2^2 \mathrm{d}t} + n^{-\frac{\beta+1}{d+2\beta}}\right\}.$$

Therefore, we have that

$$\begin{aligned}&\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n}\mathsf{W}_1(\overline{P}_{t_0}^{(k-1)},\overline{P}_{t_0}^{(k)})\\&\le C(D\vee\log n)^{a_{\mathrm{W}}}\left\{\sqrt{t_k\log n}\Big(\mathbb{E}_{\{\boldsymbol{x}^{(i)}\}_{i=1}^n}\int_{t=t_{k-1}}^{t_k}\mathbb{E}_{\boldsymbol{x}_t}\big[\|\widehat{s}(\boldsymbol{x}_t,t)-\nabla\log p_t(\boldsymbol{x}_t)\|_2^2\big]\mathrm{d}t\Big)^{1/2}+n^{-\frac{\beta+1}{d+2\beta}}\right\}.\end{aligned}$$

The constants C, a_{W} may depend on the fixed clipping exponent a_{clip}, but not on n, D, t_k, or $p_{\min}$.