INTRINSIC WASSERSTEIN RATES FOR SCORE-BASED DIFFUSION MODELS ON SMOOTH MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many high-dimensional data distributions are supported on low-dimensional manifolds, whereas diffusion models learn score functions in the ambient space. This raises a basic question: can empirical risk minimization over ambient score-network classes attain a rate governed by intrinsic dimension without first estimating a manifold representation from the observations? We study this question for distributions with strictly positive β-Hölder densities on compact d-dimensional smooth manifolds embedded in $\mathbb{R}^D$. The main technical challenge is the low-noise regime, where the score becomes singular as the noise level vanishes. To address this challenge, we construct tangent-cell ReLU approximations at moderate and high noise and a projection-centered ReLU approximation at low noise, with explicit ambient-dimension dependence. For $d > 2$, under the stated geometry, regularity, and optimization conditions, the cube-projected output of the continuous-time sampler has expected 1-Wasserstein error $D^{\mathcal{O}_\beta(d)} n^{-(\beta+1)/(d+2\beta)}$ up to logarithmic factors. The result therefore establishes the intrinsic sample-size exponent convergence rate for ambient score networks, without a preliminary data-dependent dimension-reduction step.

1 INTRODUCTION

Diffusion models learn an ambient vector field and generate samples by reversing a noising process in $\mathbb{R}^D$ (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2021; Ho et al., 2020; Vahdat et al., 2021). Many data distributions are nevertheless modeled as being supported on, or concentrated near, a nonlinear d-dimensional manifold with $d \ll D$ (Fefferman et al., 2016). This motivates the following question: can empirical risk minimization over score networks with ambient inputs and outputs attain a statistical rate governed by d without first constructing a manifold representation from the observed data?

The difficulty is not merely to cover the manifold efficiently. A manifold-supported law is singular with respect to ambient Lebesgue measure and has no ambient score at time zero. At positive time, Gaussian smoothing creates a score whose geometry changes with the noise scale. At moderate noise, tangent patches provide an intrinsic cover. Near time zero, however, a noise-scale cover would require too many local cells, while the smoothed-density denominator contains a rapidly vanishing normal Gaussian factor. Our low-noise construction controls both this singularity and the nearest-point projection without reverting to a generic D-variate approximation.

We observe n independent samples from a law P_0 supported on a compact d-dimensional smooth manifold $\mathcal{M} \subset [0,1]^D$ and having a strictly positive β-Hölder density with respect to Riemannian volume. The idealized estimator approximately minimizes separate denoising score-matching objectives over interval-specific ambient ReLU classes; the expectations over diffusion time and Gaussian corruption are evaluated exactly. It is not supplied with tangent-space, local-basis, or coordinate estimates computed from the training sample. The proof instead constructs deterministic comparison networks: tangent-cell networks at moderate and high noise, and a projection-centered low-noise network after removing the common normal Gaussian factor.

Let $\widehat{P}^{\mathrm{box}}_{t_0}$ denote the conditional output law of the early-stopped reverse SDE after projection onto $[0,1]^D$. Under the conditions of Theorem 2.2, for $d > 2$ and fixed geometry and density parameters,

$$\mathbb{E}\,\mathsf{W}_1\big(P_0, \widehat{P}^{\mathrm{box}}_{t_0}\big) \le \widetilde{\mathcal{O}}\big(D^{\mathcal{O}_\beta(d)} n^{-(\beta+1)/(d+2\beta)}\big),$$

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

where the expectation is over the training sample and optimizer randomness. Thus, the sample-size exponent depends on the intrinsic dimension, with polynomial ambient dependence. The estimator requires knowledge of d, β, and suitable geometry and density bounds. An exponential-integrator discretization preserves this rate under the grid-weighted optimization condition, while retaining exact Gaussian-corruption expectations; see Corollary H.14.

Our contributions are:

(i) an expected W_1 bound for interval-wise approximate empirical risk minimization over ambient score-network classes, with the intrinsic sample-size exponent under explicit conditions;

(ii) noise-adaptive deterministic approximation schemes based on tangent cells and a projection-centered Laplace expansion, with a finite-anchor Gauss–Newton implementation of nearest-point projection (Wong, 2001; Nocedal & Wright, 2006); and

(iii) explicit network-complexity bounds, with aggregate width and sparsity having leading sample dependence $n^{d/(d+2\beta)}$ and displayed ambient, geometry, and density factors.

Relation to prior work. Ambient-dimensional score-estimation and sampling theories give sharp guarantees under several density and functional assumptions (Lee et al., 2022; Yang & Wibisono, 2022; Benton et al., 2024; Conforti et al., 2023; Oko et al., 2023; Dou et al., 2024), but do not directly yield intrinsic rates for singular laws on nonlinear manifolds. Other work studies manifold detection, score singularities, and generalization through Wasserstein dimension (Pidstrigach, 2022; Lu et al., 2023; Li et al., 2026; Chakraborty et al., 2026); these address complementary questions.

The closest results are manifold-aware diffusion analyses. Linear-support models avoid curvature (Oko et al., 2023; Chen et al., 2023b), while some nonlinear-manifold guarantees have slower rates (Yakovlev & Puchkin, 2025). Tang & Yang (2024) obtain the same intrinsic W_1 exponent without explicitly estimating the manifold. Version 3 of Azangulov et al. (2024) obtains the same exponent with no explicit D factor in its stated W_1 rate. Its analysis first constructs anchors, local subspaces, and bases from the training sample, then performs local empirical risk minimization; it also analyzes a modified sampler. It therefore gives better ambient-dimension dependence under a different estimator and sampling interface. Zhang et al. (2026) provide a closely related projection-centered analysis; the ambient-dimension dependence of their projection-network constant is not quantified in the corresponding approximation lemma. Our finite-anchor Gauss–Newton construction implements the projection in intrinsic coordinates with explicit dimension-dependent complexity bounds, while Gaussian-factor removal controls the low-noise denominator. The resulting comparison networks enter an oracle inequality for ambient score-network ERM; the learner is not supplied with manifold coordinates. These differences concern the construction and estimator interface, not a new sample-size exponent or an improved ambient prefactor. Detailed theorem-level comparisons appear in Sec. A.1 and A.2.

Organization. Sec. 2 describes the learning pipeline and states the main result. Sec. 3 explains why two noise regimes require different constructions. Sec. 4 turns the deterministic approximants into learned scores, and Sec. 5 exposes the distribution-level error decomposition. The appendices contain all formal proofs, network-complexity bounds, sample-size checks, and detailed comparisons.

2 Preliminaries, Diffusion Models, and Main Result

We observe $n \geq 3$ independent samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \sim P_0$ in $\mathbb{R}^D$ and seek a generative distribution whose expected 1-Wasserstein error is governed by the intrinsic dimension. For $P, Q \in \mathcal{P}_1(\mathbb{R}^D)$, the Borel probability measures on $\mathbb{R}^D$ with finite first moment, let

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

where $\Gamma(P,Q)$ is the set of their couplings. This metric remains meaningful for singular manifold-supported measures and uses the ambient geometry of the sampling space; see, for example, Villani (2009) for the standard optimal-transport definition and background.

This section specifies the diffusion and learning pipeline, states the distributional assumptions used by the construction, and gives the resulting end-to-end W_1 guarantee.

Notation and conventions. Throughout, $\beta > 0$ denotes the Hölder smoothness of the data density, and $m_\beta := \lceil \beta \rceil - 1$ is the largest integer strictly smaller than β. For nonnegative a, b, write $a \lesssim b$ if $a \le Cb$ and $a \lesssim_{\log} b$ if $a \le C(1 + \log n + \log D)^c b$, with constants independent of the displayed varying quantities except for fixed parameters explicitly named. Write $a \asymp b$ when $a \lesssim b$ and $b \lesssim a$. The subscript in $\mathcal{O}_\beta$ permits dependence on fixed β, whereas $\widetilde{\mathcal{O}}$ suppresses logarithmic factors. In main-text rate displays, fixed powers of the Hölder radius B_0 are absorbed into the constants; the formal appendix statements retain them explicitly. Apart from this convention, none of the notation hides powers of n or D, geometry or density factors, or optimization error unless explicitly stated. We use $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$; see Sec. B.3 for the complete conventions.

2.1 Score-based Diffusion Models

Forward process. We use the variance-preserving Ornstein–Uhlenbeck process, a standard continuous-time noising process for score-based diffusion models (Song et al., 2021),

$$\mathrm{d}\boldsymbol{X}_t = -\boldsymbol{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t, \qquad \boldsymbol{X}_0 \sim P_0, \tag{1}$$

where $\boldsymbol{B}_t$ is a standard D-dimensional Brownian motion independent of $\boldsymbol{X}_0$, so that $\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim \mathcal{N}(m_t \boldsymbol{X}_0, \sigma_t^2 \boldsymbol{I}_D)$ with $m_t = e^{-t}$ and $\sigma_t^2 = 1 - e^{-2t}$. Let P_t denote the law of $\boldsymbol{X}_t$, let p_t denote its density for $t > 0$, and write $s^*(\boldsymbol{x}, t) = \nabla \log p_t(\boldsymbol{x})$ for the positive-time score.

Reverse process. For a terminal time $T > 0$, the time reversal $\boldsymbol{Y}_t = \boldsymbol{X}_{T-t}$ satisfies the standard reverse-time SDE (Anderson, 1982; Song et al., 2021)

$$\mathrm{d}\boldsymbol{Y}_t = \big(\boldsymbol{Y}_t + 2s^*(\boldsymbol{Y}_t, T - t)\big)\mathrm{d}t + \sqrt{2}\mathrm{d}\overline{\boldsymbol{B}}_t, \qquad \boldsymbol{Y}_0 \sim P_T, \qquad 0 \le t < T.$$

Here $\overline{\boldsymbol{B}}_t$ is a standard D-dimensional Brownian motion with respect to the reverse filtration. The endpoint $\boldsymbol{Y}_T = \boldsymbol{X}_0$ is defined by continuity, not by evaluating an ambient score at time zero. Thus, exact reverse-time sampling requires both the terminal law P_T and the score s^*, neither of which is available to the estimator.

Denoising score matching (DSM). For a time interval $I \subset (0, \infty)$, define the integrated score error

$$\mathcal{L}_I(s) := \int_I \mathbb{E}\|s(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t.$$

For $\boldsymbol{x}_0 \in \mathbb{R}^D$, define the conditional DSM loss

$$\ell_I(s, \boldsymbol{x}_0) := \int_I \mathbb{E}_{\boldsymbol{X}_t \mid \boldsymbol{X}_0 = \boldsymbol{x}_0}\Big[\Big\| s(\boldsymbol{X}_t, t) + \frac{\boldsymbol{X}_t - m_t \boldsymbol{x}_0}{\sigma_t^2} \Big\|_2^2\Big] \mathrm{d}t.$$

The DSM identity gives $\mathcal{L}_I(s) = \mathbb{E}[\ell_I(s, \boldsymbol{X}_0) - \ell_I(s^*, \boldsymbol{X}_0)]$ (Vincent, 2011). For prescribed intervals $I_1, \ldots, I_K$, the learned score $\widehat{s}$ selects interval-wise approximate empirical minimizers $\widehat{s}_k$ of the DSM objective; Sec. 4 specifies the score classes, time-selection rule, and empirical optimization excesses.

Learned reverse process. To avoid the small-time singular regime induced by manifold-supported data, the reverse process stops at $t_0 > 0$. Conditional on the learned scores, sampling follows

$$\mathrm{d}\widehat{\boldsymbol{Y}}_t = \big(\widehat{\boldsymbol{Y}}_t + 2\widehat{s}(\widehat{\boldsymbol{Y}}_t, T - t)\big)\mathrm{d}t + \sqrt{2}\mathrm{d}\widetilde{\boldsymbol{B}}_t, \qquad \widehat{\boldsymbol{Y}}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D), \qquad 0 \le t \le T - t_0,$$

where $\widetilde{\boldsymbol{B}}$ is a standard D-dimensional Brownian motion independent of the Gaussian initial state, the training sample, and optimizer randomness. Let $\widehat{P}_{t_0}$ denote its conditional endpoint law and define

$$\widehat{P}_{t_0}^{\mathrm{box}} := (\mathrm{Proj}_{[0,1]^D})_{\#} \widehat{P}_{t_0}.$$

Thus, the superscript box denotes projection onto the ambient cube, not nearest-point projection onto $\mathcal{M}$. Only the interval-specific score networks are learned; all other components are prescribed.

2.2 Data Assumptions

The assumptions below play distinct roles: positive reach provides a tubular neighborhood with a unique nearest-point projection, Hölder regularity controls local approximation error, and the density lower bound stabilizes the denominators in the score representation.

Assumption 1 (Compact smooth manifold)**.** The law P_0 is supported on a compact C^∞ d-dimensional Riemannian manifold $\mathcal{M} \subset [0,1]^D$ without boundary, isometrically embedded in $\mathbb{R}^D$, with reach $\kappa > 0$ in the sense of Federer (1959) and Riemannian volume $S_{\mathcal{M}} = \mathrm{vol}_{\mathcal{M}}(\mathcal{M})$.

Definition 2.1 (Hölder functions on a manifold)**.** We use the chartwise analogue of the standard Euclidean Hölder class (Tsybakov, 2009). Fix a finite controlled $C^{\lceil\beta\rceil}$ atlas $\{(U_i, \phi_i)\}_{i\in\mathcal{A}}$. For $\boldsymbol{\alpha} \in \{0,1,2,\ldots\}^d$, write $|\boldsymbol{\alpha}| = \sum_{j=1}^d \alpha_j$. A function $f : \mathcal{M} \to \mathbb{R}$ belongs to $\mathcal{H}^\beta(\mathcal{M}, B_0)$ if every $f^{(i)} = f \circ \phi_i^{-1}$ has continuous derivatives through order m_β and

$$\max_{|\boldsymbol{\alpha}|\le m_\beta} \|\partial^{\boldsymbol{\alpha}} f^{(i)}\|_{L^\infty(\phi_i(U_i))} \vee \max_{|\boldsymbol{\alpha}|=m_\beta} \sup_{\boldsymbol{u}\neq\boldsymbol{v}} \frac{|\partial^{\boldsymbol{\alpha}} f^{(i)}(\boldsymbol{u}) - \partial^{\boldsymbol{\alpha}} f^{(i)}(\boldsymbol{v})|}{\|\boldsymbol{u}-\boldsymbol{v}\|_2^{\beta-m_\beta}} \le B_0,$$

where $\boldsymbol{u}, \boldsymbol{v} \in \phi_i(U_i)$. This is a chartwise radius for the fixed atlas, not a supremum over arbitrary reparametrizations.

Assumption 2 (Hölder density)**.** The law P_0 has density $p_0 \in \mathcal{H}^\beta(\mathcal{M}, B_0)$ with respect to $\mathrm{vol}_{\mathcal{M}}$, where $B_0 \ge 1$ and $\int_{\mathcal{M}} p_0 \mathrm{dvol}_{\mathcal{M}} = 1$.

Assumption 3 (Positive density lower bound)**.** There is $p_{\min} \in (0,1]$ such that $p_0(\boldsymbol{y}) \ge p_{\min}$ for every $\boldsymbol{y} \in \mathcal{M}$.

For an integer q, let $\Gamma_{\mathcal{M},q} \ge e$ denote the common envelope, defined formally in Definition B.6, for the finitely many derivatives of the chosen atlas, partition of unity, and projection coordinates used by the proof. It also includes the coordinate-change bounds transferring density regularity from its defining atlas to the projection atlas. This finite-order quantity records the additional geometry dependence; reach alone is not claimed to control higher derivatives. Although Assumption 1 assumes a C^∞ manifold for convenience, the quantitative bounds use only the finite collection of derivatives recorded by $\Gamma_{\mathcal{M},q}$ at the stated order.

2.3 Main Guarantee

The cutoffs t_0, t_1, and T control early stopping, the transition between the two score representations, and terminal initialization, respectively. Define the target sample-size rate

$$\varepsilon_n := n^{-(\beta+1)/(d+2\beta)}. \tag{2}$$

For the main estimator, fix the cutoff schedule

$$t_0 \asymp n^{-2(\beta+1)/(d+2\beta)}, \qquad t_1 \asymp n^{-2/(d+2\beta)}, \qquad T = c_T \log n, \tag{3}$$

where $c_T \ge 2$ is fixed and chosen sufficiently large as specified in Corollary H.2. Specifically, take $t_1 = c_1 n^{-2/(d+2\beta)}$ and $t_0 = c_0 n^{-2(\beta+1)/(d+2\beta)}$, with fixed $0 < c_0 < c_1 \le 2^{-1-2/\beta} \wedge (1/4)$. These constants ensure the small-noise accuracy condition verified in the appendix.

Theorem 2.2 (Informal W_1 rate; formal version in Corollary H.2)**.** *Suppose Assumptions 1 to 3 hold with $d > 2$. Take $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$ as in (5), and suppose $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$. Assume the explicit reach-tube and denominator-resolution inequalities (236) and (237), and choose c_T larger than $a_0 + (\beta+1)/(d+2\beta)$ by a fixed margin. Use the cutoff schedule (3) and the interval-specific estimator of Sec. 4. If $\max_{1\le k\le K} \sqrt{\mathbb{E}\mathcal{E}_{\mathrm{opt},k}} \lesssim_{\log} \varepsilon_n$, where K is the number of time intervals and $\mathcal{E}_{\mathrm{opt},k}$ is defined in (7), then*

$$\mathbb{E}\,\mathsf{W}_1\big(P_0, \widehat{P}_{t_0}^{\mathrm{box}}\big) \lesssim_{\log} D^{\mathcal{O}_\beta(d)} \varepsilon_n,$$

when the finite-order geometry envelope, the Hölder radius B_0, and $p_{\min}^{-1}$ are fixed. The expectation is over the training sample and any optimizer randomness.

For fixed D and fixed geometry and density quantities, the sample-size exponent in Theorem 2.2 matches the known smooth-density W_1 minimax benchmark for lower-bounded densities on d-dimensional Euclidean domains (Niles-Weed & Berthet, 2022).

Finite-step discretization counterpart. Sec. H.3 and Corollary H.14 gives an aligned-grid exponential-integrator sampler with the same W_1 rate under the corresponding grid-weighted optimization condition and bounds its number of score evaluations; the conditional Gaussian-corruption expectations remain exact.

3 WHY SCORE APPROXIMATION REQUIRES TWO NOISE REGIMES

This section constructs deterministic *comparison scores* for the two noise regimes and states their approximation guarantees. A comparison score is an element of the corresponding score class used to bound the approximation term in the oracle inequality. Its construction may depend on the manifold and density, but the estimator is not supplied with the comparison function or manifold coordinates. The estimator remains an approximate empirical minimizer over the ambient score classes analyzed in Sec. 4. Each comparison score is obtained from a ReLU network with input $(\boldsymbol{x}, t) \in \mathbb{R}^D \times I$ using the fixed input truncation, affine rescaling, and coordinatewise output clipping defined in Sec. 4, where I is the corresponding time interval.

Figure 1 summarizes why the representation changes at t_1 and how the two comparison constructions enter the common statistical argument.

We next introduce the finite-order and localization parameters used by both constructions. For these constructions, choose a finite geometry order $q_{\mathrm{geom}} \in \{1, 2, \ldots\}$ satisfying

$$q_{\mathrm{geom}} \geq q_{\min}(\beta) \coloneqq \lceil \max\{\beta + 4, 6\beta + 4\} \rceil, \qquad q_\star \coloneqq q_{\mathrm{geom}} - 2. \tag{4}$$

For the simplified $d > 2$ rate, take

$$q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta) \coloneqq \lceil \max\{q_{\min}(\beta), 2 + 3(\beta + 1)\} \rceil. \tag{5}$$

This order is independent of (n, D, d) and ensures $q_\star \geq d(\beta + 1)/(d - 2)$ for every integer $d > 2$. These quantities are proof parameters for the comparison networks; they are not additional smoothness assumptions. The geometric subnetworks use smoothness q_{geom}, while density-dependent coefficients retain their actual smoothness $\beta - q$. The higher geometric order controls the cost of approximating normalized projection residuals uniformly down to t_0; the lower-time factor is displayed in (225). For an ambient point $\boldsymbol{x}$, define its distance to the scaled manifold and the active tube by

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

$$\varrho_t \coloneqq C_\varrho \sigma_t \sqrt{D \vee \log n}, \qquad A_t \coloneqq \{\boldsymbol{x} \in \mathbb{R}^D : \mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \leq \varrho_t\}. \tag{6}$$

For a sufficiently large fixed C_ϱ, Gaussian concentration for Lipschitz functions (Vershynin, 2018) makes the clipped score risk outside A_t smaller than any prescribed inverse power of n; Lemma C.3 gives the formal statement. The approximation problem is therefore concentrated on A_t. Under the small-noise reach condition, the nearest-point projection $\Pi_{\mathcal{M}}$ is single-valued at $\boldsymbol{x}/m_t$ for $\boldsymbol{x} \in A_t$ and $t \in [t_0, t_1]$. On this tube, a finite projection atlas and partition of unity express the smoothed manifold density through a Gaussian-weighted denominator $\mathcal{Q}(\boldsymbol{x}, t)$ and first moment $\mathcal{P}(\boldsymbol{x}, t)$, whose integral definitions are given in Sec. D.3.3. Defining

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

gives the centered representation

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

Approximating this centered, scaled ratio avoids separately approximating two large terms whose difference is divided by σ_t. Recall ε_n from (2) and $q_\star$ from (4), and define the cell-radius cap

$$r_{\star,n} \coloneqq c_{\mathrm{cap}} \Gamma_{\mathcal{M}, q_{\mathrm{geom}}}^{-C_\star} B_0^{-C_\star} p_{\min}^{C_\star} (D \vee \log n)^{-2} \varepsilon_n^{1/q_\star},$$

where $C_\star$ is sufficiently large and c_{cap} sufficiently small, with both constants independent of (n, D).

3.1 LOW NOISE: FACTOR OUT THE GAUSSIAN NORMAL-DISPLACEMENT TERM

Extending a tangent-cell cover down to t_0 would require on the order of $\sigma_{t_0}^{-d}$ cells, which exceeds the capacity compatible with the target statistical rate. Instead, for $\boldsymbol{x} \in A_t$ and $t \in [t_0, t_1]$, set

$$\boldsymbol{\pi}(\boldsymbol{x}, t) \coloneqq \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t), \qquad \boldsymbol{r} \coloneqq \boldsymbol{x} - m_t \boldsymbol{\pi}(\boldsymbol{x}, t), \qquad \boldsymbol{\nu} \coloneqq \boldsymbol{r}/\sigma_t.$$

In a projection chart, write $\boldsymbol{u}_i^\Pi = \phi_i(\boldsymbol{\pi}(\boldsymbol{x}, t))$ and rescale the tangential coordinate as $\boldsymbol{u} = \boldsymbol{u}_i^\Pi + \sigma_t \boldsymbol{w}$. Orthogonality of the normal residual eliminates the first-order cross term, leaving a d-dimensional Gaussian leading phase plus finite-order curvature and density corrections.

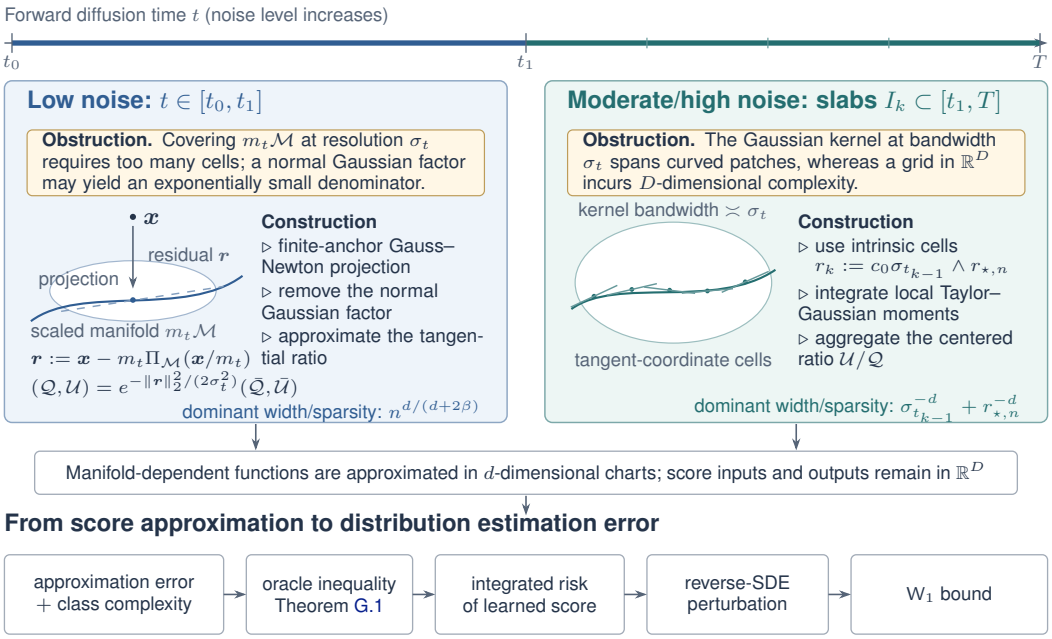

Figure 1: Noise-adapted score approximations and their role in the proof. At low noise, projection centering avoids a noise-scale tangent cover, and factoring out the common normal Gaussian term stabilizes the denominator; at moderate and high noise, tangent-cell moment expansions exploit intrinsic geometry.

The denominator itself may still be exponentially small because of the normal residual. We cancel the common factor before constructing a reciprocal:

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

After factoring out this Gaussian term, $\bar{\mathcal{Q}} \geq C^{-1}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C} p_{\min}(D \vee \log n)^{-d/4}$ on the active tube. The reciprocal network therefore operates on a polynomial, rather than exponentially small, denominator range.

We next implement $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ without approximating it as a generic nonlinear function on $\mathbb{R}^D$. The construction fixes finitely many anchors in each d-dimensional chart box. Let $\boldsymbol{z}_i = \phi_i^{-1}$ denote the chart embedding, let $\boldsymbol{J}_i(\boldsymbol{u}) = \nabla \boldsymbol{z}_i(\boldsymbol{u})$ be its Jacobian, and let $\Pi_i^{(\mathrm{box})}$ denote Euclidean projection onto the fixed chart box. The construction unrolls approximate Gauss–Newton iterations

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

and selects among the candidates by their ambient residuals. Only the manifold-dependent functions are approximated on the intrinsic variable $\boldsymbol{u} \in \mathbb{R}^d$; the remaining ambient dependence enters through explicit arithmetic, residual comparisons, and linear-algebra modules. Theorem F.13 and Corollary F.24 formalize this module.

The following informal proposition combines these components using the canonical finite-order choice in (5); its formal counterparts retain the finite ambient exponent.

Proposition 3.1 (Low-noise comparison score; informal version of Corollary F.33 and Proposition F.34)**.** *Under Assumptions 1 to 3, $D \vee p_{\min}^{-1} \leq n^{a_0}$ for a fixed $a_0 > 0$, and the small-noise conditions (236) and (237), there is a comparison score s_1 such that, for every $t \in [t_0, t_1]$,*

$$\mathbb{E}\|s_1(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \lesssim_{\log} \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} p_{\min}^{-C} D^{\mathcal{O}_\beta(d)} n^{-2\beta/(d+2\beta)} \sigma_t^{-2}.$$

The underlying ReLU network has leading width and sparsity $n^{d/(d+2\beta)}$ times explicit polynomial ambient-dimension, geometry, and density factors. Its output is clipped at order $\sigma_t^{-1}(D \vee \log n)^{1/2}$.

3.2 Moderate and High Noise: Approximation in Local Tangent Coordinates

On a dyadic time interval $I_k = [t_{k-1}, t_k) \subset [t_1, T]$, Gaussian-weighted local moments are approximated by partitioning each chart into cells of radius

$$r_k = c_0 \sigma_{t_{k-1}} \wedge r_{\star,n},$$

where $c_0 > 0$ is a sufficiently small fixed numerical constant. For a cell center $\boldsymbol{u}_\nu$ with $\boldsymbol{y}_\nu = \boldsymbol{z}_i(\boldsymbol{u}_\nu)$, let $\mathsf{P}_\nu^{\tan}$ denote the orthogonal projector onto $T_{\boldsymbol{y}_\nu}\mathcal{M}$. The affine tangent-plane center

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

separates the leading tangential Gaussian from finite-order curvature corrections. The network approximates the resulting cellwise Gaussian moments and forms their centered ratio using on the order of $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$ cells rather than a grid in $\mathbb{R}^D$. This choice of $r_{\star,n}$ makes the fixed-order phase expansion uniform over the time intervals; Proposition G.3 uses the same cap in the learned-score bound.

Proposition 3.2 (Moderate- and high-noise comparison scores; informal version of Corollary G.5)**.** *Under Assumptions 1 to 3 and $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$, every $I_k \subset [t_1, T]$ admits a comparison score s_k satisfying, for all $t \in I_k$,*

$$\mathbb{E}\|s_k(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \lesssim_{\log} \Gamma_{\mathcal{M}, q_{\mathrm{geom}}}^{C} p_{\min}^{-C} D^{\mathcal{O}_\beta(d)} \sigma_t^{-2} \varepsilon_n^2.$$

The underlying raw network's width and sparsity are bounded by $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$ times explicit polynomial factors in the ambient dimension, geometry, and density parameters.

Together, the two propositions provide deterministic comparison scores over $[t_0, T]$; Sec. 4 converts their approximation and complexity bounds into guarantees for the learned scores.

4 Statistical Guarantees for Learned Score Networks

The constructions in Sec. 3 provide deterministic benchmark elements in the score classes; they are not learned geometric modules. An oracle inequality compares each approximate empirical minimizer with the corresponding constructive score, while a covering-number bound for sparse ReLU networks controls the estimation penalty. Recall ε_n from (2) and $q_\star$ from (4); the moderate- and high-noise accuracy radius $r_{\star,n}$ is defined formally in Proposition G.3. For the continuous-time estimator, let K be the number of disjoint time intervals and use $I_1 = [t_0, t_1)$, $I_k = [t_{k-1}, t_k)$ for $2 \le k < K$, and $I_K = [t_{K-1}, T]$; the exact dyadic grid is specified in (254). Changing a single endpoint does not affect an integrated score risk. The result summaries below use the canonical finite-order choice (5).

4.1 ReLU Score Classes with Fixed Pre- and Postprocessing

We first specify the trainable ReLU class and distinguish its parameters from the fixed input preprocessing and output clipping.

For depth L, width vector $\boldsymbol{W}$, sparsity bound S, and parameter bound B, let $\mathcal{NN}(L, \boldsymbol{W}, S, B)$ denote the corresponding sparse ReLU class; the layerwise definition is given in Definition B.1. For score networks, the input and output dimensions are $D + 1$ and D, respectively.

Fix a time interval $I = [\underline{t}, \overline{t}]$. For a raw sparse-ReLU class $\mathcal{NN}(L, \boldsymbol{W}, S, B)$, the notation $\widetilde{\mathcal{NN}}(L, \boldsymbol{W}, S, B)$ denotes the corresponding postprocessed score class. Its input is first truncated coordinatewise to $[-B_x, B_x]^D \times [\underline{t}, \overline{t}]$, where $B_x = (D \vee \log n)^C \sigma_{\underline{t}}^{-1}$, and then affinely rescaled to $[0,1]^{D+1}$. The raw network is evaluated on this compact domain, after which each output coordinate is clipped at $C\sigma_t^{-1}(D \vee \log n)^{1/2}$. The input truncation, affine rescaling, and output clipping are fixed deterministic operations and are not counted as trainable ReLU parameters. Comparison networks constructed in physical coordinates are composed with the inverse affine rescaling before being placed in the raw class; (260) records the resulting parameter bounds. The off-tube estimates in Sec. 3 make the discarded input tails negligible.

Recall the time-integrated squared score error $\mathcal{L}_I(s)$ and DSM loss $\ell_I(s, \boldsymbol{x})$ from Sec. 2. Given a postprocessed class $\mathcal{F}_k$ on I_k, let $\widehat{s}_k$ be a measurable approximate empirical minimizer and define

$$\mathcal{E}_{\mathrm{opt},k} \coloneqq \mathbb{P}_n \ell_{I_k}(\widehat{s}_k) - \inf_{s \in \mathcal{F}_k} \mathbb{P}_n \ell_{I_k}(s) \ge 0, \tag{7}$$

where $\mathbb{P}_n$ is the empirical average over the training sample. All expectations involving a learned score network are over the data and any optimizer randomness. Exact ERM is the special case $\mathcal{E}_{\mathrm{opt},k} = 0$.

The objective keeps the expectations over diffusion time and Gaussian corruption exactly as written. Monte Carlo error from approximating these expectations is outside the main theorem and should not be conflated with the empirical optimization excess above.

4.2 Oracle inequality and Covering-Number Bounds

The oracle inequality separates approximation, statistical estimation, tail, and optimization errors. A covering-number bound controls the statistical estimation term. For a function class $\mathcal{F}$ and $\eta > 0$, let $\mathcal{N}(\eta, \mathcal{F}, \|\cdot\|_\infty)$ denote the minimum cardinality of an η-cover of $\mathcal{F}$ in the supremum norm. For any fixed $A > 0$ and a sufficiently large fixed A_{cov} as in Theorem G.1, the theorem gives, on each interval,

$$\mathbb{E}\mathcal{L}_{I_k}(\widehat{s}_k) \lesssim \inf_{s\in\mathcal{F}_k} \mathcal{L}_{I_k}(s) + \frac{(D\vee\log n)^2 \int_{I_k}\sigma_t^{-2}\mathrm{d}t}{n} \log\big(\mathcal{N}(n^{-A_{\mathrm{cov}}}, \mathcal{F}_k, \|\cdot\|_\infty)\vee n\big) + n^{-A} + \mathbb{E}\mathcal{E}_{\mathrm{opt},k}.$$

Because the fixed input preprocessing and coordinatewise output clipping do not increase the supremum-norm covering number, Lemma I.15 gives the direct architecture-dependent bound

$$\log\big(\mathcal{N}\big(n^{-A_{\mathrm{cov}}}, \widetilde{\mathcal{NN}}(L, \boldsymbol{W}, S, B), \|\cdot\|_\infty\big)\vee n\big) \lesssim \log n + SL\log\big(nL(\|\boldsymbol{W}\|_\infty + 1)(B\vee 1)\big). \quad (8)$$

Consequently, on a time slab I, the estimation term in the oracle inequality is bounded by

$$\frac{(D\vee\log n)^2}{n}\int_I \sigma_t^{-2}\mathrm{d}t\Big[\log n + SL\log\Big(nL(\|\boldsymbol{W}\|_\infty + 1)(B\vee 1)\Big)\Big].$$

For the time slabs used below, the integral factor is logarithmic or bounded; the subsequent rate statements substitute the corresponding values of $(L, \boldsymbol{W}, S, B)$ from the approximation constructions.

4.3 Learned Scores in the Two Noise Regimes

Substituting the two comparison networks and their architecture bounds into the oracle inequality yields the learned-score risks below.

In both summaries below, the arbitrary fixed constant A in the negligible remainder n^{-A} is chosen sufficiently large, so this remainder is absorbed by the displayed statistical term.

On the low-noise interval I_1, let $\mathcal{F}_1$ be the projection–Laplace score class and let $\widehat{s}_1$ be a measurable approximate empirical minimizer over $\mathcal{F}_1$, with empirical excess $\mathcal{E}_{\mathrm{opt},1}$. The class has leading width and sparsity scale $n^{d/(d+2\beta)}$. The informal version of Corollary G.7 is

$$\mathbb{E}\mathcal{L}_{[t_0,t_1]}(\widehat{s}_1) \lesssim_{\log} \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} p_{\min}^{-C} n^{-2\beta/(d+2\beta)} D^{\mathcal{O}_\beta(d)} + \mathbb{E}\mathcal{E}_{\mathrm{opt},1}.$$

For each moderate- or high-noise interval I_k, $k \geq 2$, let $\mathcal{F}_k$ be the tangent-cell score class and let $\widehat{s}_k$ be a measurable approximate empirical minimizer over $\mathcal{F}_k$, with empirical excess $\mathcal{E}_{\mathrm{opt},k}$. Proposition G.3 gives

$$\mathbb{E}\mathcal{L}_{I_k}(\widehat{s}_k) \lesssim_{\log} \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} p_{\min}^{-C}\Big[n^{-2(\beta+1)/(d+2\beta)} + \frac{\sigma_{t_{k-1}}^{-d} + \varepsilon_n^{-d/q_\star}}{n}\Big] D^{\mathcal{O}_\beta(d)} + \mathbb{E}\mathcal{E}_{\mathrm{opt},k}.$$

For the moderate- and high-noise classes, use cell radius $r_k = c_0\sigma_{t_{k-1}} \wedge r_{\star,n}$. A deterministic half-open time-selection rule defines the *piecewise-in-time score estimator* $\widehat{s}$ by choosing the corresponding interval-specific network. Its boundaries add no learned parameters, and the resulting collection is not claimed to be a single finite ReLU network. The aggregate width and sparsity are obtained by summing their interval-specific values, retaining both $\sigma_{t_{k-1}}^{-d}$ and $r_{\star,n}^{-d}$ explicitly. These learned-score bounds are the inputs to the distributional argument in Sec. 5.

5 From Score Estimation Error to 1-Wasserstein Error

The main theorem follows by perturbing the reverse SDE separately on each time interval. This section exposes that last conversion and, in particular, why the apparently slower small-noise score rate still yields the target distribution rate. The formal hybrid-process argument is in Sec. H. Because P_0 is supported on $[0,1]^D$ and $\mathrm{Proj}_{[0,1]^D}$ is nonexpansive, the final cube projection does not increase the 1-Wasserstein error.

Use the cutoff schedule (3) and the dyadic slab partition for moderate and high noise from (254). Then $\sqrt{t_0} \asymp \varepsilon_n$ and $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$. The following informal restatement of Corollary H.1 exposes the error decomposition underlying Theorem 2.2; the appendix corollary is the formal statement. We use the canonical finite-order choice (5) and summarize its explicit ambient exponent by $\mathcal{O}_\beta(d)$.

Proposition 5.1 (Four-term distribution-error decomposition; informal version of Corollary H.1)**.** *Assume Assumptions 1 to 3, $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$, and the two small-noise conditions (236) and (237). Take $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$ and set $q_\star = q_{\mathrm{geom}} - 2$, and use the piecewise-in-time score estimator from Sec. 4, with empirical optimization excesses defined in (7). There are positive constants c_Γ, c_p, independent of (n, D), such that*

$$\mathbb{E}\,\mathsf{W}_1\big(P_0, \widehat{P}_{t_0}^{\mathrm{box}}\big) \lesssim_{\log} \sqrt{Dt_0} + De^{-T} + \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{c_\Gamma} p_{\min}^{-c_p}\big(\varepsilon_n + n^{-1/2} + n^{-1/2}\varepsilon_n^{-d/(2q_\star)}\big) D^{\mathcal{O}_\beta(d)} + D\sum_{k=1}^{K} \sqrt{t_k\,\mathbb{E}\mathcal{E}_{\mathrm{opt},k}}.$$

The expectation is over the training sample and any optimizer randomness.

The four terms arise, respectively, from early stopping, terminal initialization, score approximation and estimation, and empirical optimization. We explain how each contribution is controlled below.

Why the low-noise regime has the target rate. Its integrated score risk is of order $n^{-2\beta/(d+2\beta)}$ up to the displayed factors. The reverse-SDE perturbation over $[t_0, t_1]$ contributes the square root of the interval length, so

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

Why moderate- and high-noise regimes have the target rate. The estimation term on time interval k contains $(\sigma_{t_{k-1}}^{-d} + \varepsilon_n^{-d/q_\star})/n$. For $d > 2$, dyadic summation over the large-noise intervals gives

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

For $d = 2$, the corresponding sum is of order $n^{-1/2}(1+\log(1/t_1))$, as accounted for in the logarithmic factors of Corollary H.1. The time intervals with $t_k \gtrsim 1$ add only logarithmic factors, while the radius cap $r_{\star,n}$ produces $n^{-1/2}\varepsilon_n^{-d/(2q_\star)}$. When $d > 2$ and $q_\star \ge d(\beta+1)/(d-2)$, both correction terms are at most a constant multiple of ε_n. In this reduction, $d > 2$ ensures $n^{-1/2} \lesssim \varepsilon_n$ and permits a finite $q_\star$ satisfying the displayed condition, so the radius-cap contribution is also absorbed into ε_n.

Finally, the choice of t_0 in (3) controls early stopping, $c_T > a_0 + (\beta+1)/(d+2\beta)$ controls terminal initialization, and the optimization tolerance in Theorem 2.2 controls the last line of Proposition 5.1. These substitutions give exactly the simplified theorem stated in Sec. 2.

6 Conclusion

We prove that approximate empirical risk minimization over ambient ReLU score-network classes can attain an intrinsic-dimensional sample-size exponent, without a preliminary manifold-estimation step. The expected 1-Wasserstein bound has polynomial ambient-dimension dependence and follows from tangent-cell approximations at moderate and high noise and a projection-centered construction at low noise. The complexity bounds apply to finitely many interval-specific networks.

The idealized pipeline assumes exact manifold support, positive density, controlled finite-order geometry, the stated small-noise sample-size conditions, accurate interval-wise optimization, and exact diffusion-time and Gaussian-corruption expectations. The main theorem is continuous-time; the finite-step result retains the rate and bounds score evaluations under grid-weighted optimization, but efficient training, near-manifold data, shared-parameter models, and sharper ambient dependence remain open.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, drafting and editing prose, consistency checks, and critical review of mathematical arguments and LaTeX artifacts. No AI-generated data or experimental results are reported. The authors take responsibility for the final content of the work, including text, claims, citations, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The learning problem, assumptions, estimator, and main guarantee are specified in Sec. 2 and 4. The appendices provide the complete proofs, construction details, network-complexity bounds, sample-size conditions, and continuous- and discrete-time error decompositions. The paper reports theoretical results rather than experiments; reproducing its conclusions requires the exact-expectation, optimization, geometry, density, and sample-size conditions stated in the corresponding results.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. doi: 10.1016/0304-4149(82)90051-5.

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

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. 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.

Jean-Daniel Boissonnat, André Lieutier, and Mathijs Wintraecken. The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *Journal of Applied and Computational Topology*, 3(1–2):29–58, 2019. doi: 10.1007/s41468-019-00029-8.

Miha Brešar and Aleksandar Mijatović. Nonasymptotic bounds for forward processes in denoising diffusions: Ornstein–uhlenbeck is hard to beat. *The Annals of Applied Probability*, 35(6):4439–4463, 2025. doi: 10.1214/25-AAP2232.

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

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

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. 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, 2023b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. 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, 2023c.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. KL convergence guarantees for score diffusion models under minimal data assumptions. *arXiv preprint arXiv:2308.12240*, 2023.

Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Provable maximum entropy manifold exploration via diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12739–12761. PMLR, 2025.

Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H. Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

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

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

Guoji Fu and Wee Sun Lee. Approximation and generalization abilities of score-based neural network generative models for sub-gaussian distributions. In *Advances in Neural Information Processing Systems*, volume 38, pp. 77963–78061, 2025.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. 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. doi: 10.1109/ISIT63088.2025.11195641.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*, 2026. doi: 10.1287/moor.2024.0769.

Yuchen Jiao, Yuchen Zhou, and Gen Li. Optimal convergence analysis of DDPM for general distributions. *arXiv preprint arXiv:2510.27562*, 2025.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22870–22882, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. 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.

Gen Li and Yuling Yan. 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.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. 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.

Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57499–57538, 2024b.

Xiang Li, Zebang Shen, Ya-Ping Hsieh, and Niao He. When scores learn geometry: Rate separations under the manifold hypothesis. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*. OpenReview.net, 2026.

Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation (extended abstract). In Nika Haghtalab and Ankur Moitra (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.

Zichen Liu, Wei Zhang, and Tiejun Li. Improving the euclidean diffusion generation of manifold data by mitigating score function singularity. In *Advances in Neural Information Processing Systems*, volume 38, pp. 110885–110917, 2025.

Yubin Lu, Zhongjian Wang, and Guillaume Bal. Mathematical analysis of singularities in the diffusion model under the submanifold assumption. *arXiv preprint arXiv:2301.07882*, 2023.

Yang Lyu, Yuchun Qian, Tan Minh Nguyen, and Xin T Tong. Resolving memorization in empirical diffusion model for manifold data in high-dimensional spaces. *arXiv preprint arXiv:2505.02508*, 2025.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022. doi: 10.1214/21-AOS2161.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, second edition, 2006. doi: 10.1007/978-0-387-400 65-5.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. 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.

Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35852–35865, 2022.

Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In Nika Haghtalab and Ankur Moitra (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.

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

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 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.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11895–11907, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440. PMLR, 2024.

Stanislas Strasman, Antonio Ocello, Claire Boyer, Sylvain Le Corff, and Vincent Lemaire. An analysis of the noise schedule for score-based generative models. *Transactions on Machine Learning Research*, 2025.

Taiji Suzuki. 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.

Jiaqi Tang and Yuling Yan. Adaptivity and convergence of probability flow odes in diffusion generative models. *arXiv preprint arXiv:2501.18863*, 2025.

Rong Tang and Yun Yang. 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.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/B13794. URL https://doi.org/10.1007/b13794.

Loring W Tu. *An Introduction to Manifolds*. Universitext. Springer, second edition, 2011. ISBN 978-1-4419-7399-3. doi: 10.1007/978-1-4419-7400-6.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302, 2021.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Breaking the curse of dimensionality: Diffusion models efficiently learn low-dimensional distributions. *Journal of Machine Learning Research*, 27(179):1–50, 2026. URL https://jmlr.org/papers/v27/25-1522.html.

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

R. Wong. *Asymptotic Approximations of Integrals*, volume 34 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2001. doi: 10.1137/1.9780898719260.

Konstantin Yakovlev and Nikita Puchkin. 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.

Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact langevin algorithm in KL divergence with application to score-based generative models. *arXiv preprint arXiv:2211.01512*, 2022.

Matthew S. Zhang, Stephen Huan, Jerry Huang, Nicholas M. Boffi, Sitan Chen, and Sinho Chewi. Sublinear iterations can suffice even for ddpms. *arXiv preprint arXiv:2511.04844*, 2025.

Zixuan Zhang, Kaixuan Huang, Tuo Zhao, Mengdi Wang, and Minshuo Chen. Diffusion model for manifold data: Score decomposition, curvature, and statistical complexity. *arXiv preprint arXiv:2603.20645*, 2026.

APPENDIX CONTENTS

A Further Discussion of Related Work **16**

A.1 Comparison with Zhang et al. . . . 16

A.2 Comparison with the Dimension-Free Support-Aware Result . . . 16

A.3 Additional Related Work . . . 16

B Notation, Geometry, and Proof Conventions **18**

B.1 Global notation and network classes . . . 18

B.2 Standard manifold conventions . . . 18

B.3 Parameter-dependence conventions . . . 18

B.4 Finite-order geometry and tube-volume conventions . . . 19

B.5 Examples satisfying the polynomial-geometry regime . . . 20

C Off-Tube Score Localization **23**

C.1 Off-manifold Gaussian tail and Fisher bounds . . . 23

D Setup for Noise-Regime-Specific Score Approximation **25**

D.1 Small- and large-noise regimes . . . 25

D.2 Noise-regime-specific approximation roadmap . . . 25

D.3 Projection atlas and chart-integral decomposition . . . 26

D.4 Centered ratio target and stability . . . 31

E Large-Noise Tangent-Cell Score Approximation **33**

E.1 Tangent-cell phase decomposition and gated chart approximation . . . 33

E.2 Scale choices for tangent-cell large-noise approximation . . . 34

F Score Approximation in the Small-Noise Regime **40**

F.1 Proof architecture and active-chart geometry . . . 40

F.2 Finite-anchor projection networks . . . 46

F.3 Projection-centered Laplace expansion . . . 59

F.4 Network assembly and final small-noise theorem . . . 66

G Score Estimation on Compact Smooth Manifolds **92**

G.1 Oracle Inequality for Score Estimation . . . 92

G.2 Time-Grid and Integral Conventions . . . 95

G.3 Score Estimation on Large-Noise Regime . . . 95

G.4 Score Estimation on Small-Noise Regime . . . 98

H W_1 Distribution Estimation **100**

H.1 Conditions for the distributional rate . . . 100

H.2 W_1 distribution estimation . . . 100

H.3 Aligned time discretization and direct path-space control . . . 105

I Technical Toolkit **110**

I.1 Neural network approximation toolkit 110

I.2 Support lemmas for generalization analysis 114

I.3 Auxiliary lemmas 116

A Further Discussion of Related Work

A.1 Comparison with Zhang et al.

Zhang et al. (2026) also study score estimation and distribution learning for smooth manifold-supported data. Their small-noise analysis uses the nearest-point projection $\Pi_{\mathcal{M}}$, and Lemma D.18 approximates the projection term by ReLU networks. The displayed network rate is intrinsic-dimensional in the approximation accuracy, but the multiplicative constant is allowed to depend on D, d, β, the density bound, and smoothness bounds for the projection maps. Consequently, that lemma does not quantify the ambient-dimension dependence of the projection-network constant.

Our construction addresses a different quantitative interface. It implements the nearest-point projection through finite-anchor Gauss–Newton iterations in d-dimensional chart coordinates, while all residual comparisons and other ambient operations are explicit. After removing the common normal Gaussian factor, the reciprocal network is evaluated only on a denominator with a displayed polynomial lower bound. Under the polynomial growth conditions imposed on the finite-order geometry and density factors, this yields an explicit finite-polynomial dependence on D. This is a constructive upper bound for our architecture, not a lower bound on alternative implementations of Zhang et al.'s result. The distinction concerns the displayed ambient prefactor, not the intrinsic sample-size exponent.

A.2 Comparison with the Dimension-Free Support-Aware Result

This comparison uses arXiv v3, revised 6 August 2026, of Azangulov et al. (2024). For $d \geq 3$, their Corollary 3.2 gives the sample-size exponent $(\alpha+1)/(2\alpha+d)$ in expected W_1, up to logarithmic factors, without an explicit ambient-dimension factor. Their Theorem 3.3 gives a dimension-free score-risk bound for the structured class $\mathcal{S}_k$ and a bound linear in D for the ambient class $\mathcal{S}'_k$. The structured estimator first constructs a data-dependent low-dimensional representation from the observations and then performs local empirical risk minimization. Their distributional result also uses the discretization and consistency construction developed in Section 6.

Our estimator has a different interface. It performs interval-wise approximate empirical risk minimization over prescribed ambient score-network classes without first supplying the optimizer with estimated local subspaces, bases, or coordinates. The geometric modules are deterministic comparison functions used to prove an oracle bound; no claim is made that the optimizer recovers those modules. Our result retains the proved finite-polynomial ambient, geometry, and density factors and does not improve the exponent or the ambient prefactor of Azangulov et al.

For sampling, our aligned-grid exponential integrator admits a direct path-space bound without a learned-score Lipschitz assumption; Corollaries H.13 and H.14 give the resulting distributional bound and polynomial score-evaluation bound under their stated conditions.

A.3 Additional Related Work

Ambient-space convergence. Early analyses imposed structural conditions such as log-Sobolev inequalities or log-concavity to obtain polynomial convergence rates (Lee et al., 2022; Yang & Wibisono, 2022; Bruno et al., 2023). Chen et al. (2023c); Lee et al. (2023) showed that L^2 score accuracy is sufficient for polynomial convergence under bounded-support or tail conditions without such functional inequalities. Conforti et al. (2023) established convergence under finite Fisher information without early stopping. Using stochastic localization, Benton et al. (2024) obtained dependence linear in the ambient dimension up to logarithmic factors, while Li & Yan (2025); Li et al. (2023) obtained faster nonasymptotic rates for ODE- and SDE-based samplers. Jiao et al. (2025) refined DDPM rates under weaker smoothness conditions. These full-dimensional analyses do not by themselves exploit the intrinsic geometry of singular manifold-supported laws.

Low-dimensional sampling guarantees. Huang et al. (2026); Potaptchik et al. (2025) proved iteration-complexity bounds governed by intrinsic rather than ambient dimension for stochastic samplers. For deterministic samplers, Tang & Yan (2025); Liang et al. (2025) established dimension-free total-variation convergence rates for probability-flow ODEs and DDIMs. George et al. (2025) identified dimension-dependent phase transitions in the dynamics. These works principally analyze

sampling given a suitable score approximation, whereas our result also controls statistical score estimation from finite data.

Intrinsic-dimension statistical guarantees. Chakraborty et al. (2026) analyze score-matching diffusion models for general low-dimensional measures through a (p, q)-Wasserstein dimension, building on empirical-measure rates such as those of Weed & Bach (2019). Their framework applies beyond smooth compact manifolds under moment assumptions and yields bounds governed by the relevant Wasserstein dimension. For lower-bounded β-smooth densities on d-dimensional Euclidean domains, the minimax W_1 exponent is $(\beta + 1)/(d + 2\beta)$ for $d > 2$ (Niles-Weed & Berthet, 2022); our theorem proves a manifold upper bound with this exponent but no separate manifold lower bound. Lyu et al. (2025) use an inertial correction with the empirical score and obtain an ambient-dimension-free W_1 rate $\mathcal{O}(n^{-2/(d+4)})$ for C^2 manifolds. Our estimator instead targets the population score through neural-network empirical risk minimization for Hölder densities.

Diffusion dynamics and schedules. Strasman et al. (2025) derived error bounds that display the effect of the noise schedule. Li et al. (2024a); Zhang et al. (2025) studied higher-order and randomized-midpoint discretizations. Brešar & Mijatović (2025) compared forward diffusions through their convergence to the invariant measure and identified advantages of the canonical Ornstein–Uhlenbeck process.

Geometric interpretations. Pidstrigach (2022) gave conditions under which score-based generative models detect the data manifold. Stanczuk et al. (2024) related diffusion models to the normal bundle and intrinsic-dimension estimation. Wang et al. (2026) interpreted diffusion training through subspace clustering and mixtures of low-rank Gaussians. Lu et al. (2023); Liu et al. (2025) analyzed the singular normal and tangential components of Euclidean diffusion scores near embedded manifolds. Li et al. (2026) showed that geometric information and density information enter at different small-noise orders. Our distribution-recovery theorem must control both the leading normal projection field and the lower-order density-dependent terms. De Santi et al. (2025) studied maximum-entropy exploration over a diffusion-defined data manifold, and Li et al. (2024b) related diffusion denoisers to optimal linear denoisers during generalization.

B Notation, Geometry, and Proof Conventions

The technical development proceeds as follows. Sec. C establishes off-tube score localization. Sec. D introduces the common projection atlas, chart-integral representation, and noise-regime setup. Sec. E and F prove the large- and small-noise approximation results, respectively. Sec. G converts these deterministic approximants into learned-score bounds, and Sec. H transfers the resulting score error to Wasserstein error. The supporting approximation, concentration, and covering results are collected in Sec. I.

B.1 Global notation and network classes

Let $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$, $\mathbb{N} := \{0, 1, 2, \ldots\}$, and $\mathbb{N}_+ := \mathbb{N} \setminus \{0\}$. For real numbers a, b, write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \mathbb{N}^k$ and $\boldsymbol{v} = (v_1, \ldots, v_k) \in \mathbb{R}^k$, set $|\boldsymbol{\alpha}| := \sum_{j=1}^k \alpha_j$ and $\boldsymbol{v}^{\boldsymbol{\alpha}} := \prod_{j=1}^k v_j^{\alpha_j}$.

Definition B.1 (Sparse ReLU network)**.** We use the standard sparse fully connected ReLU-network convention (Schmidt-Hieber, 2020). 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 $f : \mathbb{R}^{W_0} \to \mathbb{R}^{W_L}$ generated by

$$\boldsymbol{z}_0 = \boldsymbol{z}, \qquad \boldsymbol{z}_\ell = \rho\big(\boldsymbol{A}^{(\ell)} \boldsymbol{z}_{\ell-1} + \boldsymbol{b}^{(\ell)}\big) \quad (1 \leq \ell < L), \qquad f(\boldsymbol{z}) = \boldsymbol{A}^{(L)} \boldsymbol{z}_{L-1} + \boldsymbol{b}^{(L)},$$

where $\boldsymbol{A}^{(\ell)} \in \mathbb{R}^{W_\ell \times W_{\ell-1}}$ and $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{W_\ell}$ satisfy

$$\sum_{\ell=1}^{L} \Big(\|\boldsymbol{A}^{(\ell)}\|_0 + \|\boldsymbol{b}^{(\ell)}\|_0 \Big) \leq S, \qquad \max_{1 \leq \ell \leq L} \Big(\|\boldsymbol{A}^{(\ell)}\|_\infty \vee \|\boldsymbol{b}^{(\ell)}\|_\infty \Big) \leq B.$$

For score networks, $W_0 = D + 1$ and $W_L = D$. Input truncation, affine input rescaling, output clipping, time selection, and endpoint projection are fixed postprocessing operations and are not counted as trainable ReLU parameters.

B.2 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; our terminology follows Tu (2011).

Definition B.2 (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 open 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.3 (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.4 (Smooth manifold)**.** A smooth manifold is a manifold equipped with a C^∞ atlas.

Definition B.5 (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.3 Parameter-dependence conventions

The conventions here supplement the global notation and network definition in Sec. B.1. Unadorned constants $C, c > 0$ are universal. The notation $a = \mathcal{O}(b)$ means $a \leq Cb$, whereas $a = \mathcal{O}_\beta(b)$ permits dependence on the fixed smoothness β but not on $d, D, n, \sigma_t, p_{\min}$ or numerical geometry

constants. The relation $a \lesssim b$ suppresses a multiplicative constant independent of the displayed varying quantities; that constant may depend only on fixed parameters explicitly named in the statement. We write $a \lesssim_{\log} b$ if

$$a \le C(1 + \log n + \log D)^c b,$$

where $C, c > 0$ have the same permitted fixed-parameter dependence. In particular, this logarithmic relation does not hide powers of n or D, the density lower bound, reach, volume, finite-order geometry factors, or optimization error. We write $a \asymp b$ when $a \lesssim b$ and $b \lesssim a$. 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 $\widetilde{\mathcal{O}}(\cdot)$ 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.

We reserve $\varepsilon_n = n^{-(\beta+1)/(d+2\beta)}$ for the global distribution-level rate. Local approximation and estimation tolerances use descriptive subscripts and are not alternative names for ε_n. We write the smallest large-noise noise scale as $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$. The large-noise intrinsic cell radius is not identified with σ_{t_1}: on a slab $I_k = [t_{k-1}, t_k)$ (up to the final endpoint) it is $r_k = c_0 \sigma_{t_{k-1}} \wedge r_{\star,n}$, where $r_{\star,n}$ is the radius cap defined in Proposition G.3. 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, $\mathfrak{a}_D, \mathfrak{a}_{\log} \ge 1$ denote fixed finite complexity exponents for direct ambient powers and shell/covering factors, while $\mathfrak{a}_{\mathrm{geom}} \ge 1$ 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 $[\underline{t}, \bar{t}]$. 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 $\inf_{t \in [\underline{t}, \bar{t}]} \sigma_t$ rather than a formal endpoint value; the corresponding schedule lower- and upper-bound convention is stated explicitly in (11).

B.4 Finite-order geometry and tube-volume conventions

Definition B.6 (Finite-order geometry convention)**.** Fix the reach-scale projection atlas used below. For each $q \ge 1$, let $\mathcal{K}_{\mathcal{M},q} \ge e$ 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. It also controls the numbers of patches, inverse coordinate buffers, and a density-atlas transfer envelope $\mathcal{T}_{\mathcal{M},\beta} \ge e$ defined as follows. Cover each $\square_i$ by finitely many buffered convex coordinate patches whose images lie in charts ψ_α of the atlas defining $\mathcal{H}^\beta(\mathcal{M}, B_0)$, and choose a subordinate partition of unity in projection coordinates. Let $\mathcal{T}_{\mathcal{M},\beta}$ bound the number of these patches, their inverse buffers, and the $C^{\lceil \beta \rceil}$ norms of their cutoffs and of $\psi_\alpha \circ \boldsymbol{z}_i$ on the buffered patches. These quantities are finite for the fixed compatible atlas; they must be controlled separately when the manifold or the density atlas varies. The finite geometry constant $\Gamma_{\mathcal{M},q}$ in the approximation proofs may be chosen so that

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

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

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

then $\Gamma_{\mathcal{M},q_{\mathrm{geom}}} \le D^{C_{\beta,d,a_{\mathrm{geom}}}}$ 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 $\widetilde{\mathcal{O}}(\cdot)$ or in universal constants.

B.5 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 $p_{\min}^{-1} \le 1 \vee S_{\mathcal{M}}$ and $B_0 \le 1 \vee S_{\mathcal{M}}^{-1}$, 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.7 (Basic polynomial-geometry examples)**.** *Fix $d \ge 1$, a finite order $q \ge 1$, a fixed exponent $a \ge 0$, and set $q_{\mathrm{emb}} = q + 6$. The following families satisfy the polynomial-geometry regime for the reach-scale projection atlas in Definition B.6, when the density class is defined in that atlas. For another compatible density atlas, the same conclusion requires its transfer envelope $\mathcal{T}_{\mathcal{M},\beta}$ to be polynomially bounded as well. For each displayed manifold write $\kappa_D = \mathrm{reach}(\mathcal{M}_D)$. In this proposition, C^k norms of ambient-valued maps use Euclidean target and operator norms in the fixed finite atlas, including the zeroth-order norm. In particular, a translation inside $[0,1]^D$ can contribute a factor $\sqrt{D}$ to the atlas envelope.*

(i) Affine spheres. *Let $D \ge d+1$, let $E_D = \mathbb{R}^{d+1} \times \{0\}^{D-d-1}$, let $\boldsymbol{c}_D = (1/2,\dots,1/2)$, and let $0 < r_D \le 1/4$. 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 $r_D^{-1} \le D^a$, then

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

(ii) Flat product tori. *Let $D \ge 2d$, let $0 < r_{\min,D} \le r_{j,D} \le 1/4$ for $j = 1,\dots,d$, 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 $r_{\min,D}^{-1} \le D^a$, then

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

(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,*

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

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

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

(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 $D \ge m_0$, let $F_D : N \to \mathbb{R}^{D-m_0}$ be a $C^{q_{\mathrm{emb}}}$ map. Suppose $\|F_D\|_{C^{q_{\mathrm{emb}}}(N)} \le D^a$ 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 $D^{-a} \le \lambda_D \le 1$ and*

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

Then

$$S_{\mathcal{M}_D} \vee \kappa_D^{-1} \vee \mathcal{K}_{\mathcal{M}_D,q} \le C_{N,\iota,d,q,a} D^{C_{N,\iota,d,q}(1+a)}.$$

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 positive-order derivatives 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 $r_D/2$. The inverse-chart zeroth-order norm is at most $\sqrt{D}$, while the projection residual is bounded by the tube radius. Combining these bounds with derivatives of the displayed formula through order $q+4$ gives $\mathcal{K}_{\mathcal{M}_D,q} \le C_{d,q}(1+\sqrt{D}) r_D^{-C_{d,q}}$, which proves (i).

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 $r_{\min,D}$. Including the inverse-chart zeroth-order norm gives $\mathcal{K}_{\mathcal{M}_D,q} \le C_{d,q}(1+\sqrt{D}) r_{\min,D}^{-C_{d,q}}$, which proves (ii).

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 $S_{\mathcal{M}_D} \le C_N D^{C_d a}$ and $S_{\mathcal{M}_D}^{-1} \le C_N D^{C_d a}$. The reach assumption gives $\kappa_D^{-1} \le D^a$.

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 Lemma D.2 provides 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 $\mathrm{d}\iota$. Write $A_D = \mathrm{d}\Phi_D = \lambda_D(\mathrm{d}\iota, \mathrm{d}F_D)$ in a fixed base chart. For a constant $c_\iota > 0$, $\|A_D v\|_2 \ge c_\iota \lambda_D \|v\|_2$, while $\|A_D\|_{C^{q_{\mathrm{emb}}-1}} \le C_\iota D^a$. Consequently $\|(A_D^\top A_D)^{-1}\|_{\mathrm{op}} \le C_\iota D^{2a}$. Repeatedly using $\partial(G^{-1}) = -G^{-1}(\partial G)G^{-1}$ and $A_D^\dagger = (A_D^\top A_D)^{-1} A_D^\top$ gives

$$\|A_D^\dagger\|_{C^{q_{\mathrm{emb}}-1}} \le C_{N,\iota,d,q} D^{C_{d,q}a}.$$

The translation does not affect positive-order derivatives, but the zeroth-order term must be bounded separately by $\|\Phi_D\|_{C^0} \le \sqrt{D}$. Thus

$$\|\Phi_D\|_{C^{q_{\mathrm{emb}}}} \vee \|A_D^\dagger\|_{C^{q_{\mathrm{emb}}-1}} \vee \kappa_D^{-1} \le C_{N,\iota,d,q} D^{C_{d,q}a+1/2}.$$

Since $D \ge 2$, these quantities are at most $D^{a'}$ for some $a' \le C_{N,\iota,d,q}(1+a)$, after increasing the fixed constant. Applying (iii) with a' proves the displayed bound in (iv). □

Global tube-volume bound. We use the following standard reach-volume consequence only for localization and geometric estimates. 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. To prove it, map the radius-r normal disk bundle of $m\mathcal{M}$ into $\mathbb{R}^D$ by $(y,v) \mapsto y+v$. Every point of $(m\mathcal{M})_r$ has a nearest point on the compact manifold and hence belongs to this image. The reach curvature bound gives $\|S_v\|_{\mathrm{op}} \le \|v\|_2/(m\kappa)$ for the shape operator, so the absolute Jacobian is at most $(1+\|v\|_2/(m\kappa))^d$. The area formula, which gives an upper bound even when this normal map is not injective, yields

$$\mathrm{vol}((m\mathcal{M})_r) \le m^d S_{\mathcal{M}}(1+r/(m\kappa))^d \omega_{D-d} r^{D-d} \le S_{\mathcal{M}}(1 \vee \kappa^{-1})^d (1+r)^d \omega_{D-d} r^{D-d}.$$

This proves (9); the curvature bound follows from positive reach as recorded in Boissonnat et al. (2019, Lemma 5).

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 $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ is independent of $\boldsymbol{Y}$. For every $A > 0$, one may take $C_{\mathrm{dev}}(A) = 1 + \sqrt{2A}$ so that*

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

Proof. We apply the Gaussian concentration inequality for Lipschitz functions (Vershynin, 2018) to upper bound $\Pr[\mathrm{dist}(\boldsymbol{x}, m\mathcal{M}) \geq C_{\mathrm{dev}}(A)\sigma\sqrt{D \vee \log n}]$. 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:

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

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

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

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:

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

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 $r = C_{\mathrm{dev}}(A)\sigma\sqrt{D \vee \log n}$. Taking $C_{\mathrm{dev}}(A) = 1 + \sqrt{2A}$ gives $t := r - \mathbb{E}[h(\boldsymbol{z})] \geq \sqrt{2A}\,\sigma\sqrt{D \vee \log n}$. Substituting this into the concentration bound, we obtain:

$$\Pr[h(\boldsymbol{z}) \geq C_{\mathrm{dev}}(A)\sigma\sqrt{D \vee \log n}] \leq \exp\Big(-A(D \vee \log n)\Big) \leq n^{-A}.$$

This conditional bound is uniform in $\boldsymbol{y} \in \mathcal{M}$, so averaging over $\boldsymbol{Y}$ proves the claim. □

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 $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_D)$ is independent of $\boldsymbol{Y}$. Let $p_\sigma : \mathbb{R}^D \to \mathbb{R}_+$ be the density of $\boldsymbol{X}$. For every desired polynomial tail exponent $A > 0$, take $C_{\mathrm{off}}(A) := C_{\mathrm{dev}}(2A) = 1 + 2\sqrt{A}$; then*

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

Proof. Fix $A > 0$ and set $C_{\mathrm{off}}(A) = C_{\mathrm{dev}}(2A) = 1 + 2\sqrt{A}$. By $\boldsymbol{X} = m\boldsymbol{Y} + \sigma\boldsymbol{Z}$, Tweedie's formula (Efron, 2011) expresses the true score as

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

Consequently, define

$$\mathcal{E}_A := \{\boldsymbol{x} : \mathrm{dist}(\boldsymbol{x}, m\mathcal{M}) \geq C_{\mathrm{off}}(A)\sigma\sqrt{D \vee \log n}\}.$$

Then the squared score mass evaluates as follows:

$$\begin{aligned}
&\int_{\mathcal{E}_A} \Big\| \frac{\nabla p_\sigma(\boldsymbol{x})}{p_\sigma(\boldsymbol{x})} \Big\|_2^2 p_\sigma(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{E}_A} \frac{1}{\sigma^2} \big\| \mathbb{E}[\boldsymbol{Z}|\boldsymbol{X}=\boldsymbol{x}] \big\|_2^2 p_\sigma(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} && \text{(by Tweedie's formula)} \\
&= \frac{1}{\sigma^2} \mathbb{E}\Big[\|\mathbb{E}[\boldsymbol{Z}|\boldsymbol{X}]\|_2^2 \mathbb{1}\{\boldsymbol{X} \in \mathcal{E}_A\} \Big] && \text{(by definition of expectation)} \\
&\le \frac{1}{\sigma^2} \mathbb{E}\Big[\mathbb{E}[\|\boldsymbol{Z}\|_2^2|\boldsymbol{X}] \mathbb{1}\{\boldsymbol{X} \in \mathcal{E}_A\} \Big] && \text{(by Jensen's inequality)} \\
&\le \frac{1}{\sigma^2} \sqrt{\mathbb{E}[\|\boldsymbol{Z}\|_2^4] \Pr[\boldsymbol{X} \in \mathcal{E}_A]} && \text{(by Cauchy-Schwarz)} \\
&\le \frac{1}{\sigma^2} \sqrt{(D^2+2D) \Pr[\boldsymbol{X} \in \mathcal{E}_A]} && \text{(by } \mathbb{E}[\|\boldsymbol{Z}\|_2^4] = D^2+2D\text{)} \\
&\lesssim \frac{D}{\sigma^2} \sqrt{\Pr[\boldsymbol{X} \in \mathcal{E}_A]}. && (10)
\end{aligned}$$

By the definition of $C_{\mathrm{off}}(A)$, Lemma C.1 applies with exponent $2A$. Then

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

Combining this with (10) gives

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

□

Lemma C.3 (Off-manifold clipped-score error bound)**.** *Suppose Assumption 1 holds. Fix $\ell_{\mathrm{off}} \ge 1$. For every $A > 0$, let $\varrho_t = C_\varrho(A)\sigma_t\sqrt{D \vee \log n}$, with $C_\varrho(A) = 1 + 2\sqrt{A} = C_{\mathrm{off}}(A)$. For any measurable function $\widetilde{s} : \mathbb{R}^D \times [\underline{t}, \bar{t}] \to \mathbb{R}^D$ 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

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

Proof. Let $\ell_{\mathrm{off}} \ge 1$ 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 $C_\varrho(A) = C_{\mathrm{off}}(A) = 1 + 2\sqrt{A}$; this choice yields the tail exponent required below. By Lemma C.2 it holds that

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

By Lemma 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

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

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

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

□

D SETUP FOR NOISE-REGIME-SPECIFIC 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. Sec. D.2 summarizes how the two approximation constructions lead to the statistical and distributional guarantees. For the OU 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 SMALL- AND LARGE-NOISE REGIMES

Definition D.1 (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 (198). 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 unscripted parameter ε is local to this definition and is not the distribution-level rate ε_n. Target tolerances are named by role, such as coordinate accuracy ε_u, block accuracy $\varepsilon_{\mathrm{blk}}$, and scaled score accuracy $\varepsilon_{\mathrm{sc}}$. The final tangent-cell construction instead uses the direct cell-scale condition $r_{\mathrm{cell}} \le \sigma_t$ in Proposition E.3; it does not use c_{lg} or require overlap with the generic Laplace region.

D.2 NOISE-REGIME-SPECIFIC APPROXIMATION ROADMAP

There are two approximation mechanisms under Assumption 3. 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 factoring out the common normal Gaussian term, and the proof uses a projection-centered Laplace expansion together with neural approximation of nearest-point projection maps.

The proof has four stages:

Proof stage	Main output	Reference
Geometric setup	Projection charts, chart integrals, and the centered ratio $s^* = \sigma_t^{-1}\mathcal{U}/\mathcal{Q}$.	Sec. D.3
Moderate/high-noise approximation	Comparison scores from tangent-cell moment approximations and a denominator lower bound.	Theorem E.5
Small-noise approximation	Comparison scores from projection networks and a Laplace expansion after removing the common normal Gaussian factor.	Corollary F.33
Estimation and distribution recovery	Learned-score risk bounds and their conversion to W_1 error.	Sec. G and H

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 Assumption 1 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 bound (Boissonnat et al., 2019, Corollary 3) gives $\|P_{T_{\boldsymbol{z}}\mathcal{M}} - P_{T_{\boldsymbol{y}}\mathcal{M}}\|_{\mathrm{op}} \le 2\|\boldsymbol{z} - \boldsymbol{y}\|_2/\kappa$ when $\|\boldsymbol{z} - \boldsymbol{y}\|_2 < \kappa/2$. It makes tangent projection a local diffeomorphism on $\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r)$ for $r \le \kappa/8$. To verify global injectivity on this patch, suppose $\boldsymbol{z}_1, \boldsymbol{z}_2$ have the same tangent projection and put $\boldsymbol{v} = \boldsymbol{z}_2 - \boldsymbol{z}_1$. Then $\boldsymbol{v} \perp T_{\boldsymbol{y}}\mathcal{M}$ and

$$\|P_{(T_{\boldsymbol{z}_1}\mathcal{M})^\perp}\boldsymbol{v}\|_2 \ge \sqrt{1 - (2r/\kappa)^2}\|\boldsymbol{v}\|_2.$$

The reach chord inequality also bounds the left side by $\|\boldsymbol{v}\|_2^2/(2\kappa)$. If $\boldsymbol{v} \neq 0$, these inequalities contradict $\|\boldsymbol{v}\|_2 \le 2r \le \kappa/4$. Thus the projection is injective, and its local inverses form a smooth inverse on the image. 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. Take $U_{\boldsymbol{y}} := \mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r_{\mathrm{ch}})$, an open subset of $\mathcal{M}$, 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. For completeness, for $0 < r \le r_{\mathrm{ch}}$ the tangent image of $\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r)$ contains the Euclidean d-ball of radius $r/2$. Indeed, projection is a local diffeomorphism and injective, and any limit point on the patch boundary has ambient distance r from $\boldsymbol{y}$. The reach chord inequality gives tangent-coordinate norm at least $r\sqrt{1 - r^2/(4\kappa^2)} > r/2$ there. Lifting a radial segment in the tangent ball, compactness therefore prevents its lift from terminating before the segment ends. The area formula and (14) give

$$c_d r^d \le \mathrm{vol}_\mathcal{M}(\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}, r)) \le C_d r^d, \qquad 0 < r \le r_{\mathrm{ch}}. \tag{16}$$

Take a maximal $r_{\mathrm{ch}}/(32\sqrt{d})$-separated set. The balls of radius $r_{\mathrm{ch}}/(16\sqrt{d})$ cover $\mathcal{M}$, and the balls of radius $r_{\mathrm{ch}}/(64\sqrt{d})$ are disjoint. Their lower volumes from (16) prove (15).

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

$$\phi_i(\boldsymbol{y}) := \eta_i\big(\boldsymbol{V}_i^\top(\boldsymbol{y} - \boldsymbol{y}_i) + \boldsymbol{c}_i\big), \tag{17}$$

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 Lemma 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 Assumption 1, the Jacobian determinant of the inverse chart ϕ_i^{-1} satisfies, for every $\boldsymbol{u} \in \phi_i(U_i)$,*

$$\sqrt{\det(\nabla\phi_i^{-1}(\boldsymbol{u})^\top \nabla\phi_i^{-1}(\boldsymbol{u}))} \le \frac{1}{\eta_i^d}\Big(1 - \frac{2r_{\mathrm{ch}}}{\kappa}\Big)^{-d/2}. \tag{18}$$

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 (18). □

We define

$$B^{\mathrm{Jac}} := \sup_i \eta_i^{-d}(1 - 2r_{\mathrm{ch}}/\kappa)^{-d/2}. \tag{19}$$

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 $\{U_i\}_{i=1}^{C_{\mathcal{M}}}$. Using the finer cover constructed above, choose it with supports contained in $\mathcal{M} \cap \mathcal{B}(\boldsymbol{y}_i, r_{\mathrm{ch}}/(8\sqrt{d}))$ and hence with compactly supported chart pieces

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

This follows by normalizing smooth nonnegative cutoffs that are one on the covering balls of radius $r_{\mathrm{ch}}/(16\sqrt{d})$ and vanish outside the larger balls. Let

$$\mathcal{S}_i := \mathsf{supp}_{\mathcal{M}}(\rho_i) \Subset U_i, \qquad i = 1, \dots, C_{\mathcal{M}}, \tag{20}$$

be the compact support of the partition of unity on chart i. The coordinate support lies within distance $\eta_i r_{\mathrm{ch}}/(8\sqrt{d})$ of $\phi_i(\boldsymbol{y}_i)$, whereas the coordinate image contains the ball of radius $\eta_i r_{\mathrm{ch}}/2$. Thus it fits inside nested axis-parallel boxes compactly contained in $\phi_i(U_i)$, without further refinement or an increase in the atlas cardinality. 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

$$K_i^{\circ} := \prod_{r=1}^{d} [a_{i,r}^{\circ}, b_{i,r}^{\circ}], \qquad \square_i := \prod_{r=1}^{d} [a_{i,r}^{\square}, b_{i,r}^{\square}], \tag{21}$$

we have

$$\phi_i(\mathcal{S}_i) \subset \mathrm{int}(K_i^{\circ}) \subset K_i^{\circ} \Subset \mathrm{int}(\square_i) \subset \square_i \Subset \phi_i(U_i). \tag{22}$$

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

$$\bar{\rho}_i(\boldsymbol{u}) := (\rho_i \circ \phi_i^{-1})(\boldsymbol{u}), \qquad \boldsymbol{u} \in \phi_i(U_i). \tag{23}$$

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

$$\boldsymbol{z}_i(\boldsymbol{u}) := \phi_i^{-1}(\boldsymbol{u}) \in U_i \subset \mathcal{M}, \tag{24}$$

$$\boldsymbol{J}_i(\boldsymbol{u}) := \nabla \boldsymbol{z}_i(\boldsymbol{u}) \in \mathbb{R}^{D\times d}, \tag{25}$$

$$\boldsymbol{G}_i(\boldsymbol{u}) := \boldsymbol{J}_i(\boldsymbol{u})^\top \boldsymbol{J}_i(\boldsymbol{u}) \in \mathbb{R}^{d\times d}. \tag{26}$$

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

$$\begin{aligned}
p_t(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \int_{\mathcal{M}} \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2}{2\sigma_t^2}\Big) p_0(\boldsymbol{y}) \mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y}) \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_{\mathcal{M}}} \int_{\mathcal{M}} \rho_i(\boldsymbol{y}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2}{2\sigma_t^2}\Big) p_0(\boldsymbol{y}) \mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y}) \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_{\mathcal{M}}} \int_{\phi_i(U_i)} \bar{\rho}_i(\boldsymbol{u}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \phi_i^{-1}(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big) p_0(\phi_i^{-1}(\boldsymbol{u})) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} \mathrm{d}\boldsymbol{u} \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_{\mathcal{M}}} \int_{\phi_i(U_i)} \bar{\rho}_i(\boldsymbol{u}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big) p_0(\boldsymbol{z}_i(\boldsymbol{u})) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} \mathrm{d}\boldsymbol{u}
\end{aligned} \tag{27}$$

Lemma D.4 (Hölder regularity in projection coordinates)**.** *Assume $p_0 \in \mathcal{H}^\beta(\mathcal{M}, B_0)$ in the sense of Definition 2.1. Then $p_0 \circ \boldsymbol{z}_i \in \mathcal{H}^\beta(\square_i)$ for every projection chart, with norms bounded by $C_{\beta,d} B_0 \mathcal{T}_{\mathcal{M},\beta}^{C_{\beta,d}}$. Here $\mathcal{T}_{\mathcal{M},\beta}$ is the transfer envelope in Definition B.6.*

Proof. The compact set $\boldsymbol{z}_i(\square_i) \Subset U_i$ is covered by finitely many relatively compact subcharts of the atlas defining $\mathcal{H}^\beta(\mathcal{M}, B_0)$. Choose a smooth partition of unity on a neighborhood of this compact set subordinate to those subcharts. In each subchart,

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

The chain rule through order m_β and the Hölder bound of order $\beta - m_\beta$ control this composition by CB_0 on the compact support of the partition function. The transition maps are $C^{\lceil\beta\rceil}$ on buffered neighborhoods, so the derivatives and Lipschitz constants required in these composition estimates are finite. Multiplication by the smooth partition functions and zero extension across their buffered supports preserve these bounds. Each chain-rule and product term is a product of a bounded number, depending only on β, d, of derivatives controlled by $\mathcal{T}_{\mathcal{M},\beta}$ and one density derivative controlled by B_0. For pairs crossing the edge of a support, the inverse buffer bounds the Hölder quotient of the zero extension. Summing at most $\mathcal{T}_{\mathcal{M},\beta}$ pieces proves the stated polynomial bound. □

We therefore define the finite density constant

$$B_p^{\mathrm{proj}} := \max_{1 \le i \le C_{\mathcal{M}}} \|p_0 \circ \boldsymbol{z}_i\|_{\mathcal{H}^\beta(\square_i)}. \tag{28}$$

It depends on B_0, β, and the fixed atlas; it need not equal B_0. With the enlarged envelope in Definition B.6, $B_p^{\mathrm{proj}} \le C_{\beta,d} B_0 \mathcal{K}_{\mathcal{M},q}^{C_{\beta,d}}$, so its occurrence is included in the displayed geometry and density factors.

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

$$\begin{aligned}
\nabla p_t(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \int_{\mathcal{M}} \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2^2}{2\sigma_t^2}\Big)\Big(-\frac{\boldsymbol{x} - m_t \boldsymbol{y}}{\sigma_t^2}\Big) p_0(\boldsymbol{y}) \mathrm{dvol}_{\mathcal{M}}(\boldsymbol{y}) \\
&= \frac{1}{(2\pi\sigma_t^2)^{D/2}} \sum_{i=1}^{C_{\mathcal{M}}} \int_{\phi_i(U_i)} \bar{\rho}_i(\boldsymbol{u}) \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big) \\
&\quad \times \Big(-\frac{\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})}{\sigma_t^2}\Big) p_0(\boldsymbol{z}_i(\boldsymbol{u})) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} \mathrm{d}\boldsymbol{u}.
\end{aligned} \tag{29}$$

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

$$a_i^{(\mathcal{Q})}(\boldsymbol{u}) := \bar{\rho}_i(\boldsymbol{u}) \sqrt{\det \boldsymbol{G}_i(\boldsymbol{u})} p_0(\boldsymbol{z}_i(\boldsymbol{u})), \qquad \boldsymbol{a}_i^{(\mathcal{P})}(\boldsymbol{u}) := \boldsymbol{z}_i(\boldsymbol{u}) a_i^{(\mathcal{Q})}(\boldsymbol{u}). \tag{30}$$

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 compactly supported chart amplitudes admit uniform Hölder bounds and zero extensions.

Lemma D.5 (Chart-amplitude Hölder regularity)**.** *Assume Assumptions 1 and 2. Extend $a_i^{(\mathcal{Q})}$ and $a_{i,k}^{(\mathcal{P})}$ by zero outside their coordinate supports. Then, for a finite constant $B_a \le C_{\beta,d} B_p^{\mathrm{proj}} \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C$,*

$$a_i^{(\mathcal{Q})}, a_{i,k}^{(\mathcal{P})} \in \mathcal{H}^\beta(\mathbb{R}^d, B_a), \qquad 1 \le k \le D. \tag{31}$$

They satisfy

$$|a_{i,k}^{(\mathcal{P})}(\boldsymbol{u})| \le a_i^{(\mathcal{Q})}(\boldsymbol{u}) \le B_p^{\mathrm{proj}} B^{\mathrm{Jac}}. \tag{32}$$

Proof. On the compact box $\square_i$, the density in projection coordinates satisfies

$$p_0 \circ \boldsymbol{z}_i \in \mathcal{H}^\beta(\square_i, B_p^{\mathrm{proj}}), \tag{33}$$

and the geometric multipliers have $C^{\lceil\beta\rceil}$ norms controlled by $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$. Leibniz's rule bounds every derivative of their product through order m_β. For the top-order modulus, apply $|fg(u) -$

$fg(v)| \le |f(u)|\,|g(u)-g(v)|+|g(v)|\,|f(u)-f(v)|$ to each Leibniz term. This gives a Hölder radius $C_{\beta,d}B_p^{\mathrm{proj}}\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}$ on $\square_i$. The amplitudes vanish in a neighborhood of $\partial\square_i$ because the partition supports are compactly contained in the inner coordinate boxes. Their derivatives through order m_β therefore extend continuously by zero. For pairs crossing $\partial\square_i$, either both derivatives vanish, or their separation is at least the fixed distance from the support to that boundary. The reciprocal of this distance is included in the geometry envelope, so the same polynomial bound controls the extension's Hölder seminorm. Finally, $0 \le \bar{\rho}_i \le 1$, $|z_{i,k}| \le 1$, and the Jacobian bound prove (32). □

D.3.3 Chart-integral representation of density and score

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

$$F_{i,\boldsymbol{x},t}(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2. \tag{34}$$

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 Lemmas 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 noise-regime-specific approximation and estimation arguments. This same constant determines the near-manifold domain used throughout the noise-regime-specific near-manifold approximation results. The geometric tube-volume constant C_{tube} and the tail threshold C_ϱ are kept separate. 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.

$$\varrho_t := C_\varrho \sigma_t \sqrt{D \vee \log n}, \tag{35}$$

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

$$A_t := \{\boldsymbol{x} \in \mathbb{R}^D : \mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le \varrho_t\}. \tag{36}$$

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 $t \in [\underline{t}, \bar{t}]$ and $\boldsymbol{x} \in A_t$ satisfy

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

We therefore write

$$M_x^\varrho := 1 + C_\varrho \sqrt{D \vee \log n}. \tag{37}$$

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

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

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

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

$$\begin{aligned} \mathcal{U}_i(\boldsymbol{x}, t) &:= \frac{m_t \mathcal{P}_i(\boldsymbol{x}, t) - \boldsymbol{x}\mathcal{Q}_i(\boldsymbol{x}, t)}{\sigma_t} \\ &= \sigma_t^{-d} \int_{\phi_i(U_i)} \frac{m_t \boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{x}}{\sigma_t} \exp\Big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\Big) a_i^{(\mathcal{Q})}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \end{aligned} \tag{40}$$

$$\mathcal{U}(\boldsymbol{x}, t) := \sum_{i=1}^{C_{\mathcal{M}}} \mathcal{U}_i(\boldsymbol{x}, t) = \frac{m_t \mathcal{P}(\boldsymbol{x}, t) - \boldsymbol{x}\mathcal{Q}(\boldsymbol{x}, t)}{\sigma_t}. \tag{41}$$

Then

$$s^*(\boldsymbol{x},t)=\nabla\log p_t(\boldsymbol{x})=\frac{1}{\sigma_t}\frac{\mathcal{U}(\boldsymbol{x},t)}{\mathcal{Q}(\boldsymbol{x},t)}\qquad(p_t(\boldsymbol{x})>0). \tag{42}$$

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

$$s_{\mathrm{reg},C_t}(\boldsymbol{x},t):=\frac{1}{\sigma_t}\frac{\mathcal{U}(\boldsymbol{x},t)}{\mathcal{Q}(\boldsymbol{x},t)\vee C_t}. \tag{43}$$

For reference, by (27) and (29), the same regularized score can also be written in the uncentered form

$$\begin{aligned}\frac{\nabla p_t(\boldsymbol{x})}{p_t(\boldsymbol{x})\vee\rho_t}&=\frac{\frac{1}{(2\pi\sigma_t^2)^{D/2}}\sum_{i=1}^{C_{\mathcal{M}}}\int_{\phi_i(U_i)}\exp\!\big(-\frac{\|\boldsymbol{x}-m_t\phi_i^{-1}(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\big)\big(-\frac{\boldsymbol{x}-m_t\phi_i^{-1}(\boldsymbol{u})}{\sigma_t^2}\big)a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}{\frac{1}{(2\pi\sigma_t^2)^{D/2}}\sum_{i=1}^{C_{\mathcal{M}}}\int_{\phi_i(U_i)}\exp\!\big(-\frac{\|\boldsymbol{x}-m_t\phi_i^{-1}(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\big)a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\vee\rho_t}\\&=-\frac{\boldsymbol{x}}{\sigma_t^2}\frac{\sum_{i=1}^{C_{\mathcal{M}}}\int_{\phi_i(U_i)}\exp\!\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big)a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}{\sum_{i=1}^{C_{\mathcal{M}}}\int_{\phi_i(U_i)}\exp\!\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big)a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\vee(2\pi\sigma_t^2)^{D/2}\rho_t}\\&\quad+\frac{m_t}{\sigma_t^2}\frac{\sum_{i=1}^{C_{\mathcal{M}}}\int_{\phi_i(U_i)}\exp\!\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big)\boldsymbol{a}_i^{(\mathcal{P})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}}{\sum_{i=1}^{C_{\mathcal{M}}}\int_{\phi_i(U_i)}\exp\!\big(-\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2}\big)a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}\vee(2\pi\sigma_t^2)^{D/2}\rho_t}\\&=-\frac{\boldsymbol{x}}{\sigma_t^2}\frac{\mathcal{Q}(\boldsymbol{x},t)}{\mathcal{Q}(\boldsymbol{x},t)\vee C_t}+\frac{m_t}{\sigma_t^2}\frac{\mathcal{P}(\boldsymbol{x},t)}{\mathcal{Q}(\boldsymbol{x},t)\vee C_t},\end{aligned} \tag{44}$$

where

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

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

$$\rho_t=(2\pi\sigma_t^2)^{-D/2}\sigma_t^d C_t. \tag{46}$$

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

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

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

$$0\le\mathcal{Q}(\boldsymbol{x},t)\le M_0,\qquad|\mathcal{P}_k(\boldsymbol{x},t)|\le\mathcal{Q}(\boldsymbol{x},t)\le M_0,\quad k=1,\dots,D. \tag{47}$$

Proof of Lemma D.6. We first show $|\mathcal{P}_k(\boldsymbol{x},t)|\le\mathcal{Q}(\boldsymbol{x},t)$. Since $\mathcal{M}\subset[0,1]^D$, $0\le z_{i,k}(\boldsymbol{u})\le1$, 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 $\exp\!\big(-F_{i,\boldsymbol{x},t}(\boldsymbol{u})/\sigma_t^2\big)\le1$, $\sigma_t^{-d}\le\sigma_{\underline{t}}^{-d}$, and by (32) we have

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

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 (30). 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

$$K_i^{\mathrm{cell}} := \phi_i(\mathcal{S}_i). \tag{48}$$

By (22), $K_i^{\mathrm{cell}} \Subset \square_i$. 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

$$R_i(\boldsymbol{x}, t) := \sigma_t^{-d} \int_{K_i^{\mathrm{cell}}} \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big) a_i^{(R)}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}, \tag{49}$$

We use the same amplitude as a function on $\square_i$; by Lemma 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 $K_i^{\mathrm{cell}} \Subset \square_i$, set

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

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

$$Q_\nu = \boldsymbol{a}_\nu + h[0,1)^d = \boldsymbol{u}_\nu + h[-1/2, 1/2)^d, \qquad \boldsymbol{u}_\nu := \boldsymbol{a}_\nu + \tfrac{h}{2}\mathbf{1}$$

from the standard h-lattice whose closures intersect K_i^{cell}. Here $\boldsymbol{a}_\nu \in h\mathbb{Z}^d$ is the lattice corner and $\boldsymbol{u}_\nu$ is the cell center used in every subsequent Taylor expansion. Then

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

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 (Proposition G.3); 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. For $\boldsymbol{y} \in \mathcal{M}$, define

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

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

$$\frac{\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2}{\sigma_t} \le C_{\mathrm{ph}} H^{1/2}, \tag{50}$$

$$\frac{|T_\nu(\boldsymbol{x},\boldsymbol{y},t)|}{\sigma_t^2} \le C_{\mathrm{ph}}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}\, h\, H^{1/2}, \tag{51}$$

$$\frac{m_t^2 D_\nu(\boldsymbol{x},\boldsymbol{y},t)}{2\sigma_t^2} \le C_{\mathrm{ph}} H. \tag{52}$$

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

$$\frac{\|\boldsymbol{x} - m_t\boldsymbol{y}\|_2^2}{2\sigma_t^2} \ge c_{\mathrm{ph}} AH. \tag{53}$$

Proof. The chart map is C^2 with norm bounded by $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$. In this proof one may take $c_{\mathrm{ph}} = [C_d(1 + \max_i \|\boldsymbol{J}_i\|_{L^\infty(\square_i;\mathrm{op})})]^{-1}$ with sufficiently large C_d, and enlarge C_{ph} polynomially in $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$. In particular, $c_{\mathrm{ph}}^{-1} \le C_d\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C$. 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 (51). 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 (50). 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 (52) 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*

$$\Big|e^{-z} - \sum_{\ell=0}^{J} \frac{(-z)^\ell}{\ell!}\Big| \le e^{-c_{\exp}J}. \tag{54}$$

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

$$\Big|e^{-z} - \sum_{\ell=0}^{K} \frac{(-z)^\ell}{\ell!}\Big| \le C_K a^{K+1} e^a. \tag{55}$$

Proof. For (55), Taylor's theorem with integral remainder gives the claim. For (54), the Lagrange remainder is bounded by

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

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 $I_k = [t_{k-1}, t_k)$ (with T included in the last slab), the cell radius is chosen at the slab noise scale until it reaches the radius required by the fixed-order tangent-cell expansion at the target approximation accuracy. In Proposition G.3 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 ε_n, so the score error is $\sigma_t^{-2}\varepsilon_n^2$ rather than $\sigma_t^{-2} n^{-2\beta/(d+2\beta)}$. Under Assumption 3, the denominator lower bound in Lemma E.4 stabilizes the ratio step; the remaining accuracy dependence enters through the explicit finite-order cell-radius cap. Consequently the aggregate large-noise score networks have 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 Assumptions 1 and 2. Let q_{geom} satisfy (4), and let $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$ be as in Definition B.6. 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*

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

This cell-scale condition is independent of the generic Laplace thresholds in Definition D.1. Define

$$H_{\mathrm{lg}}^{\mathrm{dec}} := 1 + \log\big(C_{\mathrm{lg}}\Gamma_{\mathcal{M},q_{\mathrm{geom}}} B_0^{C_{\mathrm{lg}}} (D \vee \log n)^{C_{\mathrm{lg}}} \sigma_{\underline{t}}^{-d} q_{\mathrm{floor}}^{-1} \varepsilon_{\mathrm{rel}}^{-1}\big). \tag{56}$$

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 $c_{\mathrm{ph,cell}} > 0$:

$$\Gamma_{\mathcal{M},q_{\mathrm{geom}}} r_{\mathrm{cell}} (H_{\mathrm{lg}}^{\mathrm{dec}})^{1/2} \le c_{\mathrm{ph,cell}}, \qquad \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} r_{\mathrm{cell}}^{q_{\mathrm{geom}}-2} H_{\mathrm{lg}}^{\mathrm{dec}} \le c\varepsilon_{\mathrm{rel}}. \tag{57}$$

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 $\boldsymbol{x} \in A_t$,

$$|\mathcal{Q}_i(\boldsymbol{x}, t) - \widetilde{\mathcal{Q}}_i^{\mathrm{lg}}(\boldsymbol{x}, t)| \vee \max_{1 \le k \le D} |\mathcal{U}_{i,k}(\boldsymbol{x}, t) - \widetilde{\mathcal{U}}_{i,k}^{\mathrm{lg}}(\boldsymbol{x}, t)| \le \varepsilon_{\mathrm{rel}} \big(\mathcal{Q}_i(\boldsymbol{x}, t) + q_{\mathrm{floor}} / C_{\mathcal{M}} \big). \tag{58}$$

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

$$\begin{aligned} L \vee \log B &\le C_{\mathrm{lg}} \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^C (H_{\mathrm{lg}}^{\mathrm{dec}})^C, \\ \|\boldsymbol{W}\|_\infty \vee S &\le C_{\mathrm{lg}} \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^C D^{4(q_{\mathrm{geom}}+2)} (D \vee \log n)^{4(q_{\mathrm{geom}}+2)} \varepsilon_{\mathrm{cell}}^{-d/\beta} (H_{\mathrm{lg}}^{\mathrm{dec}})^C. \end{aligned} \tag{59}$$

The following observations explain how the proposition is used. 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 (57) holds. If the slab-scale radius does not satisfy that condition, the cell radius must be capped by $r_{\star,n}$ from Proposition G.3. 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. Write $H = H_{\mathrm{lg}}^{\mathrm{dec}}$, $\sigma_- = \sigma_{\underline{t}}$, and $r = r_{\mathrm{cell}}$. Choose $0 < c_h \le 1 \wedge c_{\mathrm{ph}}$ small enough that cubes of side $h = c_h r$ meeting K_i^{cell} are contained in $\square_i$. The chart derivative bounds and inverse coordinate buffers give $c_h^{-1} \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}$, with constants depending only on (d, q_{geom}). Since $r \le \sigma_t$, we have $h \le c_{\mathrm{ph}} \sigma_t$ throughout the stated time interval, without using c_{lg}. The number of cubes is at most $C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} r^{-d}$, and their total coordinate volume is bounded by a fixed constant. For each such cube, put $\boldsymbol{u} = \boldsymbol{u}_\nu + h\boldsymbol{v}$, with $\boldsymbol{v} \in [-1/2, 1/2]^d$, and let $\boldsymbol{z}_\nu^{\mathrm{pol}}$ be the Taylor polynomial of $\boldsymbol{z}_i$ at $\boldsymbol{u}_\nu$ of degree $q_{\mathrm{geom}} - 2$. Taylor's theorem and the geometry envelope give

$$\sup_{\boldsymbol{u} \in Q_\nu} \|\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_\nu^{\mathrm{pol}}(\boldsymbol{u})\|_2 \le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} h^{q_{\mathrm{geom}}-1}. \tag{60}$$

Here and below constants may depend on the fixed (d, q_{geom}). The vector derivative bounds used here are part of the geometric envelope; replacing them by componentwise bounds only adds the displayed polynomial ambient factors.

Localization. On A_t, all inputs lie in the fixed box $|x_k| \le R_x := 1 + C_\varrho \sqrt{D \vee \log n}$. First clip each input coordinate to this box, which leaves the target domain fixed. Approximate the center phase Φ_ν^0 uniformly on this box and the time interval to error at most $H/4$, and apply the exact piecewise-affine ramp that is one on $(-\infty, A_0 H]$ and zero on $[2A_0 H, \infty)$. Denote the resulting ReLU gate by Θ_ν. Then

$$\Theta_\nu = 1 \text{ if } \Phi_\nu^0 \le (A_0 - 1/4)H, \qquad \Theta_\nu > 0 \implies \Phi_\nu^0 < (2A_0 + 1/4)H. \tag{61}$$

In particular, the gate is exactly zero outside the latter region. All final multiplications by this gate use the zero-preserving product network in Lemma I.3. The phase approximation uses the OU schedule and arithmetic lemmas with sufficiently small internal accuracies; since $\log R_x + \log \sigma_-^{-1} \le CH$, its cost is polynomial in D and H.

Let $\boldsymbol{z} = (\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u}))/\sigma_t$. On every cell where the gate is not one, the phase bound in Lemma E.1, after increasing A_0, gives $\|\boldsymbol{z}\|_2^2 \ge cA_0 H$. The elementary bound

$$(1 + \|\boldsymbol{z}\|_2) e^{-\|\boldsymbol{z}\|_2^2/2} \le C e^{-\|\boldsymbol{z}\|_2^2/4}$$

controls the denominator and each centered numerator simultaneously. Consequently the sum of the exact contributions removed or attenuated by the gates is at most

$$C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0 \sigma_-^{-d} e^{-cA_0 H} \le \frac{\varepsilon_{\mathrm{rel}} q_{\mathrm{floor}}}{8 C_{\mathcal{M}}}, \tag{62}$$

by the definition of H and a sufficiently large fixed A_0.

Relative error from polynomial geometry. Define the exact and polynomial phases and centered factors by

$$\phi = \frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}, \qquad \phi_{\mathrm{pol}} = \frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_\nu^{\mathrm{pol}}(\boldsymbol{u})\|_2^2}{2\sigma_t^2},$$
$$b_k = \frac{m_t z_{i,k}(\boldsymbol{u}) - x_k}{\sigma_t}, \qquad b_k^{\mathrm{pol}} = \frac{m_t z_{\nu,k}^{\mathrm{pol}}(\boldsymbol{u}) - x_k}{\sigma_t}.$$

Whenever $\Theta_\nu > 0$, (60) and (61) and $h \lesssim \sigma_t$ imply

$$|\phi - \phi_{\mathrm{pol}}| + \max_k |b_k - b_k^{\mathrm{pol}}| \le \delta, \qquad |b_k| \le C\sqrt{H}, \qquad 0 \le \phi_{\mathrm{pol}} \le CH,$$
$$\delta := C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} r^{q_{\mathrm{geom}}-2} H^{1/2}.$$

Since $H \ge 1$, the bound $\delta(1+\sqrt{H}) \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} r^{q_{\mathrm{geom}}-2} H$ shows that the single power of H in (57) suffices to ensure $\delta(1+\sqrt{H}) \le \varepsilon_{\mathrm{rel}}/16$. Since $|e^{-a} - e^{-b}| \le e^{-a}(e^{|a-b|} - 1)$, we obtain

$$|e^{-\phi} - e^{-\phi_{\mathrm{pol}}}| \vee |b_k e^{-\phi} - b_k^{\mathrm{pol}} e^{-\phi_{\mathrm{pol}}}| \le \frac{\varepsilon_{\mathrm{rel}}}{4} e^{-\phi}. \tag{63}$$

Multiplication by the nonnegative amplitude $a_i^{(\mathcal{Q})}$ and integration therefore give a relative error bounded by $\varepsilon_{\mathrm{rel}}\mathcal{Q}_i/4$. This estimate retains the full Gaussian phase.

Finite polynomial moments. Set $J = \lceil C_J H \rceil$ and $P_J(z) = \sum_{j=0}^J (-z)^j/j!$. On the support of the gate, Lemma E.2 gives

$$|e^{-\phi_{\mathrm{pol}}} - P_J(\phi_{\mathrm{pol}})| \vee |b_k^{\mathrm{pol}}|\,|e^{-\phi_{\mathrm{pol}}} - P_J(\phi_{\mathrm{pol}})| \le C\sqrt{H} e^{-cJ}.$$

Choosing C_J sufficiently large makes the sum of these errors, after integration and multiplication by σ_t^{-d}, at most $\varepsilon_{\mathrm{rel}} q_{\mathrm{floor}}/(8C_{\mathcal{M}})$. Both $P_J(\phi_{\mathrm{pol}})$ and $b_k^{\mathrm{pol}} P_J(\phi_{\mathrm{pol}})$ are polynomials in the d variables $\boldsymbol{v}$ of degree at most

$$K_J := 2(q_{\mathrm{geom}} - 2)J + q_{\mathrm{geom}} - 2.$$

Store the fixed normalized moments

$$\mu_{\nu,\alpha} := \int_{[-1/2,1/2]^d} \boldsymbol{v}^\alpha a_i^{(\mathcal{Q})}(\boldsymbol{u}_\nu + h\boldsymbol{v})\mathrm{d}\boldsymbol{v}, \qquad |\alpha| \le K_J.$$

Their absolute values are at most $C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0$. The coefficients of these polynomials are functions of $(\boldsymbol{x}, m_t, \sigma_t^{-1})$. Compute them by polynomial addition and coefficient convolution in the d intrinsic variables, then take their scalar products with $(\mu_{\nu,\alpha})_\alpha$. Finally multiply by $h^d\sigma_t^{-d}$ and Θ_ν. Thus no expansion into high-degree ambient monomials is required.

Arithmetic accuracy and capacity. There are at most $(K_J+1)^d \le C(J+1)^d$ coefficients. Computing the phase costs at most CD^2 scalar operations for fixed (d, q_{geom}). The recurrence P_{j+1} through successive polynomial powers and the D centered outputs can be implemented with at most $CD^2(J+1)^{2d+3}$ additions and multiplications. These operations can be arranged with $C(J+1)$ stages and at most $CD^2(J+1)^{2d+2}$ simultaneously retained quantities. Each intermediate coefficient on the clipped input box has absolute value at most $\exp(CH^2)$: the phase coefficients have logarithmic size at most CH, and each polynomial power has order at most $J = O(H)$. Choose the approximation error of each scalar product and schedule input below

$$\eta_{\mathrm{ar}} := \exp(-C_{\mathrm{ar}}H^3)\,\varepsilon_{\mathrm{rel}} q_{\mathrm{floor}} \sigma_-^d (C_{\mathcal{M}}\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0)^{-1}.$$

For completeness, let $N_{\mathrm{op}} \le CD^2(J+1)^{2d+3}$ be the operation count and choose $R = \exp(CH^2)$ to bound every exact intermediate value. Clip every computed intermediate value to $[-R, R]$; this is exact ReLU postprocessing and cannot increase its error. If e_ℓ is the largest error after stage ℓ, the product perturbation bound in Lemma I.3 and the affine-sum bounds give

$$e_{\ell+1} \le (1 + CN_{\mathrm{op}}R)e_\ell + CN_{\mathrm{op}}\eta_{\mathrm{ar}}, \qquad e_0 \le \eta_{\mathrm{ar}}.$$

There are at most $C(J+1)$ stages, so $\max_\ell e_\ell \le \eta_{\rm ar}\exp(CH^3)$. The clipping maps fix zero and thus preserve the exact zeroing by the final gate. With $C_{\rm ar}$ sufficiently large, this bounds the error in each normalized cell integral by $\varepsilon_{\rm rel} q_{\rm floor}\sigma_-^d/(8C_{\mathcal M}V_i)$, where $V_i \ge 1+\sum_\nu h^d$ is fixed. The same induction covers the affine moment sums and the schedule factors. After exact multiplication by h^d, summing over cells costs only $\sum_\nu h^d \le V_i$. The required logarithmic accuracy is polynomial in H and is independent of h^{-1}. Outside the gate support the output is exactly zero; on its support the preceding analytic estimates apply.

More explicitly, each two-input product has depth and sparsity at most CH^3, since $\log(\eta_{\rm ar}^{-1}) + \log R \le CH^3$. The schedule subnetworks have depth and log-weight at most CH^6 and sparsity at most CDH^{12} by Lemmas I.2 and I.9. Padding parallel subnetworks by signed identity channels costs at most their retained width times the common depth. The operation and stage counts above therefore give, including all cells and their output sums,

$$L \vee \log B \le C\Gamma^C_{\mathcal M,q_{\rm geom}} B_0^C H^6, \qquad \|\boldsymbol W\|_\infty \vee S \le C\Gamma^C_{\mathcal M,q_{\rm geom}} B_0^C D^2 r^{-d} H^{2d+12}. \tag{64}$$

The constants and geometry powers may depend on $(d, q_{\rm geom})$, but the displayed powers of H do not hide such dependence. These bounds imply (59). Adding (62) and (63) and the two absolute-error budgets proves (58). Parallelizing the at most Ch^{-d} cell networks proves the asserted capacity bound. □

Lemma E.4 (Large-noise denominator lower bound)**.** *Assume Assumptions 1 to 3. There are constants $c_{\rm lb}, C_{\rm lb} > 0$, depending only on the displayed finite geometry controls, such that*

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

after enlarging the finite-order geometry constant from Definition B.6. For every $t \in [\underline t, \overline t]$ and every

$$\boldsymbol x \in A_t = \{\operatorname{dist}(\boldsymbol x, m_t\mathcal M) \le C_\varrho \sigma_t \sqrt{D \vee \log n}\},$$
$$\mathcal Q(\boldsymbol x, t) \ge q_{\rm lb} := c_{\rm lb} p_{\min} \exp\{-C_{\rm lb} C_\varrho^2 (D \vee \log n)\}. \tag{65}$$

Moreover

$$\frac{\|\mathcal U(\boldsymbol x,t)\|_2}{\mathcal Q(\boldsymbol x,t)} \le R_{\rm lb} := 1 + \Big[2\log\big(e\sigma_{\underline t}^{-d} q_{\rm lb}^{-1}\big) + C_{\rm lb}(D \vee \log n)\Big]^{1/2}. \tag{66}$$

Proof. Step 1: lower-bound the local mass. Fix $(\boldsymbol x, t)$ in the displayed tube and choose $\boldsymbol y_0 \in \mathcal M$ with $\|\boldsymbol x - m_t \boldsymbol y_0\|_2 = \operatorname{dist}(\boldsymbol x, m_t\mathcal M)$. Set $B_{\mathcal M}(\boldsymbol y, r) := \mathcal M \cap \mathcal B(\boldsymbol y, r)$ in this proof and choose $0 < a_{\rm lb} \le r_{\rm ch} \wedge 1$. By (16) and the finite-order geometry convention, $a_{\rm lb}^{-1} \le C\Gamma^C_{\mathcal M,q_{\rm geom}}$ and

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

For $\boldsymbol y \in B_{\mathcal M}(\boldsymbol y_0, a_{\rm lb}\sigma_t)$, the triangle inequality gives

$$\|\boldsymbol x - m_t \boldsymbol y\|_2 \le \operatorname{dist}(\boldsymbol x, m_t\mathcal M) + m_t a_{\rm lb}\sigma_t \le C\sigma_t\sqrt{D \vee \log n}.$$

The density lower bound therefore gives

$$\begin{aligned}\mathcal Q(\boldsymbol x,t) &= \sigma_t^{-d}\int_{\mathcal M} e^{-\|\boldsymbol x - m_t\boldsymbol y\|_2^2/(2\sigma_t^2)} p_0(\boldsymbol y)\mathrm{dvol}_{\mathcal M}(\boldsymbol y)\\ &\ge c p_{\min} e^{-CC_\varrho^2(D\vee\log n)}\\ &\qquad \times \sigma_t^{-d}\operatorname{vol}_{\mathcal M}(B_{\mathcal M}(\boldsymbol y_0, a_{\rm lb}\sigma_t)).\end{aligned}$$

The displayed small-ball volume lower bound gives $\operatorname{vol}_{\mathcal M}(B_{\mathcal M}(\boldsymbol y_0, a_{\rm lb}\sigma_t)) \ge C^{-1}\Gamma^{-C}_{\mathcal M,q_{\rm geom}}\sigma_t^d$. This proves (65) with $c_{\rm lb}^{-1} \le C\Gamma^C_{\mathcal M,q_{\rm geom}}$.

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

$$\mathrm d\nu_{\boldsymbol x,t}(\boldsymbol y) = \frac{\exp(-\|\boldsymbol z(\boldsymbol y)\|_2^2/2)}{\int_{\mathcal M}\exp(-\|\boldsymbol z(\boldsymbol y')\|_2^2/2)\mathrm dP_0(\boldsymbol y')}\mathrm dP_0(\boldsymbol y).$$

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

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

Using $\sigma_t \geq \sigma_{\underline{t}}$, $\mathcal{Q}(\boldsymbol{x},t) \geq q_{\mathrm{lb}}$, and enlarging C_{lb} gives (66). □

Theorem E.5 (Global tangent-cell large-noise score approximation)**.** *Assume Assumptions 1 to 3. Fix $0 < \varepsilon_{\mathrm{cell}} \leq 1/2$ and $0 < \varepsilon_{\mathrm{sc}} \leq 1$. Let q_{geom} be as in (4), and let $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$ be the finite geometry constant from Definition B.6. Let $r_{\mathrm{cell}} = \varepsilon_{\mathrm{cell}}^{1/\beta}$ and*

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

Let q_{lb} and R_{lb} be defined in (65) and (66), 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 $c_{\mathrm{ph,cell}}$:

$$\Gamma_{\mathcal{M},q_{\mathrm{geom}}} r_{\mathrm{cell}}(H_{\mathrm{lg}}^{\mathrm{lb}})^{1/2} \leq c_{\mathrm{ph,cell}}, \qquad \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} r_{\mathrm{cell}}^{q_{\mathrm{geom}}-2} H_{\mathrm{lg}}^{\mathrm{lb}} \leq c\varepsilon_{\mathrm{rel}}. \tag{67}$$

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 $\boldsymbol{x} \in A_t$,

$$\sigma_t\|\widetilde{s}_{\mathrm{lb}}^{\mathrm{lg}}(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_\infty \leq \varepsilon_{\mathrm{sc}}. \tag{68}$$

Proposition E.6 (Complexity of the global large-noise score network)**.** *Under the assumptions and construction of Theorem E.5, its depth L, maximum width $\|\boldsymbol{W}\|_\infty$, sparsity S, and log-weight bound $\log B$ satisfy*

$$\begin{aligned} L \vee \log B &\leq C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C}, \\ \|\boldsymbol{W}\|_\infty \vee S &\leq CD^{4(q_{\mathrm{geom}}+2)}(D\vee\log n)^{4(q_{\mathrm{geom}}+2)}\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C}\varepsilon_{\mathrm{cell}}^{-d/\beta}(H_{\mathrm{lg}}^{\mathrm{lb}})^{C}. \end{aligned} \tag{69}$$

The size bound has the following slabwise interpretation. 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 of Theorem E.5 and Proposition E.6. Step 1: approximate the unrefactored chart integrals. Apply Proposition E.3 with $q_{\mathrm{floor}} = C_{\mathrm{den}}$ and $\varepsilon_{\mathrm{rel}}$ as above. The definition of $H_{\mathrm{lg}}^{\mathrm{lb}}$ dominates the logarithmic scale $H_{\mathrm{lg}}^{\mathrm{dec}}$, and (67) implies (57). After summing over the fixed number of charts,

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

By (65), $C_{\mathrm{den}} \le \mathcal{Q}/8$, so after decreasing c_{glob},

$$|\widehat{\mathcal{Q}} - \mathcal{Q}| \vee \|\widehat{\mathcal{U}} - \mathcal{U}\|_\infty \le c\varepsilon_{\mathrm{rel}}\mathcal{Q}. \tag{70}$$

Step 2: pass from integral approximation to ratio approximation. The map $q \mapsto q \vee C_{\mathrm{den}}$ is exactly representable by a ReLU network, and on the present domain $\mathcal{Q} \ge 8C_{\mathrm{den}}$. Thus $\widehat{\mathcal{Q}} \vee C_{\mathrm{den}} \ge \mathcal{Q}/2$. Using (66) and (70),

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

The reciprocal network is applied only on the interval $[C_{\mathrm{den}}, 2M_0]$, where M_0 is the uniform chart-integral bound from Lemma D.6; the preceding approximation and $\mathcal{Q} \le M_0$ put $\widehat{\mathcal{Q}} \vee C_{\mathrm{den}}$ in this interval after decreasing c_{glob}. Let $M_U := 1 + 2M_0R_{\mathrm{lb}}$ and approximate this reciprocal with error at most $\varepsilon_{\mathrm{sc}}/(16M_U)$. Use the reciprocal lemma with parameter at most the minimum of this tolerance, C_{den}, $(2M_0)^{-1}$, and $1/2$. The numerator-reciprocal products then give a scaled-score network v with $\|v - \mathcal{U}/\mathcal{Q}\|_\infty \le \varepsilon_{\mathrm{sc}}/2$. Clip v at $1 + R_{\mathrm{lb}}$, approximate σ_t^{-1} with additive error $\varepsilon_{\mathrm{sc}}/[8(1 + R_{\mathrm{lb}})]$, and approximate the final products with additive error $\varepsilon_{\mathrm{sc}}/8$. Multiplying the resulting error by $\sigma_t \le 1$ proves the desired scaled-score tolerance. The logarithms of all these accuracies and input ranges are bounded by fixed powers of $H_{\mathrm{lg}}^{\mathrm{lb}}$, so the arithmetic costs satisfy the stated bounds. This proves (68).

Step 3: record size. The chart networks have the size in (59) with $H_{\mathrm{lg}}^{\mathrm{dec}} \le H_{\mathrm{lg}}^{\mathrm{lb}}$. The logarithms of the reciprocal input range, its requested accuracy, and the final schedule accuracy are at most $CH_{\mathrm{lg}}^{\mathrm{lb}}$. Thus the reciprocal and final schedule subnetworks have depth and log-weight at most $C(H_{\mathrm{lg}}^{\mathrm{lb}})^2$ and sparsity at most $CD(H_{\mathrm{lg}}^{\mathrm{lb}})^4$. Including identity channels and the fixed atlas sum, (64) consequently also holds for the global score network with $H = H_{\mathrm{lg}}^{\mathrm{lb}}$ and $r = r_{\mathrm{cell}}$. No additional intrinsic net is required. Therefore the width and sparsity scale as the cell count

$$N_{\mathrm{cell}} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}\varepsilon_{\mathrm{cell}}^{-d/\beta},$$

times the explicit ambient and logarithmic polynomial factor, giving (69). □

F SCORE APPROXIMATION IN THE SMALL-NOISE REGIME

With a positive density lower bound, the small-noise denominator can be controlled uniformly after factoring out the common normal Gaussian term. 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 Corollary F.33. 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 rescaled 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 Lemma F.30; the main network-assembly step is Lemma F.31, while Proposition F.27 records the local chart-level Laplace network on the active set where its approximation bounds hold.

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 (20); 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

$$C_{\mathrm{tail}}^{\max} := 1 \vee 2B_p^{\mathrm{proj}} B^{\mathrm{Jac}} \max_{1\le i\le C_{\mathcal{M}}} \sup_{0<\sigma\le 1,\ \boldsymbol{x}\in\mathbb{R}^D} \sigma^{-d} \int_{\phi_i(U_i)} \exp\Big(-\frac{\|\boldsymbol{x}-\boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{8\sigma^2}\Big)\mathrm{d}\boldsymbol{u}. \tag{71}$$

This constant is finite by the chart bi-Lipschitz bounds and the annular estimate used in the proof of Lemma 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 rescaled 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

$$\mathcal{K}_i^{(c_a)}(\varepsilon) := \big\{(\boldsymbol{x},t) : t\in[\underline{t},\bar{t}],\ \boldsymbol{x}\in A_t, \exists \boldsymbol{y}\in\mathcal{S}_i \text{ s.t. } \|\boldsymbol{x}-m_t\boldsymbol{y}\|_2 \le c_a\sigma_t\sqrt{\log\varepsilon^{-1}}\big\}. \tag{72}$$

The default active set is

$$\mathcal{K}_i(\varepsilon) := \mathcal{K}_i^{(c_\star)}(\varepsilon). \tag{73}$$

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 $c_a > 0$, $t \in [\underline{t}, \bar{t}]$, 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 $\bar{\rho}_i(\boldsymbol{u}) \neq 0$,*

$$\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2 > c_a \sigma_t \sqrt{\log \varepsilon^{-1}}. \tag{74}$$

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

$$|\mathcal{Q}_i(\boldsymbol{x}, t)| + \max_{1 \le k \le D} |\mathcal{P}_{i,k}(\boldsymbol{x}, t)| \le C_{\mathrm{tail}} \big(1 + c_a \sqrt{\log(e/\varepsilon)}\big)^d \varepsilon^{c_a^2/8}. \tag{75}$$

Proof of Lemma 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 (74).

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 $r > 0$,

$$\mathrm{vol}_d\{\boldsymbol{u} \in \phi_i(U_i) : \|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2 \le r\} \le C_i (1 + r/\sigma_t)^d \sigma_t^d. \tag{76}$$

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

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

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 $|a_{i,k}^{(\mathcal{P})}| \le |a_i^{(\mathcal{Q})}| \le B_p^{\mathrm{proj}} B^{\mathrm{Jac}}$ by (32) 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 (75). □

By the fixed choice of $c_\star$, the bound in (75) with $c_a = c_\star$, and also with any larger fixed multiplier, is below the allowed $\mathcal{O}(\varepsilon)$ tail error 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* $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$, 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 (20) to (22). 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

$$B_i^{\mathrm{buff}} := 1 \wedge \mathrm{dist}_{\mathbb{R}^D}\Big(\mathcal{S}_i,\ \mathcal{M} \setminus \phi_i^{-1}\big(\mathrm{int}(K_i^\circ)\big)\Big) > 0. \tag{77}$$

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

Auxiliary compact chart domains. Set

$$d_i^\square := \mathrm{dist}(K_i^\circ, \partial \square_i) > 0. \tag{78}$$

On this single compact set $\square_i$, define

$$L_i^{(+)} := \sup_{\boldsymbol{u}\in\square_i} \|\boldsymbol{J}_i(\boldsymbol{u})\|_{\mathrm{op}}, \qquad L_i^{(-)} := \Big(\sup_{\boldsymbol{y}\in\phi_i^{-1}(\square_i)} \|\mathrm{D}\phi_i(\boldsymbol{y})\|_{\mathrm{op}}\Big)^{-1}, \tag{79}$$

$$\lambda_i^{(-)} := \inf_{\boldsymbol{u}\in\square_i} \lambda_{\min}\big(\boldsymbol{G}_i(\boldsymbol{u})\big), \qquad \lambda_i^{(+)} := \sup_{\boldsymbol{u}\in\square_i} \lambda_{\max}\big(\boldsymbol{G}_i(\boldsymbol{u})\big), \tag{80}$$

$$\mathsf{M}_i^{(2)} := 1 \vee \sup_{\boldsymbol{u}\in\square_i} \sum_{\ell=1}^{D} \|\nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u})\|_{\mathrm{op}}. \tag{81}$$

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

$$\boldsymbol{A}_i(\boldsymbol{u}) := \boldsymbol{G}_i(\boldsymbol{u})^{-1} \boldsymbol{J}_i(\boldsymbol{u})^\top, \qquad \boldsymbol{u} \in \square_i. \tag{82}$$

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

$$L_{A,i} := 1 \vee \sup_{\boldsymbol{u}\in\square_i} \|\mathrm{D}\boldsymbol{A}_i(\boldsymbol{u})\|_{\mathrm{op}} < \infty. \tag{83}$$

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

$$L_i^{(-)} \|\boldsymbol{u} - \boldsymbol{v}\|_2 \le \|\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{z}_i(\boldsymbol{v})\|_2 \le L_i^{(+)} \|\boldsymbol{u} - \boldsymbol{v}\|_2, \tag{84}$$

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

$$\lambda_i^{(-)} \boldsymbol{I}_d \preceq \boldsymbol{G}_i(\boldsymbol{u}) \preceq \lambda_i^{(+)} \boldsymbol{I}_d. \tag{85}$$

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, use the affine extension of ϕ_i to $\mathbb{R}^D$:

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

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 (85). □

Lemma F.3 (Finite-order geometry from compact smoothness)**.** *Assume Assumption 1. Fix an integer $q \ge 1$. 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,

$$\max_{|\boldsymbol{\alpha}|\leq q+4} \|\partial^{\boldsymbol{\alpha}} \boldsymbol{z}_i\|_{L^\infty(\square_i)} \leq \Gamma_{\mathcal{M},q}, \tag{86}$$

$$\max_{|\boldsymbol{\alpha}|\leq q+4} \|\partial^{\boldsymbol{\alpha}} \bar{\rho}_i\|_{L^\infty(\square_i)} + \max_{|\boldsymbol{\alpha}|\leq q+4} \|\partial^{\boldsymbol{\alpha}} \sqrt{\det \boldsymbol{G}_i}\|_{L^\infty(\square_i)} \leq \Gamma_{\mathcal{M},q}. \tag{87}$$

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. We also require $\Gamma_{\mathcal{M},q} \geq \mathcal{K}_{\mathcal{M},q}$, thereby including the density-atlas transfer envelope. With $\mathcal{K}_{\mathcal{M},q}$ denoting the high-order atlas envelope from Definition B.6, 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.

Schedule-dependent maps such as $(\boldsymbol{x}, t) \mapsto \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t))$ are obtained by composing the geometric maps above with the separately approximated schedule m_t^{-1}. Their constants depend only on $\Gamma_{\mathcal{M},q}$ and the fixed lower bound $\underline{m}$. The constant $\Gamma_{\mathcal{M},q}$ may also be enlarged to dominate the finite atlas cardinality, coordinate buffers, tube constants, chart-volume constants, and their reciprocals.

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 (86). 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 (87) 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. Positive reach gives uniqueness of the nearest projection on the open κ-tube (Federer, 1959, Theorem 4.8). For a smooth manifold, the normal-bundle map $(y, v) \mapsto y + v$ has invertible tangential derivative $I - S_v$ when $\|v\|_2 < \kappa$, because $\|S_v\|_{\mathrm{op}} \leq \|v\|_2/\kappa$ (Boissonnat et al., 2019, Lemma 5). Its smooth local inverses agree by uniqueness, so $\Pi_{\mathcal{M}}$ is smooth on this open tube and has bounded derivatives on each compact subtube. 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 $[\underline{t}, \bar{t}]$. 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

$$\delta_{\mathrm{act}}^{(c_a)}(\varepsilon, t) := c_a \frac{\sigma_t}{m_t}\sqrt{\log \varepsilon^{-1}}, \qquad \boldsymbol{\xi}(\boldsymbol{x}, t) := \frac{\boldsymbol{x}}{m_t}. \tag{88}$$

We write

$$\delta_{\mathrm{act}}(\varepsilon, t) := \delta_{\mathrm{act}}^{(c_\star)}(\varepsilon, t) \tag{89}$$

Since the atlas is finite, define the uniform projection and Gauss–Newton margin using a fixed constant $A_{\mathrm{tail}} \geq 1$ large enough for the Gaussian polynomial-tail estimates in Lemma F.31, and writing $\delta_{\mathrm{box}} := \min_i \operatorname{dist}(K_i^\circ, \partial\square_i) > 0$,

$$\eta_\Pi := \min_{1\leq i\leq C_{\mathcal{M}}} \min\Big\{\kappa, \frac{1}{2}B_i^{\mathrm{buff}}, \frac{\lambda_i^{(-)}}{\mathsf{M}_i^{(2)}}, \frac{1}{2L_{A,i}}, \frac{c_\star \delta_{\mathrm{box}}}{A_{\mathrm{tail}}}\Big\} > 0. \tag{90}$$

The generic projection-centered constructions below use $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$; the later H-localized variants state their reach and accuracy conditions directly. Choose $\varepsilon > 0$ small enough that, for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$,

$$\delta_{\mathrm{act}}(\varepsilon, t) < \frac{1}{2}\eta_\Pi. \tag{91}$$

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

$$\mathcal{X}_i := \{\boldsymbol{\zeta} \in \mathbb{R}^D : \operatorname{dist}_{\mathbb{R}^D}(\boldsymbol{\zeta}, \mathcal{S}_i) \leq \eta_\Pi/2\}. \tag{92}$$

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

$$\big\|\nabla_{\boldsymbol{u}}^2 \Phi_i(\boldsymbol{u}; \boldsymbol{\zeta}) - \nabla_{\boldsymbol{u}}^2 \Phi_i(\boldsymbol{v}; \boldsymbol{\zeta})\big\|_{\mathrm{op}} \leq H_i \|\boldsymbol{u} - \boldsymbol{v}\|_2 \tag{93}$$

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

$$r_i := \min\Big\{d_i^{\square}/2,\ \lambda_i^{(-)}/[4(1 + H_i)]\Big\} > 0. \tag{94}$$

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

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

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

$$\mathfrak{C}_{\mathcal{M}} := \max_{1\leq i\leq C_{\mathcal{M}}} \mathfrak{C}_i < \infty. \tag{95}$$

Lemma F.4 (Nearest-point projection on the active small-noise chart)**.** *Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and (91). 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*

$$\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) := \phi_i\big(\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\big) \in K_i^\circ. \tag{96}$$

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

$$\boldsymbol{J}_i\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\big)^\top \boldsymbol{r}_i(\boldsymbol{x}, t) = 0. \tag{97}$$

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

$$\|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \leq \delta_{\mathrm{act}}(\varepsilon, t). \tag{98}$$

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

$$\operatorname{dist}(\boldsymbol{\xi}, \mathcal{M}) \leq \|\boldsymbol{\xi} - \boldsymbol{y}\|_2 \leq \delta_{\mathrm{act}}(\varepsilon, t) < \kappa. \tag{99}$$

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 $\boldsymbol{\xi}$,

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

Therefore,

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

Hence

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

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

$$\Phi_i(\boldsymbol{u}; \boldsymbol{x}, t) := \frac{1}{2}\|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{z}_i(\boldsymbol{u})\|_2^2, \qquad \boldsymbol{g}_i(\boldsymbol{u}; \boldsymbol{x}, t) := \nabla_{\boldsymbol{u}} \Phi_i(\boldsymbol{u}; \boldsymbol{x}, t). \tag{100}$$

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

$$\boldsymbol{g}_i(\boldsymbol{u}; \boldsymbol{x}, t) = \boldsymbol{J}_i(\boldsymbol{u})^\top \big(\boldsymbol{z}_i(\boldsymbol{u}) - \boldsymbol{\xi}(\boldsymbol{x}, t)\big). \tag{101}$$

Differentiating once more gives

$$\nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}; \boldsymbol{x}, t) = \boldsymbol{G}_i(\boldsymbol{u}) - \sum_{\ell=1}^{D} \big(\xi_\ell(\boldsymbol{x}, t) - z_{i,\ell}(\boldsymbol{u})\big) \nabla_{\boldsymbol{u}}^2 z_{i,\ell}(\boldsymbol{u}). \tag{102}$$

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 (97),

$$\begin{aligned}\boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t); \boldsymbol{x}, t) &= \boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top \big(\Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t)) - \boldsymbol{\xi}(\boldsymbol{x}, t)\big) \\ &= -m_t^{-1} \boldsymbol{J}_i(\boldsymbol{u}_i^{\Pi})^\top \boldsymbol{r}_i(\boldsymbol{x}, t) = 0.\end{aligned} \tag{103}$$

Lemma F.5 (Uniform invertibility near the projection)**.** *Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and (91). If $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, then*

$$\nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t); \boldsymbol{x}, t) \succeq \frac{1}{2} \lambda_i^{(-)} \boldsymbol{I}_d. \tag{104}$$

Proof. Write $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{x}, t)$ and $\boldsymbol{u}_i^{\Pi} = \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)$. By Lemma 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 (102),

$$\boldsymbol{w}^\top \nabla_{\boldsymbol{u}} \boldsymbol{g}_i(\boldsymbol{u}_i^\Pi; \boldsymbol{x}, t)\boldsymbol{w} = \boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u}_i^\Pi)\boldsymbol{w} - \sum_{\ell=1}^{D} \big(\xi_\ell - z_{i,\ell}(\boldsymbol{u}_i^\Pi)\big)\boldsymbol{w}^\top \nabla^2_{\boldsymbol{u}} z_{i,\ell}(\boldsymbol{u}_i^\Pi)\boldsymbol{w}. \tag{105}$$

Because $\boldsymbol{u}_i^\Pi \in \square_i$, (85) 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 $\mathsf{M}_i^{(2)}$:

$$\begin{aligned}
\left| \sum_{\ell=1}^{D} \big(\xi_\ell - z_{i,\ell}(\boldsymbol{u}_i^\Pi)\big)\boldsymbol{w}^\top \nabla^2_{\boldsymbol{u}} z_{i,\ell}(\boldsymbol{u}_i^\Pi)\boldsymbol{w} \right| &\leq \sum_{\ell=1}^{D} |\xi_\ell - z_{i,\ell}(\boldsymbol{u}_i^\Pi)| \|\nabla^2_{\boldsymbol{u}} z_{i,\ell}(\boldsymbol{u}_i^\Pi)\|_{\mathrm{op}} \|\boldsymbol{w}\|_2^2 \\
&\leq \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 \sum_{\ell=1}^{D} \|\nabla^2_{\boldsymbol{u}} z_{i,\ell}(\boldsymbol{u}_i^\Pi)\|_{\mathrm{op}} \|\boldsymbol{w}\|_2^2 \\
&\leq \mathrm{dist}(\boldsymbol{\xi}, \mathcal{M})\, \mathsf{M}_i^{(2)} \|\boldsymbol{w}\|_2^2.
\end{aligned} \tag{106}$$

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 (90) and (91),

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

Combining this with (105) and (106) 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 Lemma I.2 and the schedule and arithmetic lemmas in Sec. I.1.

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 $A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), t\big)$.

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.

The projection and active-chart coordinate guarantees are Theorem F.13 and Corollary F.14; the residual, Gaussian-factor, and coefficient approximations are developed in Sec. F.3 and assembled in Proposition F.27.

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

We approximate the nearest-point projection in local intrinsic coordinates using the classical Gauss–Newton method for nonlinear least squares (Nocedal & Wright, 2006, Chapter 10). 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:

$$\Pi_i^{(\mathrm{box})}(\boldsymbol{u}) := \big(a_{i,j}^{\square} + \mathsf{ReLU}(u_j - a_{i,j}^{\square}) - \mathsf{ReLU}(u_j - b_{i,j}^{\square})\big)_{j=1}^{d}. \tag{107}$$

$\Pi_i^{(\mathrm{box})}$ 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*

$$\mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t) := \Pi_i^{(\mathrm{box})}\big(\boldsymbol{u} + \boldsymbol{A}_i(\boldsymbol{u})\big(\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{z}_i(\boldsymbol{u})\big)\big), \qquad \textit{where } \boldsymbol{A}_i(\boldsymbol{u}) = \boldsymbol{G}_i(\boldsymbol{u})^{-1}\boldsymbol{J}_i(\boldsymbol{u})^{\top}. \tag{108}$$

Set

$$r_i^{\mathrm{GN}} := \min\Big\{1, \frac{1}{2}\mathrm{dist}(K_i^{\circ}, \partial\square_i), \frac{1}{4L_{A,i}L_i^{(+)}}\Big\}. \tag{109}$$

Assume $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$, $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$, and (91). Then

$$\|\mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_2 \le \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_2 \tag{110}$$

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 $\boldsymbol{h} \in \mathbb{R}^d$,

$$\begin{aligned} \mathrm{D}_{\boldsymbol{u}}\overline{\mathcal{T}}_i(\boldsymbol{u}; \boldsymbol{x}, t)[\boldsymbol{h}] &= \boldsymbol{h} + \mathrm{D}\boldsymbol{A}_i(\boldsymbol{u})[\boldsymbol{h}]\big(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\big) - \boldsymbol{A}_i(\boldsymbol{u})\boldsymbol{J}_i(\boldsymbol{u})\boldsymbol{h} \\ &= \mathrm{D}\boldsymbol{A}_i(\boldsymbol{u})[\boldsymbol{h}]\big(\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u})\big). \end{aligned} \tag{111}$$

By Lemma 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

$$\overline{\mathcal{T}}_i(\boldsymbol{u}^{\Pi}; \boldsymbol{x}, t) = \boldsymbol{u}^{\Pi}. \tag{112}$$

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

$$\Pi_i^{(\mathrm{box})}(\boldsymbol{u}^{\Pi}) = \boldsymbol{u}^{\Pi}. \tag{113}$$

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

$$\operatorname{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta_{\mathrm{act}}(\varepsilon, t) < \frac{1}{4L_{A,i}}. \tag{114}$$

Therefore, by the chart Lipschitz bound (84),

$$\begin{aligned}
\|\boldsymbol{\xi} - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 &\le \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 + \|\Pi_{\mathcal{M}}(\boldsymbol{\xi}) - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 \\
&= \operatorname{dist}(\boldsymbol{\xi}, \mathcal{M}) + \|\boldsymbol{z}_i(\boldsymbol{u}^\Pi) - \boldsymbol{z}_i(\boldsymbol{u}_s)\|_2 \\
&\le \delta_{\mathrm{act}}(\varepsilon, t) + L_i^{(+)} \|\boldsymbol{u}_s - \boldsymbol{u}^\Pi\|_2 \\
&< \frac{1}{4L_{A,i}} + L_i^{(+)} r_i^{\mathrm{GN}} \\
&\le \frac{1}{4L_{A,i}} + \frac{1}{4L_{A,i}} = \frac{1}{2L_{A,i}}.
\end{aligned} \tag{115}$$

Combining (83), (111) and (115), 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 (112),

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

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

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

□

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

$$\forall \boldsymbol{u} \in K_i^\circ, \qquad \min_{\boldsymbol{a} \in \mathcal{A}_i^{\mathrm{anc}}} \|\boldsymbol{a} - \boldsymbol{u}\|_2 \le \frac{1}{2} r_i^{\mathrm{GN}}. \tag{116}$$

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

$$\begin{aligned}
N_i &\le \prod_{r=1}^{d} \left(1 + \left\lceil \frac{\sqrt{d}\, \ell_{i,r}^\circ}{r_i^{\mathrm{GN}}} \right\rceil\right) \\
&\le \prod_{r=1}^{d} \left(2 + \frac{\sqrt{d}\, \ell_{i,r}^\circ}{r_i^{\mathrm{GN}}}\right) \\
&\le C_i^{\mathrm{anc}} (r_i^{\mathrm{GN}})^{-d}, \\
C_i^{\mathrm{anc}} &:= \prod_{r=1}^{d} \left(2 + \sqrt{d}\, \ell_{i,r}^\circ\right).
\end{aligned} \tag{117}$$

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 $m_{i,r} := \left\lceil \sqrt{d}\, \ell^{\circ}_{i,r} / r_i^{\mathrm{GN}} \right\rceil$. Choose the tensor-product grid

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

For any $\boldsymbol{u} \in K_i^{\circ}$, choose the nearest grid point in each coordinate. The coordinatewise error is at most $\ell^{\circ}_{i,r}/(2m_{i,r}) \le r_i^{\mathrm{GN}}/(2\sqrt{d})$, and therefore the Euclidean error is at most $r_i^{\mathrm{GN}}/2$. This proves (116). The grid cardinality is

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

and the final inequality in (117) follows from $1 + \lceil x \rceil \le 2 + x$, together with $r_i^{\mathrm{GN}} \le 1$. □

Define the aggregate anchor complexity

$$A_{\mathcal{M}} := \sum_{i=1}^{C_{\mathcal{M}}} |\mathcal{A}_i^{\mathrm{anc}}|. \tag{118}$$

By (117),

$$A_{\mathcal{M}} \le \sum_{i=1}^{C_{\mathcal{M}}} C_i^{\mathrm{anc}} (r_i^{\mathrm{GN}})^{-d}. \tag{119}$$

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 (91). Assume $\boldsymbol{a} \in K_i^{\circ}$ satisfies*

$$\|\boldsymbol{a} - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t)\|_2 \le \frac{1}{2} r_i^{\mathrm{GN}}. \tag{120}$$

Let $\widetilde{\mathcal{T}}_i$ be a $\square_i$-valued approximate Gauss–Newton step satisfying

$$\sup_{\boldsymbol{u} \in \square_i,\ t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon),\ (\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)} \big\| \widetilde{\mathcal{T}}_i(\boldsymbol{u}; \boldsymbol{x}, t) - \mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t) \big\|_{\infty} \le \varepsilon_{\mathrm{step}}. \tag{121}$$

Assume moreover that

$$\varepsilon_{\mathrm{step}} \le \frac{r_i^{\mathrm{GN}}}{8\sqrt{d}}. \tag{122}$$

Define the approximate iterates by

$$\widetilde{\boldsymbol{u}}_0 := \boldsymbol{a}, \qquad \widetilde{\boldsymbol{u}}_{k+1} := \widetilde{\mathcal{T}}_i(\widetilde{\boldsymbol{u}}_k; \boldsymbol{x}, t). \tag{123}$$

Then, every iterate stays in the Gauss–Newton contraction basin:

$$\widetilde{\boldsymbol{u}}_k \in \mathcal{B}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t), r_i^{\mathrm{GN}}\big) \cap \square_i, \qquad k \ge 0. \tag{124}$$

Moreover, for every $K \ge 0$,

$$\big\| \widetilde{\boldsymbol{u}}_K - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \big\|_2 \le 2^{-(K+1)} r_i^{\mathrm{GN}} + 2\sqrt{d}\, \varepsilon_{\mathrm{step}}. \tag{125}$$

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 (120),

$$e_0 = \|\boldsymbol{a} - \boldsymbol{u}^{\Pi}\|_2 \le \frac{1}{2} r_i^{\mathrm{GN}}.$$

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 (Lemma 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

$$\begin{aligned} e_{k+1} &= \|\widetilde{\mathcal{T}}_i(\widetilde{\boldsymbol{u}}_k;\boldsymbol{x},t) - \boldsymbol{u}^{\Pi}\|_2 \\ &\le \big\|\widetilde{\mathcal{T}}_i(\widetilde{\boldsymbol{u}}_k;\boldsymbol{x},t) - \mathcal{T}_i(\widetilde{\boldsymbol{u}}_k;\boldsymbol{x},t)\big\|_2 + \|\mathcal{T}_i(\widetilde{\boldsymbol{u}}_k;\boldsymbol{x},t) - \boldsymbol{u}^{\Pi}\|_2 \\ &\le \sqrt{d}\varepsilon_{\mathrm{step}} + \frac{1}{2}e_k \\ &\le \sqrt{d}\varepsilon_{\mathrm{step}} + \frac{3}{8}r_i^{\mathrm{GN}}. \end{aligned}$$

By (122),

$$\sqrt{d}\,\varepsilon_{\mathrm{step}} \le \frac{1}{8}r_i^{\mathrm{GN}}.$$

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 (124) holds.

Now that the exact contraction applies at every step, we have the recursion

$$e_{k+1} \le \frac{1}{2}e_k + \sqrt{d}\,\varepsilon_{\mathrm{step}}. \tag{126}$$

Iterating (126) 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*

$$\sup_{\boldsymbol{u}\in\square_i, t\in\mathcal{I}_{\mathrm{sm}}(\varepsilon), (\boldsymbol{x},t)\in\mathcal{K}_i(\varepsilon)} \big\|\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}}(\boldsymbol{u},\boldsymbol{x},t) - \mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t)\big\|_\infty \le \varepsilon_{\mathcal{T}}. \tag{127}$$

Moreover, for a finite chart-dependent constant $C_{i,\mathcal{T}} \ge 1$,

$$\begin{aligned} L &\lesssim \log(e(d\vee\beta))\log^2(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1}), \\ \|\boldsymbol{W}\|_\infty &\lesssim dD(d\vee\beta)(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1})^{\frac{d}{q_{\mathrm{geom}}}} + D\log^3(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1}), \\ S &\lesssim dD(d+q_{\mathrm{geom}})^{d+3}(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1})^{\frac{d}{q_{\mathrm{geom}}}}\log(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1}) + dD\log^4(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1}), \\ \log B &\lesssim \log^2(C_{i,\mathcal{T}}D\varepsilon_{\mathcal{T}}^{-1}). \end{aligned} \tag{128}$$

Proof. Recall that

$$\mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t) = \Pi_i^{(\mathrm{box})}\Big(\boldsymbol{u} + \boldsymbol{A}_i(\boldsymbol{u})\big(\boldsymbol{\xi}(\boldsymbol{x},t) - \boldsymbol{z}_i(\boldsymbol{u})\big)\Big), \qquad \boldsymbol{A}_i(\boldsymbol{u}) = \boldsymbol{G}_i(\boldsymbol{u})^{-1}\boldsymbol{J}_i(\boldsymbol{u})^\top. \tag{129}$$

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 $B_{i,\mathcal{T}} \ge 1$ such that every scalar component of $\boldsymbol{z}_i$ and $\boldsymbol{A}_i$ belongs to $\mathcal{H}^{q_{\mathrm{geom}}}(\square_i, B_{i,\mathcal{T}})$, and

$$\|\boldsymbol{z}_i(\boldsymbol{u})\|_\infty + \|\boldsymbol{A}_i(\boldsymbol{u})\|_\infty \le B_{i,\mathcal{T}}, \qquad \boldsymbol{u} \in \square_i. \tag{130}$$

Here $\|\boldsymbol{A}_i(\boldsymbol{u})\|_\infty$ denotes the maximum absolute entry of the $d \times D$ matrix $\boldsymbol{A}_i(\boldsymbol{u})$. By Lemma F.3, the Hölder radius at order q_{geom} is bounded by a fixed power of $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$. Applying Theorem I.1 at this order gives scalar size $C\delta^{-d/q_{\mathrm{geom}}}\log(C/\delta)$ at accuracy δ. All projection, coordinate, residual, and Gaussian bounds below use this geometric approximation order.

Moreover, on the active small-noise region there exists a finite constant $M_{x,i} \geq 1$ such that

$$\|\boldsymbol{x}\|_\infty \leq M_{x,i}, \qquad \|\boldsymbol{\xi}(\boldsymbol{x},t)\|_\infty = \left\|\frac{\boldsymbol{x}}{m_t}\right\|_\infty \leq M_{x,i}\underline{m}^{-1}. \tag{131}$$

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

$$C_{i,\times} := 1 \vee B_{i,\mathcal{T}} \vee M_{x,i}\underline{m}^{-1}. \tag{132}$$

Step 2: Approximate the ingredients. Let $\varepsilon_{\mathrm{aux}} \in (0, 1/2]$ be chosen later.

First, by Lemma I.7 with $M_x = M_{x,i}$, there exists a vector-valued ReLU network $\widetilde{\boldsymbol{\xi}}$ such that

$$\left\|\widetilde{\boldsymbol{\xi}}(\boldsymbol{x},t) - \boldsymbol{\xi}(\boldsymbol{x},t)\right\|_\infty \leq \varepsilon_{\mathrm{aux}} \tag{133}$$

uniformly over the active small-noise region. Its size is bounded by

$$\begin{aligned} L_\xi, \log B_\xi &\lesssim \log^2\varepsilon_{\mathrm{aux}}^{-1} + \log^2(C_{i,\times}), \\ \|\boldsymbol{W}_\xi\|_\infty &\lesssim D\big(\log^3\varepsilon_{\mathrm{aux}}^{-1} + \log^3(C_{i,\times})\big), \\ S_\xi &\lesssim D\big(\log^4\varepsilon_{\mathrm{aux}}^{-1} + \log^4(C_{i,\times})\big). \end{aligned} \tag{134}$$

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 $[0,1]^d$. For $\square_i = \prod_{r=1}^d [a_{i,r}^\square, b_{i,r}^\square]$, this rescaling has coordinates $(u_r - a_{i,r}^\square)/(b_{i,r}^\square - a_{i,r}^\square)$ and is implemented by one additional ReLU layer, which is exact on $\square_i$. Its weights and biases are bounded by

$$C_{\mathrm{aff},i} := 1 \vee \max_{1\leq r\leq d} \frac{1+|a_{i,r}^\square|}{b_{i,r}^\square - a_{i,r}^\square} \leq C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^C.$$

The last inequality follows from the coordinate and inverse-buffer bounds in the geometry envelope. The normalized-input approximants have unit weights; composing them with this shared rescaling adds at most $2d$ nonzero parameters and gives the weight bound below in the original, unscaled input coordinates. Thus there exist ReLU networks $\widetilde{\boldsymbol{z}}_i$ and $\widetilde{\boldsymbol{A}}_i$ such that

$$\|\widetilde{\boldsymbol{z}}_i - \boldsymbol{z}_i\|_{L^\infty(\square_i)} \leq \varepsilon_{\mathrm{aux}}, \qquad \|\widetilde{\boldsymbol{A}}_i - \boldsymbol{A}_i\|_{L^\infty(\square_i)} \leq \varepsilon_{\mathrm{aux}}. \tag{135}$$

Stacking the $D + dD$ scalar approximants gives size

$$\begin{aligned} L_H &\lesssim \log(e(d\vee\beta))\log(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1}), \\ \|\boldsymbol{W}_H\|_\infty &\lesssim dD(d\vee\beta)(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1})^{\frac{d}{q_{\mathrm{geom}}}}, \\ S_H &\lesssim dD(d+q_{\mathrm{geom}})^{d+3}(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1})^{\frac{d}{q_{\mathrm{geom}}}}\log(C_{i,\times}\varepsilon_{\mathrm{aux}}^{-1}), \\ B_H &\leq C_{\mathrm{aff},i}. \end{aligned} \tag{136}$$

Enlarging $C_{i,\mathcal{T}}$ to include $C_{\mathrm{aff},i}$ absorbs this additional layer and its weights in (128), without changing any accuracy exponent.

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 (133) and (135),

$$\|\widetilde{\boldsymbol{v}} - \boldsymbol{v}\|_\infty \leq 2\varepsilon_{\mathrm{aux}}. \tag{137}$$

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 $j = 1, \dots, d$ and $\ell = 1, \dots, D$, apply Lemma 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

$$\begin{aligned}&\left|\phi_{\text{mult},j,\ell}\big(\widetilde{A}_{i,j,\ell}(\boldsymbol{u}), \widetilde{v}_\ell(\boldsymbol{u}, \boldsymbol{x}, t)\big) - A_{i,j,\ell}(\boldsymbol{u}) v_\ell(\boldsymbol{u}, \boldsymbol{x}, t)\right| \\ &\quad \le \varepsilon_{\text{aux}} + 4C_{i,\times}\varepsilon_{\text{aux}} \le 5C_{i,\times}\varepsilon_{\text{aux}}. \end{aligned} \tag{138}$$

The size of each multiplication network is

$$\begin{aligned} L_{\text{mult}} &\lesssim \log(\varepsilon_{\text{aux}}^{-1}) + \log(C_{i,\times}), \\ \|\boldsymbol{W}_{\text{mult}}\|_\infty &\le 96, \\ S_{\text{mult}} &\lesssim \log(\varepsilon_{\text{aux}}^{-1}) + \log(C_{i,\times}), \\ \log B_{\text{mult}} &\lesssim \log(C_{i,\times}). \end{aligned} \tag{139}$$

Define the approximate pre-clipped update by

$$\widetilde{q}_j(\boldsymbol{u}, \boldsymbol{x}, t) := u_j + \sum_{\ell=1}^{D} \phi_{\text{mult},j\ell}\big(\widetilde{A}_{i,j\ell}(\boldsymbol{u}), \widetilde{v}_\ell(\boldsymbol{u}, \boldsymbol{x}, t)\big), \qquad j = 1, \dots, d. \tag{140}$$

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 (138) over ℓ gives

$$\|\widetilde{\boldsymbol{q}}(\boldsymbol{u}, \boldsymbol{x}, t) - \boldsymbol{q}(\boldsymbol{u}, \boldsymbol{x}, t)\|_\infty \le 5DC_{i,\times}\varepsilon_{\text{aux}}. \tag{141}$$

Now define the final network by exact clipping:

$$\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}}(\boldsymbol{u}, \boldsymbol{x}, t) := \Pi_i^{(\text{box})}\big(\widetilde{\boldsymbol{q}}(\boldsymbol{u}, \boldsymbol{x}, t)\big). \tag{142}$$

Because $\Pi_i^{(\text{box})}$ is 1-Lipschitz,

$$\begin{aligned}\left\|\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathcal{T}}}(\boldsymbol{u}, \boldsymbol{x}, t) - \mathcal{T}_i(\boldsymbol{u}; \boldsymbol{x}, t)\right\|_\infty &\le \|\widetilde{\boldsymbol{q}}(\boldsymbol{u}, \boldsymbol{x}, t) - \boldsymbol{q}(\boldsymbol{u}, \boldsymbol{x}, t)\|_\infty \\ &\le 5DC_{i,\times}\varepsilon_{\text{aux}}.\end{aligned}$$

Choose

$$\varepsilon_{\text{aux}} := \frac{\varepsilon_{\mathcal{T}}}{5DC_{i,\times}}. \tag{143}$$

Then (127) follows.

Step 4: Network size. The network is obtained by composing the following pieces: the $\boldsymbol{\xi} = \boldsymbol{x}/m_t$ network from Lemma I.7, the d-dimensional Hölder networks for $\boldsymbol{z}_i$ and $\boldsymbol{A}_i$ from Theorem I.1, the dD multiplication networks from Lemma I.3, exact affine summations, and the exact clipping map.

Substituting (143) into (134), (136) and (139) and absorbing the fixed chart constants into a single constant $C_{i,\mathcal{T}} \ge 1$ gives (128). 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 $0 < \varepsilon_{\text{loc}} \le 1$ such that (91) 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*

$$\begin{aligned}\|\boldsymbol{a} - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\|_2 &\le \frac{1}{2} r_i^{\text{GN}}, \\ \left\|\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\text{loc}}}(\boldsymbol{x}, t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\right\|_\infty &\le \varepsilon_u. \end{aligned} \tag{144}$$

Moreover, for a finite chart-dependent constant $C_{i,\Pi} \geq 1$,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3(C_{i,\Pi} D \varepsilon_u^{-1}), \\ \|\boldsymbol{W}\|_\infty &\lesssim dD(d \vee \beta)(C_{i,\Pi} D \varepsilon_u^{-1})^{\frac{d}{q_{\mathrm{geom}}}} + D \log^3(C_{i,\Pi} D \varepsilon_u^{-1}), \\ S &\lesssim dD(d + q_{\mathrm{geom}})^{d+3}(C_{i,\Pi} D \varepsilon_u^{-1})^{\frac{d}{q_{\mathrm{geom}}}} \log^2(C_{i,\Pi} D \varepsilon_u^{-1}) + dD \log^5(C_{i,\Pi} D \varepsilon_u^{-1}), \\ \log B &\lesssim \log^2(C_{i,\Pi} D \varepsilon_u^{-1}). \end{aligned} \tag{145}$$

Proof. The construction uses a fixed anchor $\boldsymbol{a}$ as the initial condition and then unrolls approximate Gauss–Newton steps. Set

$$\varepsilon_{\mathrm{step}} := \min\Big\{ \frac{r_i^{\mathrm{GN}}}{8\sqrt{d}}, \frac{\varepsilon_u}{4\sqrt{d}} \Big\}. \tag{146}$$

By Lemma F.9, applied with localization parameter $\varepsilon_{\mathrm{loc}}$, there exists a ReLU network $\widetilde{\mathcal{T}}_{i,\varepsilon_{\mathrm{step}}}$ with output in $\square_i$ such that

$$\sup_{\boldsymbol{u} \in \square_i,\ t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{loc}}),\ (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon_{\mathrm{loc}})} \Big\| \widetilde{\mathcal{T}}_{i,\varepsilon_{\mathrm{step}}}(\boldsymbol{u},\boldsymbol{x},t) - \mathcal{T}_i(\boldsymbol{u};\boldsymbol{x},t) \Big\|_\infty \leq \varepsilon_{\mathrm{step}}. \tag{147}$$

The choice (146) guarantees in particular

$$\varepsilon_{\mathrm{step}} \leq \frac{r_i^{\mathrm{GN}}}{8\sqrt{d}},$$

which is exactly the basin condition (122) required in Lemma F.8.

Let

$$K_{\varepsilon_u} := \Big\lceil \log_2\Big(\frac{r_i^{\mathrm{GN}}}{\varepsilon_u} \Big) \Big\rceil_+, \qquad \lceil a \rceil_+ := \max\{\lceil a \rceil, 0\}. \tag{148}$$

Then

$$2^{-(K_{\varepsilon_u}+1)} r_i^{\mathrm{GN}} \leq \frac{\varepsilon_u}{2}. \tag{149}$$

Indeed, if $r_i^{\mathrm{GN}} \leq \varepsilon_u$, then $K_{\varepsilon_u} = 0$ and $r_i^{\mathrm{GN}}/2 \leq \varepsilon_u/2$. If $r_i^{\mathrm{GN}} > \varepsilon_u$, then $K_{\varepsilon_u} \geq \log_2(r_i^{\mathrm{GN}}/\varepsilon_u)$, so $2^{-K_{\varepsilon_u}} r_i^{\mathrm{GN}} \leq \varepsilon_u$.

Step 1: Define the unrolled network. Define recursively

$$\widetilde{\boldsymbol{u}}_0(\boldsymbol{x},t) := \boldsymbol{a}, \qquad \widetilde{\boldsymbol{u}}_{k+1}(\boldsymbol{x},t) := \widetilde{\mathcal{T}}_{i,\varepsilon_{\mathrm{step}}}\big(\widetilde{\boldsymbol{u}}_k(\boldsymbol{x},t), \boldsymbol{x}, t\big), \qquad k = 0, \ldots, K_{\varepsilon_u} - 1. \tag{150}$$

Finally set

$$\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\mathrm{loc}}}(\boldsymbol{x},t) := \widetilde{\boldsymbol{u}}_{K_{\varepsilon_u}}(\boldsymbol{x},t). \tag{151}$$

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 $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{loc}})$, $(\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon_{\mathrm{loc}})$, and

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

Then the hypotheses of Lemma F.8 are satisfied with $K = K_{\varepsilon_u}$. Hence

$$\begin{aligned} \Big\| \widetilde{\boldsymbol{u}}_{K_{\varepsilon_u}}(\boldsymbol{x},t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x},t) \Big\|_2 &\leq 2^{-(K_{\varepsilon_u}+1)} r_i^{\mathrm{GN}} + 2\sqrt{d}\, \varepsilon_{\mathrm{step}} \\ &\leq \frac{\varepsilon_u}{2} + 2\sqrt{d} \cdot \frac{\varepsilon_u}{4\sqrt{d}} = \varepsilon_u. \end{aligned} \tag{152}$$

Since $\|\boldsymbol{v}\|_\infty \leq \|\boldsymbol{v}\|_2$, (152) implies

$$\Big\| \widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u}^{\Pi,\varepsilon_{\mathrm{loc}}}(\boldsymbol{x},t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x},t) \Big\|_\infty \leq \varepsilon_u,$$

which proves (144).

Step 3: Network size. By Lemma F.9, the one-step network with accuracy $\varepsilon_{\text{step}}$ satisfies, for a chart-dependent constant $C_{i,\mathcal{T}} \geq 1$,

$$\begin{aligned} L_{\text{step}} &\lesssim \log(e(d \vee \beta)) \log^2(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1}), \\ \|\boldsymbol{W}_{\text{step}}\|_\infty &\lesssim dD(d \vee \beta)(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1})^{\frac{d}{q_{\text{geom}}}} + D \log^3(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1}), \\ S_{\text{step}} &\lesssim dD(d + q_{\text{geom}})^{d+3}(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1})^{\frac{d}{q_{\text{geom}}}} \log(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1}) + dD \log^4(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1}), \\ \log B_{\text{step}} &\lesssim \log^2(C_{i,\mathcal{T}} D \varepsilon_{\text{step}}^{-1}). \end{aligned}$$

From (146), there exists a chart-dependent constant $C_{i,\Pi} \geq 1$ such that

$$D\varepsilon_{\text{step}}^{-1} \leq C_{i,\Pi} D \varepsilon_u^{-1}. \tag{153}$$

Also, by (148),

$$K_{\varepsilon_u} \leq C \log(C_{i,\Pi} D \varepsilon_u^{-1}). \tag{154}$$

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 (153) and (154) gives exactly (145). This completes the proof. □

$$\mathcal{K}_{\text{act}}(\varepsilon) := \bigcup_{i=1}^{C_{\mathcal{M}}} \mathcal{K}_i(\varepsilon). \tag{155}$$

For a fixed auxiliary radius multiplier $c_a > 0$, also write

$$\mathcal{K}_{\text{act}}^{(c_a)}(\varepsilon) := \bigcup_{i=1}^{C_{\mathcal{M}}} \mathcal{K}_i^{(c_a)}(\varepsilon). \tag{156}$$

Lemma F.11 (Active chart centers lie within the reach tube)**.** *Let $\delta_\Pi := \eta_\Pi/2$. Assume $t \in \mathcal{I}_{\text{sm}}(\varepsilon)$ and (91). If $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$ for some $i \in \{1, \ldots, C_{\mathcal{M}}\}$, then*

$$\operatorname{dist}\big(\boldsymbol{\xi}(\boldsymbol{x}, t), \mathcal{M}\big) \leq \delta_\Pi. \tag{157}$$

Moreover, $\delta_\Pi \leq \kappa/2$.

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

$$\|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}\|_2 \leq \delta_{\text{act}}(\varepsilon, t).$$

Therefore

$$\operatorname{dist}(\boldsymbol{\xi}(\boldsymbol{x}, t), \mathcal{M}) \leq \|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}\|_2 \leq \delta_{\text{act}}(\varepsilon, t).$$

By (91),

$$\delta_{\text{act}}(\varepsilon, t) < \frac{1}{2}\eta_\Pi = \delta_\Pi.$$

Hence (157) holds. Since $\eta_\Pi \leq \kappa$ by (90), we also have $\delta_\Pi \leq \kappa/2$. □

Lemma F.12 (Projection objective gap inside the reach tube)**.** *Assume $0 < \delta \leq \kappa/2$. Then, for every $\boldsymbol{\xi} \in \mathbb{R}^D$ satisfying $\operatorname{dist}(\boldsymbol{\xi}, \mathcal{M}) \leq \delta$ and every $\boldsymbol{y} \in \mathcal{M}$,*

$$\frac{1}{2}\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \leq \|\boldsymbol{\xi} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \leq \Big(1 + \frac{\delta}{\kappa}\Big)\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2. \tag{158}$$

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 $\boldsymbol{y} \in \mathcal{M}$,

$$\begin{aligned}\|\boldsymbol{\xi} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 &= \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) + \Pi_{\mathcal{M}}(\boldsymbol{\xi}) - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \\ &= \|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 - 2\langle \boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}), \boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\rangle. \end{aligned} \tag{159}$$

The reach bound gives the tangent-ball inequality

$$\langle \boldsymbol{v}, \boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\rangle \le \frac{\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2}{2\kappa}, \qquad \boldsymbol{y} \in \mathcal{M}, \quad \boldsymbol{v} \in N_{\Pi_{\mathcal{M}}}(\boldsymbol{\xi})\mathcal{M}, \quad \|\boldsymbol{v}\|_2 = 1. \tag{160}$$

If $\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi}) = 0$, the claim follows immediately from (159). Otherwise, applying (160) 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 (159) yields

$$\Big(1 - \frac{\delta}{\kappa}\Big)\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \|\boldsymbol{\xi} - \boldsymbol{y}\|_2^2 - \|\boldsymbol{\xi} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le \Big(1 + \frac{\delta}{\kappa}\Big)\|\boldsymbol{y} - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2.$$

Since $\delta \le \kappa/2$, the lower coefficient is at least $1/2$. This proves (158). □

Theorem F.13 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 implemented 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.13 (Neural approximation of the active projection center)**.** *Fix a localization parameter $0 < \varepsilon < 1$ such that (91) 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

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x}, t) \in \mathcal{K}_{\mathrm{act}}(\varepsilon)}} \Big\|\widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}\big(\boldsymbol{\xi}(\boldsymbol{x}, t)\big)\Big\|_2 \le \varepsilon_\Pi. \tag{161}$$

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* (119)*. For a finite geometry-dependent constant* $C_{\mathrm{proj}} \ge 1$*,* $\widetilde{\Pi}_{\varepsilon_\Pi} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ *with*

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}), \\ \|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta)(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2})^{\frac{d}{q_{\mathrm{geom}}}} + A_{\mathcal{M}} D \log^3(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}), \\ S &\lesssim A_{\mathcal{M}} d D (d + q_{\mathrm{geom}})^{d+3} (C_{\mathrm{proj}} D \varepsilon_\Pi^{-2})^{\frac{d}{q_{\mathrm{geom}}}} \log^2(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}) \\ &\quad + A_{\mathcal{M}} d D \log^5(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}), \\ \log B &\lesssim \log^2(C_{\mathrm{proj}} D \varepsilon_\Pi^{-2}). \end{aligned} \tag{162}$$

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 Lemma F.11, $\mathrm{dist}(\boldsymbol{\xi}, \mathcal{M}) \le \delta_\Pi$ and $\delta_\Pi \le \kappa/2$. Hence, Lemma F.12 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 Lemma 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 Lemma F.7, there exists $\boldsymbol{a}_\star \in \mathcal{A}_{i_\star}^{\mathrm{anc}}$ such that

$$\|\boldsymbol{a}_\star - \boldsymbol{u}_{i_\star}^{\Pi}(\boldsymbol{x}, t)\|_2 \le \frac{1}{2} r_{i_\star}^{\mathrm{GN}}. \tag{163}$$

Let

$$L_z^{\max} := \max_{1 \le i \le C_{\mathcal{M}}} L_i^{(+)}.$$

Choose

$$\varepsilon_y^{\mathrm{glob}} := \min\Big\{\frac{\varepsilon_\Pi}{8}, \sqrt{\frac{1}{32 C_{\mathrm{glob},\Pi}^{(+)}}}\varepsilon_\Pi\Big\}, \qquad \varepsilon_u^{\mathrm{glob}} := \frac{\varepsilon_y^{\mathrm{glob}}}{\sqrt{d}\, L_z^{\max}}. \tag{164}$$

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 Lemma F.10, with localization parameter ε and accuracy $\varepsilon_u^{\mathrm{glob}}$. Define the exact manifold-valued candidate

$$\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) := \boldsymbol{z}_i\Big(\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a},\varepsilon_u^{\mathrm{glob}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t)\Big). \tag{165}$$

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 $(i_\star, \boldsymbol{a}_\star)$, (163) and Lemma 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

$$\begin{aligned} \|\boldsymbol{y}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 &\le L_{i_\star}^{(+)} \Big\|\widetilde{\boldsymbol{u}}_{i_\star,\boldsymbol{a}_\star,\varepsilon_u^{\mathrm{glob}}}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_{i_\star}^{\Pi}(\boldsymbol{x}, t)\Big\|_2 \\ &\le L_z^{\max} \sqrt{d}\, \varepsilon_u^{\mathrm{glob}} = \varepsilon_y^{\mathrm{glob}}. \end{aligned} \tag{166}$$

Define

$$\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x},t) := \|\boldsymbol{\xi}(\boldsymbol{x},t) - \boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x},t)\|_2^2. \tag{167}$$

By the upper bound in (158),

$$\begin{aligned}\mathcal{D}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x},t) - \mathcal{D}_\star &\le C^{(+)}_{\mathrm{glob},\Pi}\|\boldsymbol{y}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \\ &\le C^{(+)}_{\mathrm{glob},\Pi}(\varepsilon_y^{\mathrm{glob}})^2 \le \frac{1}{32}\varepsilon_\Pi^2.\end{aligned} \tag{168}$$

Step 2: Approximate candidate outputs and objectives. For each $(i,\boldsymbol{a}) \in \mathcal{J}$, construct a ReLU network $\widetilde{\boldsymbol{y}}^{\mathrm{out}}_{i,\boldsymbol{a}}$ satisfying

$$\left\|\widetilde{\boldsymbol{y}}^{\mathrm{out}}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x},t)\right\|_2 \le \frac{\varepsilon_\Pi}{8}. \tag{169}$$

This is obtained by composing $\widetilde{\boldsymbol{u}}^{\Pi,\varepsilon}_{i,\boldsymbol{a},\varepsilon_u^{\mathrm{glob}}}$ with a $\mathcal{H}^{q_{\mathrm{geom}}}$ 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

$$\left|\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x},t)\right| \le \varepsilon_D, \qquad \varepsilon_D := \frac{1}{32}\varepsilon_\Pi^2. \tag{170}$$

Indeed, approximate $\boldsymbol{\xi}(\boldsymbol{x},t)$ by Lemma I.7, approximate $\boldsymbol{z}_i$ on $\square_i$ by Theorem I.1, then use Lemma 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 (162).

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 representable by a ReLU network using

$$\min\{a,b\} = a - \mathsf{ReLU}(a-b).$$

Let

$$\varepsilon_\Pi^{\mathrm{gate}} := \frac{1}{32}\varepsilon_\Pi^2.$$

For each $(i,\boldsymbol{a}) \in \mathcal{J}$, define

$$\omega_{i,\boldsymbol{a}}(\boldsymbol{x},t) := \mathsf{ReLU}\Big(\varepsilon_\Pi^{\mathrm{gate}} - \big(\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \widetilde{\mathcal{D}}_{\min}(\boldsymbol{x},t)\big)\Big), \tag{171}$$

and

$$W_{\mathrm{glob}}(\boldsymbol{x},t) := \sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t).$$

At least one index attains $\widetilde{\mathcal{D}}_{\min}$, and its weight equals $\varepsilon_\Pi^{\mathrm{gate}}$. Hence

$$\varepsilon_\Pi^{\mathrm{gate}} \le W_{\mathrm{glob}}(\boldsymbol{x},t) \le N_{\mathrm{cand}}\varepsilon_\Pi^{\mathrm{gate}}. \tag{172}$$

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

$$\widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x},t) < \widetilde{\mathcal{D}}_{\min}(\boldsymbol{x},t) + \varepsilon_\Pi^{\mathrm{gate}}.$$

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 (168),

$$\begin{aligned}\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x},t) &\le \widetilde{\mathcal{D}}_{i,\boldsymbol{a}}(\boldsymbol{x},t) + \varepsilon_D \\ &< \widetilde{\mathcal{D}}_{\min}(\boldsymbol{x},t) + \varepsilon_\Pi^{\mathrm{gate}} + \varepsilon_D \\ &\le \mathcal{D}_{i_\star,\boldsymbol{a}_\star}(\boldsymbol{x},t) + \varepsilon_\Pi^{\mathrm{gate}} + 2\varepsilon_D \\ &\le \mathcal{D}_\star + \varepsilon_\Pi^{\mathrm{gate}} + 3\varepsilon_D.\end{aligned} \tag{173}$$

By the definitions of $\varepsilon_\Pi^{\mathrm{gate}}$ and ε_D,

$$\varepsilon_\Pi^{\mathrm{gate}} + 3\varepsilon_D = \varepsilon_\Pi^2\Big(\frac{1}{32} + \frac{3}{32}\Big) = \frac{1}{2}\Big(\frac{\varepsilon_\Pi}{2}\Big)^2.$$

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 (158),

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2^2 \le 2(\mathcal{D}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \mathcal{D}_\star).$$

Therefore

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\|_2 < \frac{\varepsilon_\Pi}{2} \qquad \text{whenever } \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t) > 0. \tag{174}$$

Step 4: Normalize and average. Define the ideal gated output

$$\Pi_{\mathrm{gate}}(\boldsymbol{x},t) := \frac{\sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t)\widetilde{\boldsymbol{y}}_{i,\boldsymbol{a}}^{\mathrm{out}}(\boldsymbol{x},t)}{W_{\mathrm{glob}}(\boldsymbol{x},t)}.$$

The normalized weights are nonnegative and sum to one. Combining (169) and (174), for every positive-weight candidate,

$$\big\|\widetilde{\boldsymbol{y}}_{i,\boldsymbol{a}}^{\mathrm{out}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi})\big\|_2 \le \frac{\varepsilon_\Pi}{8} + \frac{\varepsilon_\Pi}{2} = \frac{5}{8}\varepsilon_\Pi.$$

Hence

$$\|\Pi_{\mathrm{gate}}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x},t))\|_2 \le \frac{5}{8}\varepsilon_\Pi. \tag{175}$$

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

$$W_{\mathrm{glob}}(\boldsymbol{x},t) = \sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t)$$

is represented exactly by one affine layer. Thus, forming W_{glob} introduces no additional approximation error. By (172), W_{glob} lies in $[\varepsilon_\Pi^{\mathrm{gate}}, N_{\mathrm{cand}}\varepsilon_\Pi^{\mathrm{gate}}]$.

Apply Lemma I.4 to approximate $1/W_{\mathrm{glob}}$ on this interval, and then use Lemma I.3 to multiply the reciprocal by each numerator coordinate

$$N_k(\boldsymbol{x},t) := \sum_{(i,\boldsymbol{a})\in\mathcal{J}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x},t)\widetilde{y}_{i,\boldsymbol{a},k}^{\mathrm{out}}(\boldsymbol{x},t), \qquad k = 1,\dots,D.$$

Since N_{cand} is finite, all candidate outputs are uniformly bounded, and $\varepsilon_\Pi^{\mathrm{gate}} \asymp \varepsilon_\Pi^2$, 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

$$\|\widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x},t) - \Pi_{\mathrm{gate}}(\boldsymbol{x},t)\|_2 \le \frac{3}{8}\varepsilon_\Pi.$$

Combining this with (175) gives

$$\|\widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x},t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x},t))\|_2 \le \varepsilon_\Pi.$$

This proves the Euclidean bound (161), and hence also its coordinatewise version.

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 $(C_{\mathrm{proj}}D\varepsilon_\Pi^{-2})^{\frac{d}{q_{\mathrm{geom}}}}$ up to logarithmic factors. The anchored local coordinate networks have scale $(CD\varepsilon_\Pi^{-1})^{d/(q_{\mathrm{geom}})}$, which is dominated by this objective-approximation scale because $\varepsilon_\Pi \le 1$. The finite ReLU minimum, ReLU gates, affine summation layers, reciprocal network, and multiplication networks add only finite atlas factors and logarithmic factors. Indeed, $W_{\mathrm{glob}} \ge \varepsilon_\Pi^{\mathrm{gate}} \asymp \varepsilon_\Pi^2$, so approximating $1/W_{\mathrm{glob}}$ only contributes logarithmic factors of order $\log(C_{\mathrm{proj}}D\varepsilon_\Pi^{-2})$. Stacking over the N_{cand} chart-anchor candidates gives the displayed $A_{\mathcal{M}}$ factors in (162). This completes the proof.

□

Corollary F.14 (Active chart coordinates from the projection-center network)**.** *Fix a chart i and a localization parameter $0 < \varepsilon < 1$ such that (91) holds for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$. For every $\varepsilon_u \in (0, 1]$, there exists a ReLU network*

$$\widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon} : \mathbb{R}^D \times [\underline{t}, \bar{t}] \to \square_i$$

such that

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)}} \left\| \widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \right\|_\infty \le \varepsilon_u. \tag{176}$$

Moreover, for a finite chart- and geometry-dependent constant $C_{i,\Pi}^{\mathrm{coord}} \ge 1$,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3\!\big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big), \\ \|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta) \big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big)^{d/(q_{\mathrm{geom}})} + A_{\mathcal{M}} D \log^3\!\big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big), \\ S &\lesssim A_{\mathcal{M}} d D (d + q_{\mathrm{geom}})^{d+3} \big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big)^{d/(q_{\mathrm{geom}})} \log^2\!\big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big) \\ &\quad + A_{\mathcal{M}} d D \log^5\!\big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big), \\ \log B &\lesssim \log^2\!\big(C_{i,\Pi}^{\mathrm{coord}} D \varepsilon_u^{-2}\big). \end{aligned} \tag{177}$$

Proof. Extend the projection chart affine formula in (17) to all of $\mathbb{R}^D$ by

$$\phi_i^{\mathrm{amb}}(\boldsymbol{y}) := \eta_i\big(\boldsymbol{V}_i^\top(\boldsymbol{y} - \boldsymbol{y}_i) + \boldsymbol{c}_i\big).$$

Since $\boldsymbol{V}_i$ has orthonormal columns and $\eta_i \le 1$, this map satisfies

$$\|\phi_i^{\mathrm{amb}}(\boldsymbol{y}) - \phi_i^{\mathrm{amb}}(\boldsymbol{y}')\|_\infty \le \|\boldsymbol{y} - \boldsymbol{y}'\|_2.$$

Set

$$\varepsilon_\Pi := \varepsilon_u / 2.$$

Apply Theorem F.13 with accuracy ε_Π. On $\mathcal{K}_i(\varepsilon) \subset \mathcal{K}_{\mathrm{act}}(\varepsilon)$ it gives

$$\big\|\widetilde{\Pi}_{\varepsilon_\Pi}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}(\boldsymbol{x}, t))\big\|_2 \le \varepsilon_\Pi.$$

Define the raw coordinate network

$$\widehat{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon} := \phi_i^{\mathrm{amb}} \circ \widetilde{\Pi}_{\varepsilon_\Pi}$$

and then clip its output coordinatewise:

$$\widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon} := \Pi_i^{(\mathrm{box})}\big(\widehat{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}\big).$$

The affine map is exact network arithmetic and the clipping map is exactly representable by a ReLU network. By Lemma F.4, $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \in K_i^\circ \subset \square_i$ on $\mathcal{K}_i(\varepsilon)$, so the clipping map fixes the target coordinate and is 1-Lipschitz. Therefore

$$\left\| \widetilde{\boldsymbol{u}}_{i,\varepsilon_u}^{\Pi,\varepsilon}(\boldsymbol{x}, t) - \boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) \right\|_\infty \le \varepsilon_\Pi \le \varepsilon_u.$$

The size bounds follow from (162) with $\varepsilon_\Pi^{-2} = 4\varepsilon_u^{-2}$. The exact affine map and coordinatewise clipping add $\mathcal{O}(dD)$ nonzero parameters; their weights are at most $C_i(1 + \sqrt{D})$, adding only $\mathcal{O}(\log(e + D))$ to the log-weight bound already included in (177). No ℓ_∞ operator norm of the ambient frame is absorbed into a constant. □

F.3 Projection-centered Laplace expansion

This block applies Laplace's method (Wong, 2001) in projection-centered manifold coordinates and 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 Lemma 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 (91). Let $\boldsymbol{w} \in \mathbb{R}^d$ satisfy $\boldsymbol{u}_i^{\Pi}(\boldsymbol{x}, t) + \sigma_t \boldsymbol{w} \in \square_i$. Then

$$\begin{aligned} \frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}_i^{\Pi} + \sigma_t \boldsymbol{w})}{\sigma_t^2} = \frac{\|\boldsymbol{r}_i\|_2^2}{2\sigma_t^2} + \frac{m_t^2}{2} \boldsymbol{w}^\top \boldsymbol{G}_i^{\Pi} \boldsymbol{w} + \sum_{\ell=1}^{m_\beta} \sigma_t^\ell \Psi_{i,\ell}\big(\boldsymbol{u}_i^{\Pi}, \boldsymbol{\nu}_i, \boldsymbol{w}, t\big) \\ + \sigma_t^{m_\beta+1} \mathrm{Rem}_i^{(F)}\big(\boldsymbol{u}_i^{\Pi}, \boldsymbol{\nu}_i, \boldsymbol{w}, t\big), \end{aligned} \tag{178}$$

where for $\ell \geq 1$,

$$\Psi_{i,\ell}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) := -m_t \, \boldsymbol{\nu}^\top \boldsymbol{z}_{i,\ell+1}(\boldsymbol{u}, \boldsymbol{w}) + \frac{m_t^2}{2} \sum_{\substack{m_1+m_2=\ell+2 \\ m_1,m_2 \geq 1}} \boldsymbol{z}_{i,m_1}(\boldsymbol{u}, \boldsymbol{w})^\top \boldsymbol{z}_{i,m_2}(\boldsymbol{u}, \boldsymbol{w}). \tag{179}$$

Moreover, for constants independent of $(\boldsymbol{x}, t, \boldsymbol{w})$,

$$\begin{aligned} \big|\Psi_{i,\ell}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t)\big| &\leq C_\ell (1 + \|\boldsymbol{\nu}\|_2)(1 + \|\boldsymbol{w}\|_2^{\ell+2}), \\ \big|\mathrm{Rem}_i^{(F)}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t)\big| &\leq C(1 + \|\boldsymbol{\nu}\|_2)(1 + \|\boldsymbol{w}\|_2^{2m_\beta+6}). \end{aligned} \tag{180}$$

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 $m_\beta + 3$ 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

$$\begin{aligned} \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi} + \sigma_t \boldsymbol{w}) &= \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi}) + \sum_{m=1}^{m_\beta+2} \sigma_t^m \boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}) + \sigma_t^{m_\beta+3} \mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}, t) \\ &= \Pi_{\mathcal{M}}(\boldsymbol{\xi}) + \sum_{m=1}^{m_\beta+2} \sigma_t^m \boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}) + \sigma_t^{m_\beta+3} \mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}, t). \end{aligned}$$

The remainder satisfies

$$\|\mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}, t)\|_2 \leq C(1 + \|\boldsymbol{w}\|_2^{m_\beta+3}). \tag{181}$$

Define

$$\Delta_i(\boldsymbol{w}) := \boldsymbol{z}_i(\boldsymbol{u}_i^{\Pi} + \sigma_t \boldsymbol{w}) - \Pi_{\mathcal{M}}(\boldsymbol{\xi}). \tag{182}$$

Then

$$\Delta_i(\boldsymbol{w}) = \sum_{m=1}^{m_\beta+2} \sigma_t^m \boldsymbol{z}_{i,m}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}) + \sigma_t^{m_\beta+3} \mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^{\Pi}, \boldsymbol{w}, t). \tag{183}$$

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

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}_i^{\Pi} + \sigma_t \boldsymbol{w})}{\sigma_t^2} = \frac{\|\boldsymbol{r}_i\|_2^2}{2\sigma_t^2} - \frac{m_t}{\sigma_t^2} \boldsymbol{r}_i^\top \Delta_i(\boldsymbol{w}) + \frac{m_t^2}{2\sigma_t^2} \|\Delta_i(\boldsymbol{w})\|_2^2. \tag{184}$$

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 (97),

$$\boldsymbol{r}_i^\top \boldsymbol{z}_{i,1}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}) = \boldsymbol{r}_i^\top \boldsymbol{J}_i(\boldsymbol{u}_i^\Pi)\boldsymbol{w} = \boldsymbol{w}^\top \boldsymbol{J}_i(\boldsymbol{u}_i^\Pi)^\top \boldsymbol{r}_i = 0. \tag{185}$$

Using $\boldsymbol{r}_i = \sigma_t \boldsymbol{\nu}_i$ and (183) and (185), we get

$$\begin{aligned} -\frac{m_t}{\sigma_t^2}\boldsymbol{r}_i^\top \Delta_i(\boldsymbol{w}) &= -m_t \sum_{m=2}^{m_\beta+2} \sigma_t^{m-1} \boldsymbol{\nu}_i^\top \boldsymbol{z}_{i,m}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}) - m_t \sigma_t^{m_\beta+2} \boldsymbol{\nu}_i^\top \mathrm{Rem}_i^{(z)}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}, t) \\ &= -m_t \sum_{\ell=1}^{m_\beta} \sigma_t^\ell \boldsymbol{\nu}_i^\top \boldsymbol{z}_{i,\ell+1}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}) + \sigma_t^{m_\beta+1} \mathrm{Rem}_i^{(\mathrm{cr})}(\boldsymbol{u}_i^\Pi, \boldsymbol{\nu}_i, \boldsymbol{w}, t). \end{aligned} \tag{186}$$

Here the term with $m = m_\beta + 2$ has order $\sigma_t^{m_\beta+1}$ and is included in the remainder, together with the Taylor remainder. Hence, using (181) and $\sigma_t \le 1$,

$$|\mathrm{Rem}_i^{(\mathrm{cr})}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t)| \le C(1 + \|\boldsymbol{\nu}\|_2)(1 + \|\boldsymbol{w}\|_2^{m_\beta+3}). \tag{187}$$

Step 3: Expansion of the quadratic term. Using (183),

$$\begin{aligned} \frac{m_t^2}{2\sigma_t^2}\|\Delta_i(\boldsymbol{w})\|_2^2 = \frac{m_t^2}{2} \sum_{m_1,m_2=1}^{m_\beta+2} &\sigma_t^{m_1+m_2-2} \\ &\times \boldsymbol{z}_{i,m_1}(\boldsymbol{u}_i^\Pi, \boldsymbol{w})^\top \boldsymbol{z}_{i,m_2}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}) + \sigma_t^{m_\beta+1} \mathrm{Rem}_i^{(\mathrm{quad})}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}, t). \end{aligned}$$

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 $\ell = 1, \dots, m_\beta$ gives

$$\begin{aligned} \frac{m_t^2}{2\sigma_t^2}\|\Delta_i(\boldsymbol{w})\|_2^2 = \frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}_i^\Pi \boldsymbol{w} + \frac{m_t^2}{2} \sum_{\ell=1}^{m_\beta} \sigma_t^\ell \sum_{\substack{m_1+m_2=\ell+2 \\ m_1,m_2 \ge 1}} &\boldsymbol{z}_{i,m_1}(\boldsymbol{u}_i^\Pi, \boldsymbol{w})^\top \boldsymbol{z}_{i,m_2}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}) \\ &+ \sigma_t^{m_\beta+1} \mathrm{Rem}_i^{(\mathrm{quad})}(\boldsymbol{u}_i^\Pi, \boldsymbol{w}, t). \end{aligned} \tag{188}$$

The omitted Taylor pairs have total power at least $\sigma_t^{m_\beta+1}$ after division by σ_t^2, and the terms involving $\mathrm{Rem}_i^{(z)}$ have even higher powers. Therefore

$$|\mathrm{Rem}_i^{(\mathrm{quad})}(\boldsymbol{u}, \boldsymbol{w}, t)| \le C(1 + \|\boldsymbol{w}\|_2^{2m_\beta+6}). \tag{189}$$

Step 4: Collect the coefficients. Substituting (186) and (188) into (184) gives (178), with $\Psi_{i,\ell}$ defined by (179) and

$$\mathrm{Rem}_i^{(F)} := \mathrm{Rem}_i^{(\mathrm{cr})} + \mathrm{Rem}_i^{(\mathrm{quad})}.$$

The bounds in (180) 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 (187) and (189). 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 $\boldsymbol{\nu}$. □

By the convention in Definition 2.1, every coefficient retained in the neural approximation has strictly positive regularity $\beta - q > 0$.

Proposition F.16 (Local Laplace expansion for rescaled 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*

$$A_{i,q,\boldsymbol{\lambda}}^{(R)} : K_i^\circ \times [\underline{t}, \overline{t}] \to \mathbb{R}, \qquad 0 \le q \le m_\beta, \qquad \boldsymbol{\lambda} \in \mathbb{N}^D, \quad \|\boldsymbol{\lambda}\|_1 \le q,$$

such that, for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon)$ and every $(\boldsymbol{x}, t) \in \mathcal{K}_i(\varepsilon)$,

$$R_i(\boldsymbol{x}, t) = \exp\Big(-\frac{\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big) \sum_{q=0}^{m_\beta} \sigma_t^q \sum_{\|\boldsymbol{\lambda}\|_1 \le q} A_{i,q,\boldsymbol{\lambda}}^{(R)}\big(\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t), t\big)\boldsymbol{\nu}_i(\boldsymbol{x}, t)^{\boldsymbol{\lambda}} + \overline{\mathrm{Rem}}_i^{(R)}(\boldsymbol{x}, t), \tag{190}$$

where

$$\big|\overline{\mathrm{Rem}}_i^{(R)}(\boldsymbol{x},t)\big| \le C\sigma_t^{\beta}\exp\Big(-c\frac{\|\boldsymbol{r}_i(\boldsymbol{x},t)\|_2^2}{\sigma_t^2}\Big). \tag{191}$$

Moreover,

$$A_{i,q,\boldsymbol{\lambda}}^{(R)}(\cdot,t) \in \mathcal{H}^{\beta-q}(K_i^{\circ}, C_{\mathrm{coef}}), \qquad t \in [\underline{t},\overline{t}],$$

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

$$R_i(\boldsymbol{x},t) = \int_{\Omega_{i,\boldsymbol{x},t}} \exp\Big(-F_{i,\boldsymbol{x},t}(\boldsymbol{u}^{\Pi}+\sigma_t\boldsymbol{w})/\sigma_t^2\Big) a_i^{(R)}(\boldsymbol{u}^{\Pi}+\sigma_t\boldsymbol{w})\mathrm{d}\boldsymbol{w}. \tag{192}$$

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 Lemma 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 Lemma 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 (93) and (94). Then, whenever $\|\boldsymbol{u} - \boldsymbol{u}^{\Pi}\|_2 \le r_i$,

$$\nabla_{\boldsymbol{u}}^2\Phi_i(\boldsymbol{u};\boldsymbol{\xi}) \succeq \frac{1}{4}\lambda_i^{(-)}\boldsymbol{I}_d.$$

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 $\sigma_t\|\boldsymbol{w}\|_2 \le r_i$,

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}^{\Pi}+\sigma_t\boldsymbol{w})}{\sigma_t^2} \ge \frac{\|\boldsymbol{r}\|_2^2}{2\sigma_t^2} + c_0\|\boldsymbol{w}\|_2^2 \tag{193}$$

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

$$\mathcal{C}_i^{(R)} := \Big\{(\boldsymbol{\zeta}, \boldsymbol{u}) \in \mathcal{X}_i^\circ \times S_i^{(R)} : \|\boldsymbol{u} - \boldsymbol{u}_i^\Pi(\boldsymbol{\zeta})\|_2 \geq r_i\Big\}.$$

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 $\boldsymbol{\xi}$,

$$\Phi_i(\boldsymbol{u}; \boldsymbol{\xi}) - \Phi_i(\boldsymbol{u}^\Pi; \boldsymbol{\xi}) \geq c_1$$

whenever $\boldsymbol{u} \in S_i^{(R)}$ and $\|\boldsymbol{u} - \boldsymbol{u}^\Pi\|_2 \geq r_i$. Therefore, on this part of the support,

$$\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u})}{\sigma_t^2} \geq \frac{\|\boldsymbol{r}\|_2^2}{2\sigma_t^2} + \frac{\underline{m}^2 c_1}{\sigma_t^2}. \tag{194}$$

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

$$\sigma_t R_\sigma \leq r_i \leq d_i^\square/2.$$

Since $\boldsymbol{u}^\Pi \in K_i^\circ$, this implies

$$\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w} \in \square_i \subset \phi_i(U_i), \qquad \|\boldsymbol{w}\|_2 \leq R_\sigma.$$

Thus the local part of (192) is over the full ball $\{\|\boldsymbol{w}\|_2 \leq R_\sigma\}$.

Write $R_i = I_{\mathrm{loc}} + I_{\mathrm{tail}}$, where I_{loc} is the integral over $\|\boldsymbol{w}\|_2 \leq R_\sigma$. On $R_\sigma < \|\boldsymbol{w}\|_2 \leq r_i/\sigma_t$, use (193); on $\|\boldsymbol{w}\|_2 > r_i/\sigma_t$, use (194) on the support of $a_i^{(R)}$. Since $a_i^{(R)}$ is bounded and $S_i^{(R)}$ has finite Lebesgue measure,

$$\begin{aligned}|I_{\mathrm{tail}}| &\leq Ce^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} \int_{\|\boldsymbol{w}\|_2 > R_\sigma} e^{-c_0\|\boldsymbol{w}\|_2^2} \mathrm{d}\boldsymbol{w} + C\sigma_t^{-d} e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-c/\sigma_t^2} \\ &\leq Ce^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-c_2 R_\sigma^2} + C\sigma_t^{-d} e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} e^{-c/\sigma_t^2}. \end{aligned} \tag{195}$$

Choose A so large that $e^{-c_2 R_\sigma^2} \leq C\sigma_t^\beta$. Also, after decreasing the small-noise threshold if necessary, $\sigma_t^{-d} e^{-c/\sigma_t^2} \leq C\sigma_t^\beta$. Hence

$$|I_{\mathrm{tail}}| \leq C\sigma_t^\beta \exp\Big(-c\frac{\|\boldsymbol{r}\|_2^2}{\sigma_t^2}\Big). \tag{196}$$

Step 3: Expand the phase on the local window. On $\|\boldsymbol{w}\|_2 \leq R_\sigma$, Lemma F.15 and the m_β-convention give the truncated form

$$\begin{aligned}\frac{F_{i,\boldsymbol{x},t}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w})}{\sigma_t^2} &= \frac{\|\boldsymbol{r}\|_2^2}{2\sigma_t^2} + \frac{m_t^2}{2} \boldsymbol{w}^\top \boldsymbol{G}^\Pi \boldsymbol{w} + \sum_{\ell=1}^{m_\beta} \sigma_t^\ell \Psi_{i,\ell}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) \\ &\quad + \sigma_t^\beta \widehat{\mathrm{Rem}}_i^{(F)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t). \end{aligned} \tag{197}$$

Indeed, the remainder from Lemma F.15 has order $\sigma_t^{m_\beta+1}$. Since $m_\beta < \beta \leq m_\beta + 1$ and $\sigma_t \leq 1$, this is bounded by σ_t^β times the same polynomial; when $\beta \in \mathbb{N}_+$, the two powers coincide. Thus

$$|\widehat{\mathrm{Rem}}_i^{(F)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t)| \leq C(1 + \|\boldsymbol{\nu}\|_2)(1 + \|\boldsymbol{w}\|_2^{2m_\beta+6})$$

on the local window. Set

$$Z_\sigma := -\sum_{\ell=1}^{m_\beta} \sigma_t^\ell \Psi_{i,\ell}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) - \sigma_t^\beta \widehat{\mathrm{Rem}}_i^{(F)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t).$$

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 $\mathcal{I}_{\mathrm{sm}}(\varepsilon)$, $\log \varepsilon^{-1} \lesssim \log(e/\sigma_t)$. Combining this with $\|\boldsymbol{w}\|_2 \le R_\sigma$ and (180), 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 m_β, gives

$$e^{Z_\sigma} = \sum_{\ell=0}^{m_\beta} \sigma_t^\ell \mathcal{E}_{i,\ell}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \mathrm{Rem}_i^{(\exp)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t),$$

where $\mathcal{E}_{i,0} = 1$, and for $1 \le \ell \le m_\beta$,

$$\mathcal{E}_{i,\ell}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) := \sum_{r=1}^{\ell} \frac{(-1)^r}{r!} \sum_{\substack{\ell_1 + \cdots + \ell_r = \ell \\ \ell_j \ge 1}} \Psi_{i,\ell_1}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) \cdots \Psi_{i,\ell_r}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t).$$

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

$$a_i^{(R)}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}) = \sum_{m=0}^{m_\beta} \sigma_t^m \mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{w}) + \sigma_t^\beta \mathrm{Rem}_i^{(\mathrm{amp})}(\boldsymbol{u}^\Pi, \boldsymbol{w}),$$

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

$$|\mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}, \boldsymbol{w})| \le C(1 + \|\boldsymbol{w}\|_2^m), \qquad |\mathrm{Rem}_i^{(\mathrm{amp})}(\boldsymbol{u}, \boldsymbol{w})| \le C(1 + \|\boldsymbol{w}\|_2^\beta).$$

This is Taylor's theorem through order m_β with the $\mathcal{H}^\beta$ remainder. When $\beta \in \mathbb{N}_+$, the convention $m_\beta = \beta - 1$ makes this the standard Lipschitz top-derivative remainder of order β. Multiplying gives

$$e^{Z_\sigma} a_i^{(R)}(\boldsymbol{u}^\Pi + \sigma_t \boldsymbol{w}) = \sum_{q=0}^{m_\beta} \sigma_t^q \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \mathrm{Rem}_i^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t),$$

where

$$\mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t) := \sum_{\substack{m+\ell=q \\ 0 \le m, \ell \le m_\beta}} \mathcal{A}_{i,m}^{(R)}(\boldsymbol{u}, \boldsymbol{w}) \mathcal{E}_{i,\ell}(\boldsymbol{u}, \boldsymbol{\nu}, \boldsymbol{w}, t).$$

All omitted products contain either an explicit amplitude or exponential remainder, or have total algebraic order at least $m_\beta + 1$. Since $m_\beta + 1 \ge \beta$, 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

$$I_{\mathrm{loc}} = e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)} \sum_{q=0}^{m_\beta} \sigma_t^q \int_{\|\boldsymbol{w}\|_2 \le R_\sigma} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}^\Pi \boldsymbol{w}} \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) \mathrm{d}\boldsymbol{w} + E_{\mathrm{loc}}.$$

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 $0 \le q \le m_\beta$, 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,

$$\begin{aligned} &e^{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)}\sigma_t^q \left| \int_{\|\boldsymbol{w}\|_2 > R_\sigma} e^{-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}^{\Pi}\boldsymbol{w}} \mathcal{U}_{i,q}^{(R)}(\boldsymbol{u}^{\Pi},\boldsymbol{\nu},\boldsymbol{w},t)\mathrm{d}\boldsymbol{w} \right| \\ &\quad \le Ce^{-c\|\boldsymbol{\nu}\|_2^2} e^{-c_3 R_\sigma^2/2} \le C\sigma_t^{\beta} e^{-c\|\boldsymbol{\nu}\|_2^2}. \end{aligned}$$

Combining this extension error with the bounds for E_{loc} and I_{tail} proves (190) and (191), 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 $t \in [\underline{t},\bar{t}]$. 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

$$A_{i,q,\boldsymbol{\lambda}}^{(R)}(\cdot,t) \in \mathcal{H}^{\beta-q}(K_i^\circ, C_{\mathrm{coef}}), \qquad t \in [\underline{t},\bar{t}].$$

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 $C_{\mathrm{Lap}}^{\max} \ge 1$ to dominate the constants in the Laplace remainder bound (191), 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 Proposition F.16 holds whenever $0 < \sigma_t \le \bar{\sigma}_{\mathrm{Lap}}$: in particular $\sigma_t R_\sigma \le r_i$, the compact-away tail satisfies $\sigma_t^{-d} e^{-c/\sigma_t^2} \lesssim \sigma_t^{\beta}$, and the exponential correction obeys $|Z_\sigma| \le 1/2$, 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 Definition D.1 are fixed so that

$$C_{\mathrm{sm}} \le \bar{\sigma}_{\mathrm{Lap}}, \qquad C_{\mathrm{Lap}}^{\max} C_{\mathrm{sm}}^{\beta} \le \frac{1}{6}, \qquad 4c_{\mathrm{lg}} \le C_{\mathrm{sm}}. \tag{198}$$

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, active-chart gates with explicit validity regions, 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 $0 < \varepsilon \le 1$ such that (91) 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.13. Then there exist ReLU networks

$$\widetilde{\boldsymbol{r}}_{i,\varepsilon_{\mathrm{res}}}, \widetilde{\boldsymbol{\nu}}_{i,\varepsilon_{\mathrm{res}}} : \mathbb{R}^D \times [\underline{t}, \overline{t}] \to \mathbb{R}^D$$

such that

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)}} \max\Big\{ \big\|\widetilde{\boldsymbol{r}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x}, t) - \boldsymbol{r}_i(\boldsymbol{x}, t)\big\|_\infty, \big\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x}, t) - \boldsymbol{\nu}_i(\boldsymbol{x}, t)\big\|_\infty \Big\} \le \varepsilon_{\mathrm{res}}.$$

Moreover,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big), \\ \|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta) \big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big)^{d/(q_{\mathrm{geom}})} + A_{\mathcal{M}} D \log^3\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big), \\ S &\lesssim A_{\mathcal{M}} d D (d + q_{\mathrm{geom}})^{d+3} \big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big)^{d/(q_{\mathrm{geom}})} \log^2\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big) \\ &\quad + A_{\mathcal{M}} d D \log^5\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big), \\ \log B &\lesssim \log^2\big(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2}\big). \end{aligned} \tag{199}$$

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

$$\begin{aligned} \|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2 &= m_t \mathrm{dist}\big(\boldsymbol{\xi}(\boldsymbol{x}, t), \mathcal{M}\big) \\ &\le m_t \|\boldsymbol{\xi}(\boldsymbol{x}, t) - \boldsymbol{y}\|_2 = \|\boldsymbol{x} - m_t \boldsymbol{y}\|_2 \\ &\le c_\star \sigma_t \sqrt{\log(e/\varepsilon)}. \end{aligned} \tag{200}$$

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.13, 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.13, we use the smoothness q_{geom} of the geometric chart maps, hence its non-logarithmic factor is controlled by

$$\big(C_{\mathrm{proj}} D \varepsilon_{\mathrm{cen}}^{-2}\big)^{d/(q_{\mathrm{geom}})}.$$

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 Lemma 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 Lemma 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

$$|\phi_{1/\sigma}(t) - \sigma_t^{-1}| \le \varepsilon_\sigma, \qquad \varepsilon_\sigma := \frac{\varepsilon_{\mathrm{res}}}{8C_{\nu,\varepsilon}}.$$

For each j, use Lemma 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

$$\big\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_{\mathrm{res}}}(\boldsymbol{x},t) - \boldsymbol{\nu}_i(\boldsymbol{x},t)\big\|_\infty \le \varepsilon_{\mathrm{res}}.$$

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 (162) 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 $\underline{\sigma_t}^{-1}$, all bounded by powers of $\log(C_{\mathrm{res}} D \varepsilon_{\mathrm{res}}^{-2})$. Summing the components gives (199). □

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 (91) 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 Lemma F.17. For every $\varepsilon_G \in (0,1]$, define

$$\varepsilon_\nu := \frac{\varepsilon_G}{12 D^{1/2} C_{\nu,\varepsilon}}, \qquad C_{\mathrm{gau}} := e \vee \Big(144 D^2 C_{\mathrm{res}} C_{\nu,\varepsilon}^2\Big).$$

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,

$$\begin{aligned} L &\lesssim \log(e(d \vee \beta)) \log^3(C_{\mathrm{gau}} \varepsilon_G^{-2}), \\ \|\boldsymbol{W}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta) \big(C_{\mathrm{gau}} \varepsilon_G^{-2}\big)^{d/(q_{\mathrm{geom}})} + A_{\mathcal{M}} D \log^3(C_{\mathrm{gau}} \varepsilon_G^{-2}), \\ S &\lesssim A_{\mathcal{M}} d D (d + q_{\mathrm{geom}})^{d+3} \big(C_{\mathrm{gau}} \varepsilon_G^{-2}\big)^{d/(q_{\mathrm{geom}})} \log^2(C_{\mathrm{gau}} \varepsilon_G^{-2}) \\ &\quad + A_{\mathcal{M}} d D \log^5(C_{\mathrm{gau}} \varepsilon_G^{-2}), \\ \log B &\lesssim \log^2(C_{\mathrm{gau}} \varepsilon_G^{-2}). \end{aligned} \tag{201}$$

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 Lemma 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 (200),

$$\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le c_\star \sqrt{\log(e/\varepsilon)} \le C_{\nu,\varepsilon}.$$

Step 2: Approximate the exponent. Let

$$\varepsilon_{\mathrm{sq}} := \frac{\varepsilon_G}{3D}.$$

For each $\ell = 1, \dots, D$, apply Lemma 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

$$\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu}(\boldsymbol{x},t) - \boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le D^{1/2}\varepsilon_\nu \le \frac{\varepsilon_G}{12 C_{\nu,\varepsilon}}.$$

Using

$$\left|\|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu}\|_2^2 - \|\boldsymbol{\nu}_i\|_2^2\right| \le \|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu} - \boldsymbol{\nu}_i\|_2 \big(2\|\boldsymbol{\nu}_i\|_2 + \|\widetilde{\boldsymbol{\nu}}_{i,\varepsilon_\nu} - \boldsymbol{\nu}_i\|_2\big),$$

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 Lemma 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 (199) 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 (201) 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 $i \in \{1, \dots, C_{\mathcal{M}}\}$, $R \in \mathcal{R}$, $0 \le q \le m_\beta$, 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 (21) and (22). 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

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}, t) := \sum_{e \in \mathcal{E}_{i,q,\boldsymbol{\lambda}}^{(R)}} m_t^{r_e} B_{i,q,\boldsymbol{\lambda},e}^{(R)}(\boldsymbol{u}), \qquad (\boldsymbol{u}, t) \in \square_i^A \times [\underline{t}, \overline{t}], \tag{202}$$

is an extension of $A_{i,q,\boldsymbol{\lambda}}^{(R)}$:

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}, t) = A_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}, t), \qquad (\boldsymbol{u}, t) \in K_i^\circ \times [\underline{t}, \overline{t}]. \tag{203}$$

With $S_i^{(R)} := \mathsf{supp} a_i^{(R)}$, the extension can be chosen so that

$$\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}, t) = 0, \qquad (\boldsymbol{u}, t) \in \left(\square_i^A \setminus S_i^{(R)}\right) \times [\underline{t}, \overline{t}]. \tag{204}$$

Moreover,

$$\sup_{t \in [\underline{t}, \overline{t}]} \left\| \overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\cdot, t) \right\|_{\mathcal{H}^{\beta-q}(\square_i^A)} \le C_{A,q}. \tag{205}$$

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 (179), 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

$$m_t^a H(\boldsymbol{u}) \partial^\alpha a_i^{(R)}(\boldsymbol{u}) \int_{\mathbb{R}^d} e^{-\frac{m_t^2}{2} \boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u}) \boldsymbol{w}} \boldsymbol{w}^\rho \mathrm{d}\boldsymbol{w}, \tag{206}$$

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 (206) gives an explicit finite formula for the coefficient on K_i°. At retained order q, each phase factor of order $h \ge 1$ has tangential degree at most $h + 2$ and schedule degree at most two, while an amplitude derivative of order h has tangential degree h. Consequently $0 \le a \le 2q$ and $|\rho| \le 3q$ in (206), so every nonzero schedule term may be chosen with

$$r_e = a - d - |\rho|, \qquad |r_e| \le d + 3q. \tag{207}$$

In particular, the zeroth-order coefficient contains the factor m_t^{-d}. 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 $B^{(R)}_{i,q,\boldsymbol{\lambda},e}$ 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 $\mathcal{E}^{(R)}_{i,q,\boldsymbol{\lambda}}$ to be a singleton with $r_e = 0$ and $B^{(R)}_{i,q,\boldsymbol{\lambda},e} \equiv 0$.

Since $a_i^{(R)} \in \mathcal{H}^\beta$ on the coordinate domain and the geometric factors are smooth on $\square_i$, each directly defined $B^{(R)}_{i,q,\boldsymbol{\lambda},e}$ belongs to $\mathcal{H}^{\beta-q}(\square_i^A)$. The formula agrees with the original coefficient on K_i°, giving (203). Moreover, if $\boldsymbol{u} \notin S_i^{(R)} = \mathrm{supp} a_i^{(R)}$, 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 (204). Taking $C_{A,q}$ to dominate the finitely many $\mathcal{H}^{\beta-q}$-norms of the directly defined factors proves (205) 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 Lemma F.19, and assume $\beta - q > 0$. Let $\square_i^A$, $\mathcal{E}^{(R)}_{i,q,\boldsymbol{\lambda}}$, r_e, $B^{(R)}_{i,q,\boldsymbol{\lambda},e}$, and $\overline{A}^{(R)}_{i,q,\boldsymbol{\lambda}}$ be the objects introduced in Lemma F.19. For every $0 < \varepsilon_A \le 1/2$, there is a ReLU network*

$$\widetilde{A}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_A} : \mathbb{R}^d \times [\underline{t}, \overline{t}] \to \mathbb{R}$$

such that

$$\sup_{(\boldsymbol{u},t)\in \square_i^A \times [\underline{t},\overline{t}]} \left| \widetilde{A}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_A}(\boldsymbol{u},t) - \overline{A}^{(R)}_{i,q,\boldsymbol{\lambda}}(\boldsymbol{u},t) \right| \le \varepsilon_A. \tag{208}$$

Moreover, for a finite constant $C_{A,\mathrm{nn}} \ge 1$,

$$\begin{aligned} L, \log B &\lesssim \log(e(d \vee (\beta - q))) \log^2(C_{A,\mathrm{nn}} \varepsilon_A^{-1}), \\ \|\boldsymbol{W}\|_\infty &\lesssim (d \vee (\beta - q))(C_{A,\mathrm{nn}} \varepsilon_A^{-1})^{d/(\beta-q)} + \log^3(C_{A,\mathrm{nn}} \varepsilon_A^{-1}), \\ S &\lesssim (d + \beta - q + 2)^{d+4} (C_{A,\mathrm{nn}} \varepsilon_A^{-1})^{d/(\beta-q)} \log(C_{A,\mathrm{nn}} \varepsilon_A^{-1}) + \log^4(C_{A,\mathrm{nn}} \varepsilon_A^{-1}). \end{aligned} \tag{209}$$

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

$$\varepsilon_B := \frac{\varepsilon_A}{8 N_A M_A}.$$

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 Lemma I.2 and form the fixed positive power by Lemma I.3. If $r_e < 0$, approximate $1/m_t$ by Lemma I.6 and form the fixed positive power $(1/m_t)^{-r_e}$ by Lemma 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 (208). Adding the intrinsic Hölder-approximation cost, the logarithmic schedule cost, and the fixed product/summation overhead gives (209). □

Corollary F.21 (Weighted coefficient approximation on the active branch)**.** *Fix $i, R, q, \boldsymbol{\lambda}$ as in Lemma F.19, and assume $\beta - q > 0$. Let $0 < \varepsilon \le 1/2$ be a localization parameter for which (91) holds on $\mathcal{I}_{\mathrm{sm}}(\varepsilon)$. Assume the small-noise branch satisfies*

$$\sigma_t \le \sigma_\star(\varepsilon) \qquad \textit{for every } t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon),$$

For every $0 < \varepsilon_{\mathrm{term}} \le 1/2$, define

$$\varepsilon_{A,q} := \frac{1}{2} \wedge c_{A,q}\, \varepsilon_{\mathrm{term}}\, \sigma_\star(\varepsilon)^{-q} \big(\log(e/\varepsilon)\big)^{-q/2}, \tag{210}$$

with $c_{A,q} > 0$ sufficiently small. Then there is a ReLU network

$$\widetilde{A}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}} : \mathbb{R}^D \times [\underline{t}, \bar{t}] \to \mathbb{R}$$

with

$$\widetilde{A}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}} \in \mathcal{NN}(L_{\mathcal{A}}, \boldsymbol{W}_{\mathcal{A}}, S_{\mathcal{A}}, B_{\mathcal{A}})$$

such that

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)}} \Big| \widetilde{A}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}}(\boldsymbol{x},t) - A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big) \Big| \le \varepsilon_{A,q}. \tag{211}$$

Consequently,

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon) \\ (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon)}} \sigma_t^q |\boldsymbol{\nu}_i(\boldsymbol{x},t)^{\boldsymbol{\lambda}}| \Big| \widetilde{A}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon_{A,q}}(\boldsymbol{x},t) - A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big) \Big| \le \varepsilon_{\mathrm{term}}. \tag{212}$$

Corollary F.22 (Capacity of the weighted coefficient approximation)**.** *Work in the setting of Corollary F.21, with $\varepsilon_{A,q}$ defined by (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*

$$\begin{aligned}
L_{\mathcal{A}} &\lesssim \log(e(d \vee \beta)) \log^3\big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big) + \log(e(d \vee (\beta - q))) \log^2\big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big), \\
\|\boldsymbol{W}_{\mathcal{A}}\|_\infty &\lesssim A_{\mathcal{M}} d D (d \vee \beta) \big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big)^{d/(q_{\mathrm{geom}})} + A_{\mathcal{M}} D \log^3\big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big) \\
&\quad + (d \vee (\beta - q)) \big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big)^{d/(\beta - q)} + \log^3\big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big), \\
S_{\mathcal{A}} &\lesssim A_{\mathcal{M}} d D (d + q_{\mathrm{geom}})^{d+3} \big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big)^{d/(q_{\mathrm{geom}})} \log^2\big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big) \\
&\quad + A_{\mathcal{M}} d D \log^5\big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big) \\
&\quad + (d + \beta - q + 2)^{d+4} \big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big)^{d/(\beta - q)} \log\big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big) + \log^4\big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big), \\
\log B_{\mathcal{A}} &\lesssim \log^2\big(C_{\mathcal{A},q} D \varepsilon_{A,q}^{-2/\alpha_q}\big) + \log^2\big(C_{\mathcal{A},q} \varepsilon_{A,q}^{-1}\big).
\end{aligned} \tag{213}$$

In particular, for fixed intrinsic, geometric, schedule, and coefficient constants, the total width and sparsity are bounded, up to logarithmic factors, by

$$\varepsilon_{A,q}^{-d/(\beta - q)} + A_{\mathcal{M}} D^{1 + d/(q_{\mathrm{geom}})} \varepsilon_{A,q}^{-2d/(\alpha_q q_{\mathrm{geom}})}. \tag{214}$$

If $\sigma_\star(\varepsilon) \asymp \varepsilon^{1/\beta}$ and $\varepsilon_{\mathrm{term}} \asymp \varepsilon$, then the two ε-dependent powers in (214) 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 and the displayed polynomial ambient factor.

Proof of Corollaries F.21 and F.22. Use the fixed buffered box K_i° and the intermediate box $\square_i^A$ from Lemma F.19. Let

$$\alpha_q := (\beta - q) \wedge 1.$$

Apply Corollary F.14 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 Lemma F.20 with accuracy $\varepsilon_{A,q}/2$, and compose the resulting coefficient network with this coordinate network. Since

$$\overline{A}^{(R)}_{i,q,\boldsymbol{\lambda}}(\cdot,t) \in \mathcal{H}^{\beta - q}(\square_i^A, C_{A,q})$$

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 (203) 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 (211).

On $\mathcal{K}_i(\varepsilon)$, (200) 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 (212).

The coefficient network contributes the terms in (209) with accuracy $\varepsilon_{A,q}/2$. The coordinate network is obtained from Corollary F.14 with accuracy $\varepsilon_u = c\varepsilon_{A,q}^{1/\alpha_q}$; substituting this value into (177) gives the terms with scale $D\varepsilon_{A,q}^{-2/\alpha_q}$. The final composition and the finite summations change the displayed quantities only by fixed constants. After increasing $C_{\mathcal{A},q}$, this proves (213). In particular, including its ambient output and fixed-atlas factors, the coordinate network's non-logarithmic contribution is bounded by

$$A_{\mathcal{M}}D(D\varepsilon_u^{-2})^{d/(q_{\mathrm{geom}})} \lesssim A_{\mathcal{M}}D^{1+d/(q_{\mathrm{geom}})}\varepsilon_{A,q}^{-2d/(\alpha_q q_{\mathrm{geom}})}.$$

Together with the coefficient contribution $\varepsilon_{A,q}^{-d/(\beta-q)}$, this proves (214).

Finally, since $\sigma_\star(\varepsilon) \asymp \varepsilon^{1/\beta}$ and $\varepsilon_{\mathrm{term}} \asymp \varepsilon$, for sufficiently small ε, the minimum in (210) 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 (214) gives

$$\varepsilon_{A,q}^{-d/(\beta-q)} \lesssim \varepsilon^{-d/\beta}\big(\log(e/\varepsilon)\big)^{qd/(2(\beta-q))}.$$

For the geometric term, substitution gives

$$\varepsilon_{A,q}^{-2d/(\alpha_q q_{\mathrm{geom}})} \lesssim \varepsilon^{-2d(\beta-q)/(\beta\alpha_q q_{\mathrm{geom}})}\big(\log(e/\varepsilon)\big)^{qd/(\alpha_q q_{\mathrm{geom}})}.$$

Since $(\beta-q)/\alpha_q = \max\{\beta-q,1\} \le \beta \vee 1$ and $q_{\mathrm{geom}} \ge 2(\beta\vee 1)$, its accuracy exponent is at most d/β. This proves the displayed polylogarithmic rate. □

Corollary F.23 (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

$$\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le c_a\sqrt{\log(e/\varepsilon)}$$

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 Lemmas F.4 and F.11 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 chart-coordinate network in Corollary F.14 can be chosen so that its error bound holds uniformly on $\mathcal{K}_i^{(c_a)}(\varepsilon)$.

(iii) Theorem F.13 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 Lemmas F.17 and F.18 and Corollary 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

$$\|\boldsymbol{x} - m_t \boldsymbol{y}\|_2 \le c_a \sigma_t \sqrt{\log(e/\varepsilon)}, \qquad \boldsymbol{y} \in \mathcal{S}_i,$$

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

$$\|\boldsymbol{r}_i(\boldsymbol{x}, t)\|_2 \le c_a \sigma_t \sqrt{\log(e/\varepsilon)}, \qquad \|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \le c_a \sqrt{\log(e/\varepsilon)}.$$

Thus every place where $C_{\nu,\varepsilon}$ appears may use $2\sigma_{\underline{t}}^{-1} + c_a\sqrt{\log(e/\varepsilon)}$. For the global projection theorem, the original argument uses only the following two facts: for every active pair $(\boldsymbol{x}, t)$, there is 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.13 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.13. Composing that fixed-c_a projection-center network with the affine extension of the chart map ϕ_i and the exact box clipping layer gives the fixed-c_a version of Corollary 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 chart-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.24 (H-active component approximation)**.** *Fix $c_a > 0$, $H \ge 1$, and $0 < \varepsilon \le 1/2$. Assume $[t_0, t_1] \subset [\underline{t}, \overline{t}]$, $0 < t_0 < t_1 \le 1$, and*

$$c_a \frac{\sigma_t}{m_t}\sqrt{H} \le \eta_\Pi/2, \qquad t \in [t_0, t_1].$$

Set $\sigma_H^\star := \sup_{t\in[t_0,t_1]} \sigma_t$. For

$$A_t(H) := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le C_\varrho \sigma_t \sqrt{H}\},$$

define

$$\mathcal{K}_i^{(c_a,H)} := \Big\{(\boldsymbol{x}, t) : \boldsymbol{x} \in A_t(H),\ \exists \boldsymbol{y} \in \mathcal{S}_i \text{ with } \|\boldsymbol{x} - m_t\boldsymbol{y}\|_2 \le c_a\sigma_t\sqrt{H}\Big\}.$$

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

$$\|\boldsymbol{\nu}_i(\boldsymbol{x}, t)\|_2 \le c_a\sqrt{H}.$$

Moreover, the component approximation statements in Corollaries F.14 and F.21, Theorem F.13, and Lemmas F.17 and F.18 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 $H + \log(C\varepsilon^{-1})$. For fixed c_a, these replacements preserve the same non-logarithmic intrinsic costs and introduce only additional fixed polynomial powers of H.

Corollary F.25 (H-active weighted-coefficient tolerance)**.** *Work under the assumptions of Corollary F.24. More explicitly, for the weighted-coefficient statement, fix the same indices $i, R, q, \boldsymbol{\lambda}$ and the same term accuracy $0 < \varepsilon_{\mathrm{term}} \le 1/2$ as in Corollary 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 (212) holds on $\mathcal{K}_i^{(c_a,H)}$ with $\varepsilon_{A,q}$ replaced by $\varepsilon_{A,q}^{H}$.

Proof of Corollaries F.24 and F.25. The proof of Corollary F.23 depends on the active set only through the reach condition and the residual bound. No Laplace remainder or threshold C_{sm} is used in these component constructions. Indeed, $m_t \ge e^{-1}$, $\sigma_t \ge \sigma_{t_0} > 0$, and the active reach bound places $\boldsymbol{x}/m_t$ in a fixed buffered neighborhood of $\mathcal{S}_i$. These are precisely the range and conditioning assumptions used in the coordinate iteration, objective weights, and arithmetic networks; their lower-time dependence remains explicit. 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 (200). The chart-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 $|\boldsymbol{\nu}_i^{\boldsymbol{\lambda}}| \le C_q H^{q/2}$ when $\|\boldsymbol{\lambda}\|_1 \le q$. The definition of $\varepsilon_{A,q}^{H}$ and the bound $\sigma_t \le \sigma_H^\star$ give the same weighted error proof as Corollary 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.26 (H-active chart gate from objective weights)**.** *Fix a chart index i, a radius multiplier $c_a \ge c_\star$, $H \ge 1$, $0 < \varepsilon \le 1/2$, and a coordinate tolerance $0 < \eta \le 1$. Let $[t_0, t_1] \subset [\underline{t}, \overline{t}]$ and $0 < t_0 < t_1 \le 1$, and assume*

$$c_a \frac{\sigma_t}{m_t}\sqrt{H} \le \eta_\Pi/2, \qquad t \in [t_0, t_1].$$

Set $A_t(H) := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le C_\varrho \sigma_t \sqrt{H}\}$. Define $\mathcal{K}_i^{(c_a,H)}$ from this $A_t(H)$ as in Corollary F.24. 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

$$\Gamma_i^H(\boldsymbol{x}, t) = 1, \qquad \|\breve{\boldsymbol{u}}_i(\boldsymbol{x}, t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\|_\infty \le \eta, \qquad \|\breve{\boldsymbol{u}}_i^A(\boldsymbol{x}, t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\|_\infty \le \eta. \tag{215}$$

If $\Gamma_i^H(\boldsymbol{x}, t) > 0$, then $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t) \in U_i$ and, with $\boldsymbol{u}_i^\Pi(\boldsymbol{x}, t) := \phi_i(\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t))$,

$$\|\breve{\boldsymbol{u}}_i(\boldsymbol{x}, t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\|_\infty \le \eta. \tag{216}$$

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 $\eta^{-2d/q_{\mathrm{geom}}}$ and fixed-atlas arithmetic.

Proof. We repeat the objective-weight part of the proof of Theorem F.13, using the H-active extension from Corollary F.24, but we keep the chart identity of each candidate. Let

$$\mathcal{J} := \{(j, \boldsymbol{a}) : 1 \le j \le C_{\mathcal{M}},\ \boldsymbol{a} \in \mathcal{A}_j^{\mathrm{anc}}\}, \qquad N_{\mathrm{cand}} := |\mathcal{J}|.$$

Set

$$\delta_i^U := \mathrm{dist}_{\mathbb{R}^D}\big(\boldsymbol{z}_i(\square_i), \mathcal{M} \setminus U_i\big) > 0, \qquad C_+ := 1 + \frac{\eta_\Pi}{2\kappa} \le \frac{3}{2}, \qquad L_z^{\max} := 1 \vee \max_j L_j^{(+)},$$

with $\operatorname{dist}(A, \emptyset) = +\infty$. Set $c_{\mathrm{cert}} := (1/8) \wedge (\delta_i^U/8)$, put $\varepsilon_{\mathrm{cert}} = c_{\mathrm{cert}}\eta$, and allocate the internal coordinate accuracy

$$\varepsilon_u^{\mathrm{cert}} := \frac{\varepsilon_{\mathrm{cert}}}{8\sqrt{dC_+}\, L_z^{\max}}.$$

For each $(j, \boldsymbol{a}) \in \mathcal{J}$, form the anchored coordinate candidate $\widetilde{\boldsymbol{u}}_{j,\boldsymbol{a}}$ at accuracy $\varepsilon_u^{\mathrm{cert}}$, 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

$$|\widetilde{\mathcal{D}}_{j,\boldsymbol{a}} - \|\boldsymbol{x}/m_t - \boldsymbol{y}_{j,\boldsymbol{a}}\|_2^2| \le \gamma/64, \qquad \gamma := \varepsilon_{\mathrm{cert}}^2, \tag{217}$$

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 every such chart j, Corollary F.24 and Lemmas F.7 and F.10 give an anchor $\boldsymbol{a}_\star$ satisfying

$$\|\boldsymbol{y}_{j,\boldsymbol{a}_\star} - \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)\|_2 \le L_z^{\max}\sqrt{d}\,\varepsilon_u^{\mathrm{cert}} = \frac{\varepsilon_{\mathrm{cert}}}{8\sqrt{C_+}}.$$

Since $\operatorname{dist}(\boldsymbol{x}/m_t, \mathcal{M}) \le \eta_\Pi/2$, the upper objective-gap bound in Lemma F.12 therefore gives

$$\|\boldsymbol{x}/m_t - \boldsymbol{y}_{j,\boldsymbol{a}_\star}\|_2^2 - \operatorname{dist}(\boldsymbol{x}/m_t, \mathcal{M})^2 \le C_+\|\boldsymbol{y}_{j,\boldsymbol{a}_\star} - \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)\|_2^2 \le \gamma/64.$$

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

$$\gamma \le W_{\mathrm{glob}} \le N_{\mathrm{cand}}\gamma. \tag{218}$$

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 (218) gives

$$\frac{\Omega_i}{W_{\mathrm{glob}}} \ge \tau_0, \qquad \tau_0 := \frac{1}{2N_{\mathrm{cand}}}. \tag{219}$$

Conversely, suppose $\Omega_i > 0$. Then some chart-i candidate has positive weight. The same calculation as (173) and (174), with the tolerances in (217), implies

$$\|\boldsymbol{y}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)\|_2 \le 2\varepsilon_{\mathrm{cert}} \tag{220}$$

for every positive-weight chart-i candidate. Indeed, its exact objective gap is at most $(1+3/64)\gamma$, and the lower bound in Lemma F.12 gives squared distance at most $2(1+3/64)\gamma < 4\gamma$. Our choice of c_{cert} ensures $2\varepsilon_{\mathrm{cert}} \le \eta/4$ and, when $\delta_i^U < \infty$, $2\varepsilon_{\mathrm{cert}} \le \delta_i^U/4$. Thus (220) forces $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t) \in U_i$ whenever $\Omega_i > 0$. The affine extension $\phi_i(y) = \eta_i(\boldsymbol{V}_i^\top(y - \boldsymbol{y}_i) + \boldsymbol{c}_i)$ is globally 1-Lipschitz because $\eta_i \le 1$, so every positive-weight chart-i coordinate candidate satisfies

$$\|\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t) - \boldsymbol{u}_i^\Pi(\boldsymbol{x}, t)\|_\infty \le 2\varepsilon_{\mathrm{cert}} \le \eta/4. \tag{221}$$

Define the ideal chart coordinate average

$$\boldsymbol{u}_{i,\mathrm{id}}(\boldsymbol{x}, t) := \frac{\sum_{\boldsymbol{a}\in\mathcal{A}_i^{\mathrm{anc}}} \omega_{i,\boldsymbol{a}}(\boldsymbol{x}, t)\widetilde{\boldsymbol{u}}_{i,\boldsymbol{a}}(\boldsymbol{x}, t)}{\Omega_i(\boldsymbol{x}, t)}$$

whenever $\Omega_i > 0$. If $\Omega_i/W_{\mathrm{glob}} \ge \tau_0/32$, then (218) gives $\Omega_i \ge \tau_0\gamma/32$, and (221) gives $\|\boldsymbol{u}_{i,\mathrm{id}} - \boldsymbol{u}_i^\Pi\|_\infty \le \eta/4$. Implement the denominator as $\Omega_i \vee (\tau_0\gamma/64)$, approximate the reciprocal and

products to total coordinatewise error at most $\eta/4$, and apply the exact clipping map $\Pi_i^{(\mathrm{box})}$ as the final layer. The implemented coordinate $\breve{\boldsymbol{u}}_i$ maps into $\square_i$ and has error at most $\eta/2$, hence satisfies (216), whenever the gate below is nonzero, because the ideal weighted average belongs to the convex box $\square_i$ and is within $\eta/4$ of $\boldsymbol{u}_i^{\Pi}$. Compare the clipped arithmetic output with this ideal average, which clipping fixes; this argument also applies when the true coordinate lies just outside $\square_i$.

It remains only to construct the binary-safe chart gate. Approximate $\Omega_i/W_{\mathrm{glob}}$ with error at most $\tau_0/32$ on the interval in (218). Let $\theta_{\mathrm{cert}} : \mathbb{R} \to [0,1]$ 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 (219) 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 $\Omega_i/W_{\mathrm{glob}} \geq \tau_0/32$, so the clipped denominator in the coordinate average is inactive and the coordinate error bound 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 establishes the neural Laplace approximation 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 rescaled construction below adds a separate active-chart gate before using inactive-chart neural outputs. This separation ensures that no local chart-coordinate network is used outside the region where its approximation guarantee is available.

Proposition F.27 (Projection–Laplace chart approximation for small noise)**.** *Assume Assumptions 1 and 2. Let $0 < \varepsilon_{\mathrm{sm}} \leq 1/2$, and assume the localization condition (91), as well as the same condition with $3c_\star$ in place of $c_\star$, holds for every $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$. For every chart i and every $R \in \mathcal{R}$, there exists a ReLU network $\widetilde{R}_i^{\mathrm{sm}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ such that*

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}}) \\ (\boldsymbol{x},t) \in \mathcal{K}_i^{(3c_\star)}(\varepsilon_{\mathrm{sm}})}} \left| R_i(\boldsymbol{x},t) - \widetilde{R}_i^{\mathrm{sm}}(\boldsymbol{x},t) \right| \leq \varepsilon_{\mathrm{sm}}. \tag{222}$$

In addition, the true inactive tail satisfies

$$\sup_{\substack{t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}}) \\ \boldsymbol{x} \in A_t,\ (\boldsymbol{x},t) \notin \mathcal{K}_i(\varepsilon_{\mathrm{sm}})}} |R_i(\boldsymbol{x},t)| \leq \varepsilon_{\mathrm{sm}}/6. \tag{223}$$

Proposition F.28 (Complexity of the projection–Laplace chart network)**.** *Under the assumptions and construction of Proposition F.27, let $C_{\mathrm{size,sm}} \geq e$ be a fixed constant large enough to absorb the chart, the bounds on $m_t^{\pm 1}$, coefficient, and arithmetic constants, independently of $\sigma_{\underline{t}}$, and define*

$$\Lambda_{\mathrm{sm}} := e \vee C_{\mathrm{size,sm}} D \varepsilon_{\mathrm{sm}}^{-1} \vee \underline{t}^{-1}. \tag{224}$$

For finite exponents $\mathfrak{a}_{\mathrm{sm}}$ and $\mathfrak{p}_{\beta,d}$ depending only on $(\beta, d, q_{\mathrm{geom}})$, the same construction can be chosen so that, up to fixed geometry-dependent constants,

$$\begin{aligned} L + \log B &\lesssim \log(e(d \vee \beta)) \log^3 \Lambda_{\mathrm{sm}}, \\ \|\boldsymbol{W}\|_\infty + S &\lesssim D^{\mathfrak{a}_{\mathrm{sm}}} A_{\mathcal{M}} \left[\varepsilon_{\mathrm{sm}}^{-d/\beta} + (\sigma_{\underline{t}}^{-4} \varepsilon_{\mathrm{sm}}^{-2})^{d/q_{\mathrm{geom}}} \right] \log^{\mathfrak{p}_{\beta,d}} \Lambda_{\mathrm{sm}}. \end{aligned} \tag{225}$$

The expanded depth, width, sparsity, and log-weight bounds are recorded in Lemma F.29.

Lemma F.29 (Small-noise chart size bound)**.** *Work in the setting of Proposition F.27. For $0 \leq q \leq m_\beta$, 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 Proposition F.27 may be chosen so that the following bounds hold. More explicitly, one may take*

$$\mathfrak{a}_{\mathrm{sm}} = \left\lceil \frac{m_\beta\, d}{\beta - m_\beta} + \frac{d}{\beta \vee 1} + 3 \right\rceil.$$

The logarithmic factor $\log \Lambda_{\mathrm{sm}}$ *below includes the ambient dimension and the lower-time dependence of the schedule networks.*

$$
\begin{aligned}
L &\lesssim \log(e(d \vee \beta)) \log^3 \Lambda_{\mathrm{sm}}, \\
\|\boldsymbol{W}\|_\infty &\lesssim D^{\mathfrak{a}_{\mathrm{sm}}} \frac{(D+d+\beta)^{m_\beta+1}}{m_\beta!} \Bigg[\left(C_{\mathrm{size,sm}} \sigma_{\underline{t}}^{-4} \varepsilon_{\mathrm{sm}}^{-2}\right)^{d/(q_{\mathrm{geom}})} \\
&\qquad + \sum_{q=0}^{m_\beta} \Bigg\{ \left(C_{\mathrm{size,sm}} \varepsilon_{\mathrm{sm}}^{-2(1-q/\beta)/\alpha_q} \log^{q/\alpha_q} \Lambda_{\mathrm{sm}}\right)^{d/(q_{\mathrm{geom}})} \\
&\qquad + \left(C_{\mathrm{size,sm}} \varepsilon_{\mathrm{sm}}^{-(1-q/\beta)} \log^{q/2} \Lambda_{\mathrm{sm}}\right)^{d/(\beta-q)} \Bigg\} \Bigg] \log^{3+d/2} \Lambda_{\mathrm{sm}}, \\
S &\lesssim D^{\mathfrak{a}_{\mathrm{sm}}} A_{\mathcal{M}} \frac{dD(D+d+\beta)^{m_\beta}(d+q_{\mathrm{geom}})^{d+4}}{m_\beta!} \Bigg[\left(C_{\mathrm{size,sm}} \sigma_{\underline{t}}^{-4} \varepsilon_{\mathrm{sm}}^{-2}\right)^{d/(q_{\mathrm{geom}})} \\
&\qquad + \sum_{q=0}^{m_\beta} \Bigg\{ \left(C_{\mathrm{size,sm}} \varepsilon_{\mathrm{sm}}^{-2(1-q/\beta)/\alpha_q} \log^{q/\alpha_q} \Lambda_{\mathrm{sm}}\right)^{d/(q_{\mathrm{geom}})} \\
&\qquad + \left(C_{\mathrm{size,sm}} \varepsilon_{\mathrm{sm}}^{-(1-q/\beta)} \log^{q/2} \Lambda_{\mathrm{sm}}\right)^{d/(\beta-q)} \Bigg\} \Bigg] \log^{5+d/2} \Lambda_{\mathrm{sm}}, \\
\log B &\lesssim \log^2 \Lambda_{\mathrm{sm}}.
\end{aligned}
\tag{226}
$$

The remaining consequences of the complexity bound are recorded separately. The factors $\log^{3+d/2} \Lambda_{\mathrm{sm}}$ and $\log^{5+d/2} \Lambda_{\mathrm{sm}}$ include the logarithmic overhead of the projection-coordinate gate, the residual ramp, and the final products. The normalized-residual and Gaussian subnetworks contribute the explicit factor $(\sigma_{\underline{t}}^{-4} \varepsilon_{\mathrm{sm}}^{-2})^{d/q_{\mathrm{geom}}}$. This factor arises from the required projection accuracy, even though the construction uses no $\sigma_{\underline{t}}$-scale mesh. The fixed localization constant $c_\star$ is chosen from the annular Gaussian-tail bound in Lemma F.1 and is independent of $\sigma_{\underline{t}}$. For fixed d, β and fixed geometry, the bracketed non-logarithmic factor in (226) is

$$
\lesssim \left[\varepsilon_{\mathrm{sm}}^{-d/\beta} + (\sigma_{\underline{t}}^{-4} \varepsilon_{\mathrm{sm}}^{-2})^{d/q_{\mathrm{geom}}}\right] \log^{\mathfrak{p}_{\beta,d}} \Lambda_{\mathrm{sm}}
$$

after the finite order q_{geom} in (4) is fixed, and one may take

$$
\mathfrak{p}_{\beta,d} = c_{\mathrm{amb}} (q_{\mathrm{geom}}+2)^2 \Big(1 + d + \frac{(m_\beta+1)d}{\beta - m_\beta}\Big). \tag{227}
$$

Here $c_{\mathrm{amb}} \geq 1$ is a sufficiently large universal exponent constant, fixed after the arithmetic constructions. Lemma F.32 derives a common upper bound $C(q_{\mathrm{geom}}+2)^2(1+d+(m_\beta+1)d/(\beta-m_\beta))$ for the ambient exponents of both comparison constructions and their statistical transfer. Choose c_{amb} sufficiently large that the comparison-network depth, log-weight, and size prefactors each use exponent at most $\mathfrak{p}_{\beta,d}/8$, and their squared approximation-error prefactors exponent at most $\mathfrak{p}_{\beta,d}/4$. The unused margin accounts for multiplication of depth and sparsity in the entropy bound, time integration, and the distributional transfer. In particular, the corrected multiplication bound in Lemma I.3 changes fixed-degree products only by fixed constants and logarithmic factors. The large-noise polynomial convolutions use two-input products; the additional factor $\log d$ in a d-fold product and the linear-degree range term are retained in the schedule bounds of Lemmas I.8 and I.9. Consequently these arithmetic costs introduce no additional power of the sample size or of an approximation tolerance, and are covered by the same exponent margin. The common exponent is independent of $D, n, \sigma_t, p_{\min}$ and the numerical geometry constants. In the main $d > 2$ regime we take the canonical choice $q_{\mathrm{geom}} = q_{\mathrm{opt}}(\beta)$ from (5); then $q_{\mathrm{geom}} = \mathcal{O}_\beta(1)$ and the displayed formula gives $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$. If $\sigma_{\underline{t}} \gtrsim \varepsilon_{\mathrm{sm}}^{(\beta+1)/\beta}$, the choice $q_{\mathrm{geom}} \geq 6\beta + 4$ bounds the additional term by $C\varepsilon_{\mathrm{sm}}^{-d/\beta}$. This relation holds for the final statistical cutoffs. The factor $D^{\mathfrak{a}_{\mathrm{sm}}}$ records the retained-term allocations and coordinatewise normal accuracies; the lower-time dependence remains explicit until these cutoffs are substituted.

Proof of Propositions F.27 and F.28 and Lemma F.29. Fix i and $R \in \mathcal{R}$. The number of retained $(q, \boldsymbol{\lambda})$-terms is

$$\sum_{q=0}^{m_\beta} \#\{\boldsymbol{\lambda} \in \mathbb{N}^D : \|\boldsymbol{\lambda}\|_1 \le q\} = \sum_{q=0}^{m_\beta} \binom{D+q}{q} = \binom{D+m_\beta+1}{m_\beta}.$$

Also, since $d < D$ and $m_\beta < \beta$,

$$\binom{D+m_\beta+1}{m_\beta}(D+d+\beta) \le \frac{(D+d+\beta)^{m_\beta+1}}{m_\beta!},$$

which is the simpler ambient-dimension factor used in the width bound. Similarly,

$$\binom{D+m_\beta+1}{m_\beta} dD(d+q_{\mathrm{geom}})^{d+4} \le \frac{dD(D+d+\beta)^{m_\beta}(d+q_{\mathrm{geom}})^{d+4}}{m_\beta!},$$

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 $q \in \{0, \dots, m_\beta\}$ is a finite set and the coefficient network sizes contain only fixed powers of $(\varepsilon^{\mathrm{sm}}_{A,q})^{-1}$, all additional powers of $\binom{D+m_\beta+1}{m_\beta}$ are bounded by $D^{\mathfrak{a}_{\mathrm{sm}}}$ after increasing $\mathfrak{a}_{\mathrm{sm}}(\beta, d)$. Indeed,

$$\binom{D+m_\beta+1}{m_\beta} \le C_\beta D^{m_\beta}, \qquad \max_{0\le q\le m_\beta}\left\{\frac{d}{\beta-q}, \frac{d}{((\beta-q)\wedge 1)(\beta\vee 1)}\right\} \le \frac{d}{\beta-m_\beta}.$$

Thus the retained-term allocation contributes at most $D^{m_\beta d/(\beta-m_\beta)}$, up to fixed (β, d)-dependent constants, before the other displayed ambient factors are included. The residual and coefficient-coordinate constructions contribute at most $D^{d/(q_{\mathrm{geom}})}$ 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 Lemma F.18; for the rescaled 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. The internal residual scale is at most $C_{\mathrm{res}}D^2$ up to logarithms, where $C_{\mathrm{res}} \lesssim C_{\mathrm{proj}}\sigma_{\underline{t}}^{-2}\log(e/\varepsilon_{\mathrm{sm}})$. For the Gaussian factor, $C_{\mathrm{gau}} \lesssim CD^2\sigma_{\underline{t}}^{-4}\log^2(e/\varepsilon_{\mathrm{sm}})$. The first term of the size bound retains this larger lower-time factor. 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

$$\varepsilon^{\mathrm{sm}}_{\mathrm{term}} := c_{\mathrm{term}}\varepsilon_{\mathrm{sm}}\binom{D+m_\beta+1}{m_\beta}^{-1}, \qquad \sigma_{\mathrm{sm}} := C_{\mathrm{sm}}\varepsilon_{\mathrm{sm}}^{1/\beta}.$$

For $0 \le q \le m_\beta$, define

$$\alpha_q := (\beta-q)\wedge 1, \qquad \varepsilon^{\mathrm{sm}}_{A,q} := \frac{1}{2} \wedge c^{\mathrm{sm}}_{A,q}\varepsilon_{\mathrm{sm}}^{1-q/\beta}\big(\log\Lambda_{\mathrm{sm}}\big)^{-q/2},$$

where $c^{\mathrm{sm}}_{A,q} > 0$ is chosen no larger than the fixed quantity

$$c_{A,q}c_{\mathrm{term}}C_{\mathrm{sm}}^{-q}\binom{D+m_\beta+1}{m_\beta}^{-1},$$

and $c_{A,q}$ is the constant from Corollary F.21. By the fixed choice of $c_\star$ in the active-set definition, and after decreasing the fixed numerical tolerances below if necessary, the inactive tail in Lemma 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 Proposition F.16. On the inactive region inside A_t, use zero. Thus define

$$R^{\mathrm{sm}}_{i,\mathrm{app}}(\boldsymbol{x}, t) := \begin{cases} \exp\!\big(-\dfrac{\|\boldsymbol{r}_i(\boldsymbol{x},t)\|_2^2}{2\sigma_t^2}\big)\displaystyle\sum_{q=0}^{m_\beta}\sigma_t^q \\ \qquad \times \displaystyle\sum_{\|\boldsymbol{\lambda}\|_1\le q} A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big)\boldsymbol{\nu}_i(\boldsymbol{x},t)^{\boldsymbol{\lambda}}, & (\boldsymbol{x},t) \in \mathcal{K}_i(\varepsilon_{\mathrm{sm}}), \\ 0, & (\boldsymbol{x},t) \notin \mathcal{K}_i(\varepsilon_{\mathrm{sm}}). \end{cases}$$

By Proposition F.16 and Lemma F.1, uniformly over $t \in \mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})$ and $\boldsymbol{x} \in A_t$,

$$|R_i(\boldsymbol{x}, t) - R^{\mathrm{sm}}_{i,\mathrm{app}}(\boldsymbol{x}, t)| \le \varepsilon_{\mathrm{sm}}/3. \tag{228}$$

The active part follows because the remainder in (191) is bounded by $C^{\max}_{\mathrm{Lap}}\sigma_t^{\beta}$, and the threshold C_{sm} in Definition 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 (223).

Step 2: Projection-coordinate gate. Let $\square_i^A$ be the coefficient-extension box from Lemma 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 Lemma 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 Corollary F.23, 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 a fixed neighborhood of K_i° compactly contained in $\square_i^A$, 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

$$\Big[0, \big(\frac{3c_\star}{2}\big)^2 \log(e/\varepsilon_{\mathrm{sm}})\Big]$$

and equal to zero on

$$\Big[(2c_\star)^2 \log(e/\varepsilon_{\mathrm{sm}}), \infty\Big).$$

Let $\widetilde{\boldsymbol{u}}_i^{\Pi}$ be the chart-coordinate network obtained from the fixed-radius version of Corollary F.14, and let $\widetilde{\boldsymbol{\nu}}_i$ be the normalized-residual network from the fixed-radius version of Lemma F.17. Write

$$\ell_{\mathrm{gate}} := \log(e/\varepsilon_{\mathrm{sm}}), \qquad \varepsilon_{\nu,\mathrm{gate}} := c_{\mathrm{gate}}\big(1 \wedge D^{-1/2}\ell_{\mathrm{gate}}^{1/2}\big).$$

Choose the projection-coordinate accuracy and normal-residual accuracy so that, on $\mathcal{K}_i^+(\varepsilon_{\mathrm{sm}})$,

$$\|\widetilde{\boldsymbol{u}}_i^{\Pi} - \boldsymbol{u}_i^{\Pi}\|_\infty \le c_{\mathrm{gate}}\ell_{\mathrm{gate}}^{-1}, \qquad \|\widetilde{\boldsymbol{\nu}}_i - \boldsymbol{\nu}_i\|_\infty \le \varepsilon_{\nu,\mathrm{gate}},$$

where $c_{\mathrm{gate}} > 0$ is a sufficiently small geometric constant. Form $\|\widetilde{\boldsymbol{\nu}}_i\|_2^2$ using the multiplication lemma and define

$$\Theta_i^{\mathrm{sm}}(\boldsymbol{x}, t) := \min\Big\{\theta_i^A\big(\widetilde{\boldsymbol{u}}_i^{\Pi}(\boldsymbol{x}, t)\big), \theta_{\mathrm{res}}\big(\|\widetilde{\boldsymbol{\nu}}_i(\boldsymbol{x}, t)\|_2^2\big)\Big\}, \tag{229}$$

with the minimum implemented exactly by ReLU, and with the squared norm replaced by its arithmetic network, clipped to $[0, \infty)$, inside the residual ramp. The normal-residual accuracy gives

$$\|\widetilde{\boldsymbol{\nu}}_i - \boldsymbol{\nu}_i\|_2 \le c_{\mathrm{gate}}\ell_{\mathrm{gate}}^{1/2},$$

On the enlarged active region, $\|\boldsymbol{\nu}_i\|_2 \le 3c_\star\ell_{\mathrm{gate}}^{1/2}$. Hence

$$\big|\|\widetilde{\boldsymbol{\nu}}_i\|_2^2 - \|\boldsymbol{\nu}_i\|_2^2\big| \le (6c_\star c_{\mathrm{gate}} + c_{\mathrm{gate}}^2)\ell_{\mathrm{gate}}.$$

Choose c_{gate} so that $6c_\star c_{\mathrm{gate}} + c_{\mathrm{gate}}^2 \le c_\star^2/4$, and approximate the squared norm with additive error at most $c_\star^2\ell_{\mathrm{gate}}/4$. The nonnegative clipping cannot increase this error. The total squared-norm error is then at most $c_\star^2\ell_{\mathrm{gate}}/2$. On $\mathcal{K}_i(\varepsilon_{\mathrm{sm}})$, the implemented value is therefore at most $(3/2)c_\star^2\ell_{\mathrm{gate}} < (3c_\star/2)^2\ell_{\mathrm{gate}}$, so the residual ramp is exactly one. The coordinate ramp is also one after decreasing c_{gate} according to its fixed buffer. These analytic tolerances use only ℓ_{gate}; their implementation costs are bounded using the larger scale $\log \Lambda_{\mathrm{sm}}$. 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 error. 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 satisfy their error bounds on $\mathcal{K}_i^{+}(\varepsilon_{\mathrm{sm}})$, and the analytic tail bound (223) 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 enlarged active region where those bounds hold, not from a $\sigma_{\underline{t}}$-mesh of $\mathcal{S}_i$. Its size is included in the projection-residual term of (225).

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 Lemma 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 $q \le m_\beta$ and $\|\boldsymbol{\lambda}\|_1 \le q$, approximate

$$A^{(R)}_{i,q,\boldsymbol{\lambda}}\big(\boldsymbol{u}_i^{\Pi}(\boldsymbol{x},t),t\big)$$

by the fixed-radius version of Corollary F.21, using Corollary F.23, with term accuracy

$$\varepsilon^{\mathrm{sm}}_{\mathrm{term}}.$$

Here and below, Corollary F.21 and the residual and Gaussian lemmas are used through Corollary F.23 with $c_a = 3c_\star$. Approximate $\boldsymbol{\nu}_i$ by Lemma F.17, form the monomial $\boldsymbol{\nu}_i^{\boldsymbol{\lambda}}$ by Lemma I.3, approximate σ_t^q by the constant-one network when $q = 0$ and by Lemma I.8 when $q \ge 1$, and approximate the Gaussian factor by Lemma F.18. In the case $q \ge 1$, the accuracy parameter passed to Lemma I.8 is chosen as the minimum of the required schedule accuracy, $\underline{t}/2$, and $1/2$. This makes the admissibility condition in Lemma 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 (226). Internal accuracies are chosen so that each retained term contributes at most $\varepsilon^{\mathrm{sm}}_{\mathrm{term}}$ to the final error after multiplication by the bounded remaining factors. The bounds

$$\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le C\sqrt{\log \Lambda_{\mathrm{sm}}}, \qquad \sigma_t \le C_{\mathrm{sm}}\varepsilon_{\mathrm{sm}}^{1/\beta},$$

are exactly the bounds used in Corollary 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 allowing one additional $\mathcal{O}(\varepsilon_{\mathrm{sm}})$ error for 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 (204) gives $\overline{A}^{(R)}_{i,q,\boldsymbol{\lambda}}(\boldsymbol{u}^{\Pi},t) = 0$ for every retained coefficient. The implemented coefficient network is evaluated at the approximate coordinate $\widetilde{\boldsymbol{u}}_i^{\Pi}$ used in $\widetilde{\mathcal{A}}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon^{\mathrm{sm}}_{A,q}}$, not exactly at $\boldsymbol{u}^{\Pi}$. By the fixed-radius coordinate approximation in Corollary F.23, the choice $\varepsilon_u \asymp (\varepsilon^{\mathrm{sm}}_{A,q})^{1/((\beta-q)\wedge 1)}$, the $(\beta-q)$-Hölder bound in (205), and the intrinsic coefficient-network error,

$$\Big|\widetilde{\mathcal{A}}^{(R)}_{i,q,\boldsymbol{\lambda},\varepsilon^{\mathrm{sm}}_{A,q}}(\boldsymbol{x},t)\Big| \le C\|\widetilde{\boldsymbol{u}}_i^{\Pi} - \boldsymbol{u}^{\Pi}\|_\infty^{(\beta-q)\wedge 1} + \varepsilon^{\mathrm{sm}}_{A,q} \le C\varepsilon^{\mathrm{sm}}_{A,q}.$$

In both cases, the bounds

$$\|\boldsymbol{\nu}_i(\boldsymbol{x},t)\|_2 \le C\sqrt{\log \Lambda_{\mathrm{sm}}}, \qquad \sigma_t^q \le \sigma^q_{\mathrm{sm}},$$

and the definition of $\varepsilon^{\mathrm{sm}}_{A,q}$ imply that each retained term is bounded by its termwise error tolerance. 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

$$\sup_{\substack{t\in\mathcal{I}_{\mathrm{sm}}(\varepsilon_{\mathrm{sm}})\\(\boldsymbol{x},t)\in\mathcal{K}_i^{+}(\varepsilon_{\mathrm{sm}})}}\left|\widetilde{R}^{\mathrm{sm}}_i(\boldsymbol{x},t)-R^{\mathrm{sm}}_{i,\mathrm{app}}(\boldsymbol{x},t)\right|\le 2\varepsilon_{\mathrm{sm}}/3. \tag{230}$$

Combining (228) and (230) proves (222).

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 $\varepsilon^{\mathrm{sm}}_{A,q}$, and enlarging $C_{\mathrm{size,sm}}$ to absorb the fixed coefficient constants gives (226); the additional powers of the retained-term count are absorbed by $D^{\mathfrak{a}_{\mathrm{sm}}}$. The projection-coordinate gate in (229) consists of a fixed chart coordinate bump, one residual-norm computation, and one scalar residual ramp. The residual and Gaussian costs are bounded by $(C\sigma_{\underline{t}}^{-4}\varepsilon_{\mathrm{sm}}^{-2})^{d/q_{\mathrm{geom}}}$ times the displayed ambient and logarithmic factors. For coefficient degree q, the geometric coordinate cost has accuracy exponent $2(1-q/\beta)/(\alpha_q q_{\mathrm{geom}})\le 1/\beta$, while the density cost has exponent $(1-q/\beta)/(\beta-q)=1/\beta$. Here $\alpha_q=(\beta-q)\wedge 1$ and $q_{\mathrm{geom}}\ge 2(\beta\vee 1)$. These inequalities prove (225) with its explicit lower-time term. The geometric projection terms use the smoothness q_{geom}, 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

Lemma F.30 is the key lower-bound step in the small-noise branch. Without first factoring out this term, the denominator contains $\exp\{-\|\boldsymbol{r}\|_2^2/(2\sigma_t^2)\}$ 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.30 (Small-noise rescaled denominator lower bound)**.** *Assume Assumptions 1 and 3. 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

$$\boldsymbol{y}_\star\in\operatorname*{arg\,min}_{\boldsymbol{y}\in\mathcal{M}}\|\boldsymbol{x}-m_t\boldsymbol{y}\|_2,\qquad \boldsymbol{r}_\star:=\boldsymbol{x}-m_t\boldsymbol{y}_\star.$$

Then

$$\exp\Big(\frac{\|\boldsymbol{r}_\star\|_2^2}{2\sigma_t^2}\Big)\mathcal{Q}(\boldsymbol{x},t)\ge c_Q p_{\min}\big(1+C_\varrho\sigma_t\sqrt{H}\big)^{-d/2}\ge C_Q^{-1}p_{\min}H^{-d/4}. \tag{231}$$

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 rescaled 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 $\phi_{i_\star}(\mathcal{S}_{i_\star})\Subset K^\circ_{i_\star}$. 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

$$B_{\mathbb{R}^d}(\boldsymbol{u}_\star,r_Q)\subset K^\circ_{i_\star},\qquad a^{(\mathcal{Q})}_{i_\star}(\boldsymbol{u})=\bar\rho_{i_\star}(\boldsymbol{u})p_0(\boldsymbol{z}_{i_\star}(\boldsymbol{u}))\sqrt{\det\boldsymbol{G}_{i_\star}(\boldsymbol{u})}\ge c_a p_{\min}$$

for every $\boldsymbol{u} \in B_{\mathbb{R}^d}(\boldsymbol{u}_\star, r_Q)$. By Definition B.6 and Lemma F.3, these constants may be chosen with $r_Q^{-1} \vee c_a^{-1} \le C\mathrm{T}^C_{\mathcal{M},q_{\mathrm{geom}}}$.

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,

$$\begin{aligned} \exp\Big(\frac{\|\boldsymbol{r}_\star\|_2^2}{2\sigma_t^2}\Big)\mathcal{Q}(\boldsymbol{x}, t) &\ge \sigma_t^{-d}\int_{B(\boldsymbol{u}_\star, r_Q)} \exp\Big(-\frac{\|\boldsymbol{x} - m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}_\star\|_2^2}{2\sigma_t^2}\Big) a_i^{(\mathcal{Q})}(\boldsymbol{u})\mathrm{d}\boldsymbol{u} \\ &\ge c_a p_{\min} a_\star^{-d/2}\int_{\|\boldsymbol{w}\|_2 \le R_Q} e^{-C\|\boldsymbol{w}\|_2^2}\mathrm{d}\boldsymbol{w} \\ &\ge c_Q p_{\min}(1 + \|\boldsymbol{r}_\star\|_2)^{-d/2}. \end{aligned}$$

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 (231). 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.31 (H-localized small-noise chart approximation after rescaling)**.** *Assume Assumptions 1 and 2. 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 $[t_0, t_1] \subset [\underline{t}, \bar{t}]$, $0 < t_0 < t_1 \le 1$, and $\sigma_{t_1} \le \varepsilon_{\mathrm{sm}}^{1/\beta}$, and that, for every $t \in [t_0, t_1]$,

$$3c_\star\frac{\sigma_t}{m_t}\sqrt{H} \le \eta_\Pi/2. \tag{232}$$

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 (232). For $R \in \{\mathcal{Q}, \mathcal{P}_1, \dots, \mathcal{P}_D\}$, set

$$\bar{R}_i(\boldsymbol{x}, t) := \exp\Big(\frac{\|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big)R_i(\boldsymbol{x}, t).$$

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

$$\sup_{\substack{t\in[t_0,t_1]\\ \boldsymbol{x}\in A_t(H)}} |\widehat{\bar{R}}^{\mathrm{sm}}_{i,H}(\boldsymbol{x},t)-\bar{R}_i(\boldsymbol{x},t)| \le C\varepsilon_{\mathrm{sm}}H^C. \tag{233}$$

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

$$|\widehat{\bar{\mathcal{Q}}}^{\mathrm{sm}}_H(\boldsymbol{x},t)-\bar{\mathcal{Q}}(\boldsymbol{x},t)| \vee \|\widehat{\bar{\mathcal{U}}}^{\mathrm{sm}}_H(\boldsymbol{x},t)-\bar{\mathcal{U}}(\boldsymbol{x},t)\|_\infty \le C\varepsilon_{\mathrm{sm}}H^C \tag{234}$$

uniformly over $t\in[t_0,t_1]$ and $\boldsymbol{x}\in A_t(H)$. The networks have the intrinsic size factors in (225), with an additional polynomial in H, D, and $\log\Lambda_{\mathrm{sm}}$. Each exponent is at most $\mathfrak{p}_{\beta,d}/8$ after choosing the universal c_{amb} sufficiently large. The polynomial includes the fixed-atlas sum, the D centered coordinates, and the coefficient tolerances displayed in the proof.

For $H = D\vee\log n$, the additional H^C factors are absorbed by the powers $(D\vee\log n)^{\mathfrak{p}_{\beta,d}}$ used below.

Proof. Step 1: H-active localization and reach control. For $c_a>0$, define

$$\mathcal{K}_i^{(c_a,H)} := \Big\{(\boldsymbol{x},t):\boldsymbol{x}\in A_t(H),\ \exists \boldsymbol{y}\in\mathcal{S}_i \text{ with } \|\boldsymbol{x}-m_t\boldsymbol{y}\|_2\le c_a\sigma_t\sqrt{H}\Big\}.$$

Because $c_\star\ge 2C_\varrho$ 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

$$\operatorname{dist}(\boldsymbol{x}/m_t,\mathcal{M}) \le 3c_\star\frac{\sigma_t}{m_t}\sqrt{H} \le \eta_\Pi/2.$$

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 $\|\boldsymbol{\nu}\|_2\le 3c_\star\sqrt{H}$.

Step 2: coercivity after rescaling. Let $\Phi_i(\boldsymbol{u};\boldsymbol{\xi})=\frac12\|\boldsymbol{\xi}-\boldsymbol{z}_i(\boldsymbol{u})\|_2^2$, 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 Lemma F.5, with the H-reach condition above, gives

$$\nabla^2_{\boldsymbol{u}}\Phi_i(\boldsymbol{u}^\Pi;\boldsymbol{\xi}) \succeq \frac12\lambda_i^{(-)}\boldsymbol{I}_d.$$

Hence there are constants $r_i^0>0$ and $c_0>0$ such that

$$\Phi_i(\boldsymbol{u};\boldsymbol{\xi})-\Phi_i(\boldsymbol{u}^\Pi;\boldsymbol{\xi}) \ge c_0\|\boldsymbol{u}-\boldsymbol{u}^\Pi\|_2^2$$

whenever $\boldsymbol{u}\in\square_i$ and $\|\boldsymbol{u}-\boldsymbol{u}^\Pi\|_2\le r_i^0$. A quantitative global bound follows directly from reach: for $\boldsymbol{\pi}=\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$ and every $\boldsymbol{y}\in\mathcal{M}$,

$$\|\boldsymbol{x}-m_t\boldsymbol{y}\|_2^2-\|\boldsymbol{x}-m_t\boldsymbol{\pi}\|_2^2 \ge m_t^2\Big(1-\frac{\|\boldsymbol{x}/m_t-\boldsymbol{\pi}\|_2}{\kappa}\Big)\|\boldsymbol{y}-\boldsymbol{\pi}\|_2^2 \ge \tfrac12 e^{-2}\|\boldsymbol{y}-\boldsymbol{\pi}\|_2^2.$$

Here the first inequality is the normal-chord reach inequality (Federer, 1959), and the second uses $\eta_\Pi\le\kappa$, the assumed reach condition, and $t_1\le 1$. Because ϕ_i is an orthogonal projection coordinate, $\|\boldsymbol{z}_i(\boldsymbol{u})-\boldsymbol{z}_i(\boldsymbol{u}^\Pi)\|_2\ge\|\boldsymbol{u}-\boldsymbol{u}^\Pi\|_2$. Consequently we obtain the uniform rescaled lower bound

$$\frac{\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u}^\Pi+\sigma_t\boldsymbol{w})\|_2^2-\|\boldsymbol{r}\|_2^2}{2\sigma_t^2}\ge c\|\boldsymbol{w}\|_2^2$$

for all $\boldsymbol{u}^\Pi+\sigma_t\boldsymbol{w}\in S_i^{(R)}$, without decreasing a geometry-dependent noise threshold. Consequently, for $R_H:=A_{\mathrm{tail}}\sqrt{H}$,

$$\int_{\|\boldsymbol{w}\|_2>R_H}\exp\Big(-\frac{\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u}^\Pi+\sigma_t\boldsymbol{w})\|_2^2-\|\boldsymbol{r}\|_2^2}{2\sigma_t^2}\Big)(1+\|\boldsymbol{w}\|_2)^C\mathrm{d}\boldsymbol{w}\le C\varepsilon_{\mathrm{sm}}H^{-C},$$

where we used $H \geq \log(e/\varepsilon_{\mathrm{sm}})$ and the fixed choice of A_{tail} to dominate the finitely many polynomial powers used below. The margin in (90) ensures $\sigma_t R_H \leq \delta_{\mathrm{box}}/6$. Thus $\boldsymbol{u}^\Pi + s\boldsymbol{w} \in \square_i$ for every $0 \leq s \leq \sigma_t$ and $\|\boldsymbol{w}\|_2 \leq R_H$, as required for the Taylor expansion.

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 rescaled 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 rescaled 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 (232) holds,

$$\mathrm{dist}(\boldsymbol{x}_s/m_t, \mathcal{M}) \leq s\|\boldsymbol{\nu}\|_2/m_t \leq 3c_\star\sigma_t m_t^{-1}\sqrt{H} \leq \eta_\Pi/2.$$

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 Lemma F.12 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 m_β, 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

$$G_R(\sigma_t, \boldsymbol{w}) = \sum_{q=0}^{m_\beta} \sigma_t^q G_{R,q}(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t) + \sigma_t^\beta \mathcal{E}_R(\boldsymbol{u}^\Pi, \boldsymbol{\nu}, \boldsymbol{w}, t),$$

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 Proposition F.16, because both constructions

differentiate the same recentered integrand after the common normal Gaussian factor has been removed. Integrating the Taylor expansion on $\|\boldsymbol{w}\|_2 \le R_H$ and adding the tail from Step 2 yields

$$\bar{R}_i(\boldsymbol{x},t) = \sum_{q=0}^{m_\beta} \sigma_t^q \sum_{\|\boldsymbol{\lambda}\|_1 \le q} A^{(R)}_{i,q,\boldsymbol{\lambda}}(\boldsymbol{u}^\Pi, t)\boldsymbol{\nu}^{\boldsymbol{\lambda}} + \mathrm{Rem}^{(R)}_{i,H}(\boldsymbol{x},t),$$

Here $|\mathrm{Rem}^{(R)}_{i,H}(\boldsymbol{x},t)| \le C(\sigma_t^\beta + \varepsilon_{\mathrm{sm}})H^C$. Since $\sigma_t \le \sigma_{t_1} \le \varepsilon_{\mathrm{sm}}^{1/\beta}$, this remainder is bounded by $C\varepsilon_{\mathrm{sm}}H^C$, with the finite geometry and density factors retained in 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

$$\bar{\mathcal{U}}_{i,k}(\boldsymbol{x},t) = \sum_{q=0}^{m_\beta} \sigma_t^q \sum_{\|\boldsymbol{\lambda}\|_1 \le q+1} A^{(\bar{U})}_{i,k,q,\boldsymbol{\lambda}}(\boldsymbol{u}^\Pi,t)\boldsymbol{\nu}^{\boldsymbol{\lambda}} + \mathrm{Rem}^{(\bar{U})}_{i,k,H}(\boldsymbol{x},t),$$

where $|\mathrm{Rem}^{(\bar{U})}_{i,k,H}(\boldsymbol{x},t)| \le C\varepsilon_{\mathrm{sm}}H^C$. 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

$$m_t^a P(\boldsymbol{\nu},\boldsymbol{w})H(\boldsymbol{u}^\Pi)\partial^\alpha a_i^{(\mathcal{Q})}(\boldsymbol{u}^\Pi)\exp\Big(-\frac{m_t^2}{2}\boldsymbol{w}^\top \boldsymbol{G}_i(\boldsymbol{u}^\Pi)\boldsymbol{w}\Big), \qquad |\alpha| \le q,$$

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 Lemma 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 $\boldsymbol{u} \notin S_i^{(\mathcal{Q})} = \mathrm{supp} a_i^{(\mathcal{Q})}$, 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 Lemma F.19. There are at most $D\binom{D+m_\beta+2}{m_\beta+1}$ centered terms. For degree q, allocate coefficient accuracy

$$\varepsilon^U_{A,q} = \frac{1}{2} \wedge \frac{c\varepsilon_{\mathrm{sm}}\sigma_{\mathrm{sm}}^{-q}H^{-(q+1)/2}}{D\binom{D+m_\beta+2}{m_\beta+1}}, \qquad \sigma_{\mathrm{sm}} = \varepsilon_{\mathrm{sm}}^{1/\beta}.$$

Then $\sigma_t^q|\boldsymbol{\nu}^\lambda|\varepsilon^U_{A,q}$ is at most the total tolerance divided by the term count. Its non-logarithmic dependence on $\varepsilon_{\mathrm{sm}}$ is $\varepsilon_{\mathrm{sm}}^{1-q/\beta}$, so the density-network cost remains $\varepsilon_{\mathrm{sm}}^{-d/\beta}$. The additional term count and the factor $H^{(q+1)/2}$ contribute polynomial ambient factors, including their powers $d/(\beta-q)$. The enlarged exponent (227) includes these factors. 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: active chart gate and network construction. 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 Lemma F.26 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 an accurate approximation to the chart-i coordinate of $\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)$. For

the coefficient functions in Steps 3 and 4, use the extensions from Lemma F.19 and the centered-numerator analogues verified in Step 4, evaluate them at the clipped coordinate $\breve{\boldsymbol{u}}_i^A$, and multiply the retained expansion by Γ_i^H. Use the global normal-residual network from Corollary F.24 on the active union $\bigcup_j \mathcal{K}_j^{(3c_\star,H)}$, which contains $A_t(H)$ by Step 1. Use the weighted coefficient tolerances from (210) and Step 4. Choose each coefficient-coordinate tolerance proportional to $(\varepsilon_{A,q})^{1/\alpha_q}$. Allocate accuracy $c\varepsilon_{\mathrm{sm}}H^{-C}$ to each remaining monomial, schedule power, product, normal-residual operation, and arithmetic gate operation, after division by the displayed term count. Multiplication by Γ_i^H uses the zero-preserving network in Lemma I.3. On $\mathcal{K}_i^{(3c_\star,H)}$ the gate equals one, the coordinate and normal-residual errors satisfy their stated bounds, and the Taylor expansions from Steps 3 and 4 are approximated with total error $C\varepsilon_{\mathrm{sm}}H^C$. The objective gate reuses the projection networks. Its coordinate tolerance has the same $\varepsilon_{\mathrm{sm}}$ exponent as the weighted coefficient construction, and all additional factors are polynomial in H and D. Together with the explicit residual cost in (225), the degree-dependent tolerances in Step 4 give the stated size bounds.

Step 6: inactive and transition regions after factoring out the common normal Gaussian term. 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

$$\|\boldsymbol{x}-m_t\boldsymbol{z}_i(\boldsymbol{u})\|_2^2-\|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2\ge cc_\star^2\sigma_t^2H.$$

The reach-tube objective gap also gives

$$\|\boldsymbol{x}-m_t\boldsymbol{y}\|_2^2-\|\boldsymbol{r}\|_2^2\ge\tfrac12 m_t^2\|\boldsymbol{y}-\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)\|_2^2.$$

Split the exponential into two equal factors, use the inactive phase gap for one factor, and integrate the other using the uniform local volume bound and annuli. The resulting integral is at most $C\sigma_t^d e^{-cc_\star^2H}$; the prefactor σ_t^{-d} therefore cancels. For the centered numerator, $|m_ty_k-x_k|/\sigma_t\le C\sqrt{H}+m_t\|\boldsymbol{y}-\Pi_{\mathcal{M}}(\boldsymbol{x}/m_t)\|_2/\sigma_t$, and the additional radial factor is absorbed by the same Gaussian integral. Thus the true inactive contributions are bounded by $C\varepsilon_{\mathrm{sm}}H^C$, uniformly in the lower time cutoff. 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 $\Gamma_i^H>0$, Step 5 guarantees 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 $\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\boldsymbol{u}_i^\Pi,t)=0$, and the Hölder estimate plus the 1-Lipschitz clipping to $\square_i^A$ gives $|\overline{A}_{i,q,\boldsymbol{\lambda}}^{(R)}(\breve{\boldsymbol{u}}_i^A,t)|\le C(\varepsilon_u^{\mathrm{cert}})^{(\beta-q)\wedge1}$. 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 error tolerance $c\varepsilon_{\mathrm{sm}}H^{-C}$ after multiplication by σ_t^q and the $H^{q/2}$ normal-monomial bound. Combining the active approximation, the true inactive tail, and the established off-support network control proves (233). Summing over the fixed atlas and using the direct centered-numerator networks from Step 4 proves (234). The displayed network-size statement follows from the size bounds in Step 5, the fixed atlas summation, and the coordinatewise parallelization for $\bar{\mathcal{U}}$. □

Lemma F.32 (Uniform ambient-exponent bound)**.** *Assume Assumptions 1 to 3 and $D\vee p_{\min}^{-1}\le n^{a_0}$ for fixed $a_0>0$. For the small-noise claims, also assume the reach and denominator-resolution conditions (236) and (237). Use the statistical cutoffs in (3), the geometric order in (4), and the radius cap in Proposition G.3. Set $H=D\vee\log n$, $q=q_{\mathrm{geom}}$, $\ell=m_\beta$, and $a=\beta-\ell\in(0,1]$. There is a universal constant C_0 such that, with*

$$E=C_0(q+2)^2\big(1+d+(\ell+1)d/a\big),\qquad G=C\Gamma_{\mathcal{M},q}^C B_0^C p_{\min}^{-C},$$

the comparison networks may be chosen with

$$L_k\vee\log B_k\le GH^E,\qquad \|\boldsymbol{W}_k\|_\infty\vee S_k\le GN_kH^E,\qquad \begin{cases}N_1=n^{d/(d+2\beta)},\\ N_k=\sigma_{t_{k-1}}^{-d}+\varepsilon_n^{-d/q_\star},\quad k\ge2.\end{cases}\tag{235}$$

Here and in the proof the constants and powers in G may be enlarged and may depend on the fixed (β, d, q, a_0), but C_0 is independent of these parameters. The squared scaled-score approximation errors have ambient exponent at most $4E$. Under the stated reach and denominator-resolution conditions, taking $c_{\mathrm{amb}} \geq 64C_0$ in (227) also covers the entropy, time-integration, and Wasserstein-transfer exponents.

Proof. We keep multiplicative constants depending on intrinsic dimension in G, rather than converting them into powers of H. The following counts bound the exponents of H uniformly in d.

For the large-noise construction, (64) and (261) give depth and log-weight at most GH^6 and size at most $GD^2 r_k^{-d} H^{2d+12}$. The explicit cap gives

$$r_k^{-d} \leq G\big(\sigma_{t_{k-1}}^{-d} + H^{2d}\varepsilon_n^{-d/q_\star}\big),$$

so the total size is at most $GN_k H^{4d+14}$. The choice $\varepsilon_{\mathrm{sc}} = \varepsilon_n/\sqrt{D}$ makes the squared scaled-score approximation error at most $C\varepsilon_n^2$ before final clipping. In particular, the unspecified geometry powers in the cap do not enter the ambient exponent.

For the small-noise construction, put $\varepsilon = \varepsilon_{\mathrm{sm}}$ and $b = (3\ell + 5)/2$. The centered expansion has at most $C_\beta D^{\ell+2}$ terms, including its D output coordinates. At retained order $j \leq \ell$, its coefficient accuracy in Step 4 of Lemma F.31 can therefore be chosen with

$$\varepsilon_{A,j}^{-1} \leq G\varepsilon^{-(1-j/\beta)} H^b, \qquad \alpha_j = (\beta - j) \wedge 1 \geq a.$$

The two non-logarithmic costs in (214) are consequently bounded by

$$\begin{aligned} \varepsilon_{A,j}^{-d/(\beta-j)} &\leq G\varepsilon^{-d/\beta} H^{bd/a}, \\ D^{1+d/q}\varepsilon_{A,j}^{-2d/(\alpha_j q)} &\leq G\varepsilon^{-d/\beta} H^{1+d/q+2bd/(aq)}. \end{aligned}$$

The second inequality uses $q \geq 2(\beta \vee 1)$ and $(\beta - j)/\alpha_j \leq \beta \vee 1$. Parallelizing all terms adds at most $\ell + 2$ to these ambient exponents. The finite sum over retained orders changes the multiplicative constant, not the exponent.

For the remaining small-noise modules, all normal monomials have degree at most $\ell + 1$ and all their first derivatives are bounded by $GH^{\ell/2}$ on the active set. Thus coordinatewise normal accuracy $cG^{-1}\varepsilon H^{-2(\ell+2)}$ suffices after summing the retained terms. Substitution into the residual/projection bounds contributes at most $C(\ell + 2)d/q$ additional ambient powers, while retaining the factor $(\sigma_{t_0}^{-4}\varepsilon^{-2})^{d/q}$ already displayed in (225). The certified chart gate uses the same coefficient-coordinate accuracies and adds no further inverse-tolerance power by Lemma F.26. The schedule powers must be counted separately from the normal monomials. By (207), their exponents satisfy $|r_e| \leq d + 3\ell$ for the uncentered coefficients. For a centered coefficient, a derivative of order h of $N_k(s, \boldsymbol{w})$ contributes tangential degree at most $h + 1$ and schedule degree at most one, so the same count gives $|r_e| \leq d + 3\ell + 1$. On $[t_0, t_1]$, $m_t \geq e^{-1}$; hence these degrees and ranges depend on (d, β) but not on (D, n). For $k = |r_e| \geq 2$ and accuracy $0 < \delta \leq 1/2$, approximate the appropriate base m_t or m_t^{-1} to accuracy $\delta/(4ke^{k-1})$ and apply Lemma I.3 with range e and product accuracy $\delta/2$. Its perturbation bound gives total error at most δ, while Lemmas I.2 and I.6 give

$$L_{\mathrm{sch}} \vee \log B_{\mathrm{sch}} \leq C_{d,\beta} \log^2(e/\delta), \qquad \|\boldsymbol{W}_{\mathrm{sch}}\|_\infty \vee S_{\mathrm{sch}} \leq C_{d,\beta} \log^4(e/\delta).$$

The cases $k = 0, 1$ use a constant or the base network directly. Because the allocated accuracy logarithms are bounded by GH, these schedule modules contribute at most GH^4 per retained term; their degree-dependent constants are absorbed into G, not into additional powers of H. The remaining products have degree at most $C(q + 2)$, and the remaining logarithmic powers in (213) and (226), after the coefficient costs above are accounted for, are at most $C(q + 2) + d/2$. Together with signed identity channels and parallelization, these arithmetic costs add at most $C(q + 2)^2(1 + d)$ to the ambient exponent. Both intrinsic factors are at most $Cn^{d/(d+2\beta)}$ by (251). Since $b \leq 4(\ell + 1)$, these contributions establish (235) after a universal enlargement of C_0.

The bounds on the scalar Taylor remainders used for the small-noise integrals involve at most $\ell + 2$ powers of the normal variable, multiplied by Gaussian-integrable polynomial factors in the tangential variable. Integrating those factors contributes constants depending on (d, β), not additional powers

of H. Hence the integral errors are bounded by $G\varepsilon H^{q+2}$, a conservative bound also covering the allocated network errors and the inactive tails. The denominator floor is $G^{-1}H^{-d/4}$, and the centered ratio is bounded by $GH^{1/2}$. The ratio perturbation estimate and $\sqrt{D} \le \sqrt{H}$ therefore give squared scaled-score error at most $G\varepsilon^2 H^{2q+5+d/2}$, which is covered by H^{4E}. Reciprocal and final schedule accuracies enter their network sizes only logarithmically; the factor $H^{d/4}$ in the reciprocal range does not require a new density or projection approximation. Final score clipping and off-tube tails are handled by Lemmas C.3 and I.11 and do not increase the asserted exponent budget.

Finally, $\log N_k \le C \log n$, so (8) and (235) bound the entropy by $GN_k H^{3E+1}$. The oracle prefactor $H^2 \int_{I_k} \sigma_t^{-2} \mathrm{d}t$ adds at most three ambient powers, since the integral is at most $C \log n$. There are at most $C \log n$ slabs, with $t_k \le C \log n$; the Wasserstein transfer uses their square-root risks and the factor $H\sqrt{t_k \log n}$. These operations, the aggregate network sum, and the corresponding grid-weighted bounds require at most $8E + 8$ ambient powers in total. Choose $C_0 \ge 1$ and $c_{\mathrm{amb}} \ge 64 C_0$. Then $\mathfrak{p}_{\beta,d} \ge 64E$ supplies the earlier $\mathfrak{p}_{\beta,d}/8$ capacity and $\mathfrak{p}_{\beta,d}/4$ squared-error margins, as well as all subsequent transfers. For $q = q_{\mathrm{opt}}(\beta)$, this proves $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$ without hiding a dimension-dependent exponent in an unspecified constant. □

Final small-noise parameters. For this construction set $[\underline{t}, \bar{t}] = [t_0, t_1]$, so the OU schedule bound may be taken as $\underline{m} = e^{-1}$, independently of n, D. The localization constants are chosen on this interval, not on the terminal interval $[t_0, T]$. They can be chosen with $c_\star + A_{\mathrm{tail}} \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C$: the annular bounds, Gaussian moments, and the finitely many retained Taylor derivatives have prefactors polynomial in these envelopes, and their decay constants have polynomially bounded reciprocals. Explicitly, a bound $K(1+H)^M e^{-ar^2 H}$ for $H \ge 1$ is at most e^{-H} when $ar^2 \ge 2 + M + \log(K \vee 1)$, since $\log(1+H) \le H$ and $\log(K \vee 1) \le H \log(K \vee 1)$. Any additional fixed power of r in a prefactor is absorbed by increasing r polynomially in $K \vee a^{-1} \vee 1$. Apply this first to $c_\star$ and then to A_{tail}; their powers in the construction are therefore included in the displayed geometry and density factors. Put $H := D \vee \log n$, let q_{geom} be as in (4), and, after this finite order is fixed, take

$$\mathfrak{p}_{\beta,d} = c_{\mathrm{amb}}(q_{\mathrm{geom}} + 2)^2 \Big(1 + d + \frac{(m_\beta + 1)d}{\beta - m_\beta}\Big),$$

with the same universal c_{amb} and the exponent margin specified after (227). 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 $C_{\mathrm{den,sm}} \ge 1$ be a fixed constant large enough to dominate the reciprocal and product constants in the small-noise ratio step. The small-noise reach and denominator-resolution conditions are

$$3c_\star \underline{m}^{-1} \sigma_{t_1} \sqrt{H} \le \eta_\Pi / 2 \tag{236}$$

and

$$C_{\mathrm{den,sm}} p_{\min}^{-1} \Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C n^{-\beta/(d+2\beta)} H^{\mathfrak{p}_{\beta,d}/2 - 1/2} \le 1. \tag{237}$$

For the OU schedule used here, $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$, so (236) is implied by

$$D \vee \log n \le c\, \underline{m}^2 (\eta_\Pi / c_\star)^2\, n^{2/(d+2\beta)}.$$

Corollary F.33 (Score approximation on small noise regime $[t_0, t_1]$)**.** *Assume Assumptions 1 to 3, $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$, and (236) and (237). There exists a raw ReLU comparison score $s_{1,0}$, with postprocessed score s_1, on $[t_0, t_1]$ such that for*

$$\begin{gathered} A_t := \{\mathrm{dist}(\boldsymbol{x}, m_t \mathcal{M}) \le C_\varrho \sigma_t \sqrt{H}\}, \\ \|\sigma_t s_{1,0}(\boldsymbol{x}, t) - \sigma_t s^*(\boldsymbol{x}, t)\|_2 \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}/2}, \\ \boldsymbol{x} \in A_t, \qquad t \in [t_0, t_1]. \end{gathered} \tag{238}$$

or equivalently,

$$\begin{gathered} \|s_{1,0}(\boldsymbol{x}, t) - s^*(\boldsymbol{x}, t)\|_2^2 \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} \sigma_t^{-2}, \\ \boldsymbol{x} \in A_t, \qquad t \in [t_0, t_1]. \end{gathered} \tag{239}$$

For any prescribed fixed $A > 0$, the postprocessed score satisfies, uniformly for $t \in [t_0, t_1]$,

$$\mathbb{E}\|s_1(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} H^{\mathfrak{p}_{\beta,d}} \sigma_t^{-2} + n^{-A} \sigma_t^{-2}. \tag{240}$$

Its clipping and tube constants may depend on A, and

$$\sup_{\boldsymbol{x} \in \mathbb{R}^D} \|s_1(\boldsymbol{x}, t)\|_\infty \le C \sigma_t^{-1} (D \vee \log n)^{1/2}, \qquad t \in [t_0, t_1], \tag{241}$$

Proposition F.34 (Complexity of the small-noise comparison score)**.** *Under the assumptions of Corollary F.33, the underlying raw network may be chosen with*

$$L \vee \log B \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}, \tag{242}$$

and

$$\|\boldsymbol{W}\|_\infty \vee S \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}. \tag{243}$$

Proof of Corollary F.33 and Proposition F.34. Step 1: remove the common Gaussian factor from the chart integrals. On the tube

$$A_t := \{\mathrm{dist}(\boldsymbol{x}, m_t\mathcal{M}) \le C_\varrho \sigma_t \sqrt{H}\},$$

the reach condition (236) gives

$$\mathrm{dist}(\boldsymbol{x}/m_t, \mathcal{M}) \le C_\varrho \frac{\sigma_t}{m_t}\sqrt{H} \le \eta_\Pi/2.$$

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 rescaled denominator and centered numerator by

$$\bar{\mathcal{Q}}(\boldsymbol{x}, t) := \exp\Big(\frac{\|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big)\mathcal{Q}(\boldsymbol{x}, t), \qquad \bar{\mathcal{U}}(\boldsymbol{x}, t) := \exp\Big(\frac{\|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big)\mathcal{U}(\boldsymbol{x}, t). \tag{244}$$

Equivalently, for $R \in \{\mathcal{Q}, \mathcal{P}_1, \ldots, \mathcal{P}_D\}$, define

$$\bar{R}_i(\boldsymbol{x}, t) := \exp\Big(\frac{\|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big) R_i(\boldsymbol{x}, t). \tag{245}$$

Using (49), this removes exactly the common normal Gaussian factor:

$$\bar{R}_i(\boldsymbol{x}, t) = \sigma_t^{-d} \int_{K_i^{\mathrm{cell}}} \exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big) a_i^{(R)}(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}. \tag{246}$$

Thus all exponential factors in the lower-bound small-noise approximation are displayed explicitly: the original chart Gaussian

$$\exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2}{2\sigma_t^2}\Big),$$

the multiplicative factor used to remove the common Gaussian term

$$\exp\Big(\frac{\|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big),$$

and their product

$$\exp\Big(-\frac{\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}(\boldsymbol{x}, t)\|_2^2}{2\sigma_t^2}\Big).$$

Consequently,

$$\bar{\mathcal{Q}} = \sum_i \bar{\mathcal{Q}}_i, \qquad \bar{\mathcal{P}}_k = \sum_i \bar{\mathcal{P}}_{i,k}, \qquad \bar{\mathcal{U}}_k = \frac{m_t \bar{\mathcal{P}}_k - x_k \bar{\mathcal{Q}}}{\sigma_t}. \tag{247}$$

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 rescaled denominator. By Lemma F.30, the density lower bound gives the following floor for the rescaled denominator:

$$\bar{\mathcal{Q}}(\boldsymbol{x}, t) \ge C^{-1} \Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min} (D \vee \log n)^{-d/4}, \qquad (\boldsymbol{x}, t) \in A_t, \quad t \in [t_0, t_1], \tag{248}$$

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 (66), used with $\underline{t} = t_0$ and the assumption $p_{\min}^{-1} \le n^{a_0}$, gives

$$\|\bar{\mathcal{U}}(\boldsymbol{x},t)\|_2 \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} H^{1/2} \bar{\mathcal{Q}}(\boldsymbol{x},t) \qquad ((\boldsymbol{x},t) \in A_t). \tag{249}$$

Step 3: approximate the rescaled 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 $d \ge 1$, $d < D$, and hence $D \ge 2$. Use Lemma F.31 with $\varepsilon_{\mathrm{sm}}$ and $H = D \vee \log n$. The constants fixed in (3) give

$$\sigma_{t_1} \le \sqrt{2t_1} \le 2^{-1/\beta} n^{-1/(d+2\beta)} \le (\tfrac{1}{2} \wedge e\, n^{-\beta/(d+2\beta)})^{1/\beta} = \varepsilon_{\mathrm{sm}}^{1/\beta}.$$

Thus the H-localized lemma applies without imposing the earlier generic Laplace threshold C_{sm}; its geometry dependence is carried by the explicit reach condition and error factors. The reach condition needed by that lemma is exactly (236). This is the point where the rescaled 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

$$\|\boldsymbol{x} - m_t \boldsymbol{z}_i(\boldsymbol{u})\|_2^2 - \|\boldsymbol{r}(\boldsymbol{x},t)\|_2^2 \ge c\sigma_t^2 H$$

on their supports and are exponentially negligible after the rescaling. After summing over charts, this gives networks $\widehat{\bar{\mathcal{Q}}}$ and $\widehat{\bar{\mathcal{U}}}$ satisfying

$$|\widehat{\bar{\mathcal{Q}}}(\boldsymbol{x},t) - \bar{\mathcal{Q}}(\boldsymbol{x},t)| \vee \|\widehat{\bar{\mathcal{U}}}(\boldsymbol{x},t) - \bar{\mathcal{U}}(\boldsymbol{x},t)\|_\infty \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C n^{-\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}/2 - d/4 - 1/2} \tag{250}$$

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 (234) is dominated by $(D \vee \log n)^{\mathfrak{p}_{\beta,d}/2 - d/4 - 1/2}$. With $\varepsilon_{\mathrm{sm}} = \frac{1}{2} \wedge e\, n^{-\beta/(d+2\beta)}$ and $\sigma_{t_0} \asymp n^{-(\beta+1)/(d+2\beta)}$, the two terms in (225) satisfy

$$\varepsilon_{\mathrm{sm}}^{-d/\beta} \lesssim n^{d/(d+2\beta)}, \qquad (\sigma_{t_0}^{-4} \varepsilon_{\mathrm{sm}}^{-2})^{d/q_{\mathrm{geom}}} \lesssim n^{(6\beta+4)d/[q_{\mathrm{geom}}(d+2\beta)]} \le n^{d/(d+2\beta)}. \tag{251}$$

This proves the claimed intrinsic size exponent while retaining the lower-time cost throughout the construction. The extra $\log \underline{t}^{-1}$ from (224) is absorbed into $(D \vee \log n)^{\mathfrak{p}_{\beta,d}}$ because $\underline{t} = t_0$ and (3) gives $\log t_0^{-1} \lesssim \log n$.

Step 4: form the stable ratio. By (248),

$$\bar{\mathcal{Q}}(\boldsymbol{x},t)^{-1} \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min}^{-1} (D \vee \log n)^{d/4} \qquad ((\boldsymbol{x},t) \in A_t).$$

The denominator-resolution condition (237) and (250) 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 $C_{\mathrm{den,sm}}$. Set

$$\lambda_{\mathrm{den}} := \frac{c}{2} \Gamma^{-C}_{\mathcal{M},q_{\mathrm{geom}}} p_{\min} (D \vee \log n)^{-d/4}.$$

Then $\widehat{\bar{\mathcal{Q}}} \vee \lambda_{\mathrm{den}} \ge \lambda_{\mathrm{den}}$. Together with (249) and (250), the standard reciprocal and product networks applied to $\widehat{\bar{\mathcal{U}}}/(\widehat{\bar{\mathcal{Q}}} \vee \lambda_{\mathrm{den}})$ give a network $\widehat{v}_{\mathrm{sm}}^{\mathrm{lb}}$ such that

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

By (66), 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

$$\sigma_t \|s^*(\boldsymbol{x},t)\|_\infty \le \sigma_t \|s^*(\boldsymbol{x},t)\|_2 \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} (D \vee \log n)^{1/2}, \qquad (\boldsymbol{x},t) \in A_t.$$

Clip $\widehat{v}_{\mathrm{sm}}^{\mathrm{lb}}$ at $\mathcal{R}_{\mathrm{sm}} := C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} \sqrt{H}$, approximate σ_t^{-1} by Lemma I.9, and implement their coordinatewise products with total scaled error at most $c\varepsilon_{\mathrm{sm}}$. For example, inverse-schedule accuracy $c\varepsilon_{\mathrm{sm}}/(\sqrt{D}\mathcal{R}_{\mathrm{sm}})$ and product accuracy $c\varepsilon_{\mathrm{sm}}/\sqrt{D}$ suffice because $\sigma_t \le 1$; replace either tolerance by

its minimum with $1/2$ when invoking an arithmetic lemma. This gives the raw ReLU score network $s_{1,0}$ with the asserted tube error. Define s_1 by the prescribed input preprocessing and coordinatewise output clipping at $C\sigma_t^{-1}\sqrt{H}$. The internal clipping at $\mathcal{R}_{\mathrm{sm}}$ fixes the scaled target on A_t, so (238) and (239) hold for $s_{1,0}$. The final clipping radius need not fix the true score pointwise. Instead, Lemma I.11 with $u = A' \log n$ bounds its on-tube mean-square error by twice the raw tube bound plus $8\sigma_t^{-2} n^{-A'}$. The preprocessing is the identity on this tube after undoing the affine input rescaling, as verified in the proof of Proposition G.3. On A_t^c, Lemma C.3 applies to the clipped envelope. Choose $A' > A$ sufficiently large, and increase the fixed clipping and tube constants, to obtain (240). The output clipping itself gives (241).

Step 5: record size. The reciprocal, products, clipping, and schedule subnetworks 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 (242) and (243) 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 postprocessed score classes $\widetilde{\mathcal{NN}}$, the oracle inequality for their approximate minimizers, 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. All expectations involving a learned score network are over the training sample and any optimizer randomness. On a large-noise slab $I_k = [t_{k-1}, t_k)$ (up to the final endpoint), 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 radius required to make the fixed-order tangent-cell expansion uniform also on order-one-noise slabs at the target approximation accuracy. 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 Assumption 3. In all size bounds in this section, the factor $\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}$ may be read as

$$C(1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}_{\mathcal{M},q_{\mathrm{geom}}}^{\mathfrak{a}_{\mathrm{geom}}}$$

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 Theorem E.5 and Corollary F.33 and evaluate their covering-number bounds. Readers interested only in the final rates can read Proposition G.3 and Theorem G.6 first and return to the oracle proof for concentration details.

G.1 ORACLE INEQUALITY FOR SCORE ESTIMATION

Theorem G.1 (Oracle inequality for postprocessed score estimation)**.** *Fix $A > 0$ and $a_{\mathrm{poly}} \geq 1$, and suppose that*

$$n^{-a_{\mathrm{poly}}} \leq \underline{t} < \overline{t} \leq n^{a_{\mathrm{poly}}} \qquad \text{and} \qquad D \leq n^{a_{\mathrm{poly}}}.$$

Let $\widehat{s}$ be a measurable approximate empirical minimizer over the postprocessed 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 DSM loss $\ell_{[\underline{t},\overline{t}]}$, with empirical excess (where $\mathbb{P}_n$ denotes the empirical average)

$$\mathcal{E}_{\mathrm{opt}} := \mathbb{P}_n\ell_{[\underline{t},\overline{t}]}(\widehat{s}) - \inf_{s\in\mathcal{NN}_{\underline{t}}} \mathbb{P}_n\ell_{[\underline{t},\overline{t}]}(s) \geq 0.$$

Then there exists a fixed $A_{\mathrm{cov}} = A_{\mathrm{cov}}(A, a_{\mathrm{poly}}) > 0$ such that the expected excess risk satisfies

$$\begin{aligned}
&\mathbb{E}\Big[\int_{\underline{t}}^{\overline{t}} \mathbb{E}_{\boldsymbol{X}_t}[\|\widehat{s}(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2]\mathrm{d}t\Big] \\
&\lesssim \inf_{s\in\mathcal{NN}_{\underline{t}}} \int_{\underline{t}}^{\overline{t}} \mathbb{E}_{\boldsymbol{X}_t}\big[\|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\big]\mathrm{d}t \\
&\quad + \frac{(D\vee\log n)^2 \int_{\underline{t}}^{\overline{t}} \sigma_t^{-2}\mathrm{d}t \log\big(\mathcal{N}(n^{-A_{\mathrm{cov}}}, \mathcal{NN}_{\underline{t}}, \|\cdot\|_\infty)\vee n\big)}{n} + n^{-A} + C\mathbb{E}\mathcal{E}_{\mathrm{opt}}.
\end{aligned}$$

Proof. Step 1: reduce the DSM risk to the time-integrated squared score error. 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-matching 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 $\mathbb{P}_n \ell_{[\underline{t},\bar{t}]}(s)$ is also an empirical minimizer of $\mathbb{P}_n\{\ell_{[\underline{t},\bar{t}]}(s) - \ell_{[\underline{t},\bar{t}]}(s^*)\}$.

Step 2: record the envelope and variance bounds. The postprocessed definition of $\widetilde{\mathcal{NN}}$ and Lemma I.13 imply

$$\sup_{s,s' \in \mathcal{NN}_{\underline{t}}} \sup_{\boldsymbol{x} \in \mathbb{R}^D} |\ell_{[\underline{t},\bar{t}]}(s,\boldsymbol{x}) - \ell_{[\underline{t},\bar{t}]}(s',\boldsymbol{x})| \le C(D \vee \log n)^2 \int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t.$$

Moreover, applying Lemma I.12 to $s, s' \in \mathcal{NN}_{\underline{t}}$, using the same output envelope, and then averaging over $\boldsymbol{X}_0$, gives

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}_0}\big[\{\ell_{[\underline{t},\bar{t}]}(s,\boldsymbol{X}_0) - \ell_{[\underline{t},\bar{t}]}(s',\boldsymbol{X}_0)\}^2\big] \\
&\le C(D \vee \log n)^2 \Big(\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t\Big) \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t} \|s(\boldsymbol{X}_t,t) - s'(\boldsymbol{X}_t,t)\|_2^2 \mathrm{d}t \\
&\le C(D \vee \log n)^2 \Big(\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t\Big)\{\mathcal{L}_{[\underline{t},\bar{t}]}(s) + \mathcal{L}_{[\underline{t},\bar{t}]}(s')\}. \qquad (252)
\end{aligned}$$

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 $\eta = n^{-A_{\mathrm{cov}}}$, where A_{cov} will be chosen below, 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 every function in the postprocessed class first applies this input truncation, and exact output clipping is nonexpansive, this is also a global η-net for the postprocessed functions on $\mathbb{R}^D \times [\underline{t},\bar{t}]$. For fixed $u, v \in \mathcal{N}_\eta$, Bernstein's inequality and (252) give, with probability at least $1 - e^{-r}$,

$$\begin{aligned}
&|(\mathbb{P}_n - \mathbb{E}_{\boldsymbol{X}_0})\{\ell_{[\underline{t},\bar{t}]}(u) - \ell_{[\underline{t},\bar{t}]}(v)\}| \\
&\le C\Big[\sqrt{\frac{(D \vee \log n)^2 \big(\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t\big)\{\mathcal{L}_{[\underline{t},\bar{t}]}(u) + \mathcal{L}_{[\underline{t},\bar{t}]}(v)\} r}{n}} + \frac{(D \vee \log n)^2 \big(\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t\big) r}{n}\Big].
\end{aligned}$$

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 postprocessed class. If u is an η-net approximation of s, then the pointwise bound in Lemma I.12, together with the clipped envelope, yields

$$\sup_{\boldsymbol{x} \in \mathbb{R}^D} |\ell_{[\underline{t},\bar{t}]}(s,\boldsymbol{x}) - \ell_{[\underline{t},\bar{t}]}(u,\boldsymbol{x})| \le C\sqrt{D(\bar{t} - \underline{t})(D \vee \log n)^2 \int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t}\, \eta.$$

Thus

$$\mathbb{E}_{\boldsymbol{X}_0} |\ell_{[\underline{t},\bar{t}]}(s) - \ell_{[\underline{t},\bar{t}]}(u)| + \mathbb{P}_n |\ell_{[\underline{t},\bar{t}]}(s) - \ell_{[\underline{t},\bar{t}]}(u)| \le n^{-A_2}$$

deterministically for any prescribed $A_2 > 0$ after taking A_{cov} sufficiently large as a function of A_2 and a_{poly}. Indeed, under the displayed polynomial-regime assumptions and the OU schedule, the prefactor multiplying η is bounded by $n^{C a_{\mathrm{poly}}}$ for a universal constant C. Since

$$\mathcal{L}_{[\underline{t},\bar{t}]}(s) - \mathcal{L}_{[\underline{t},\bar{t}]}(u) = \mathbb{E}_{\boldsymbol{X}_0}\{\ell_{[\underline{t},\bar{t}]}(s,\boldsymbol{X}_0) - \ell_{[\underline{t},\bar{t}]}(u,\boldsymbol{X}_0)\},$$

the same estimate also gives $|\mathcal{L}_{[\underline{t},\bar{t}]}(s) - \mathcal{L}_{[\underline{t},\bar{t}]}(u)| \le n^{-A_2}$. Hence the finite-cover Bernstein display applies to arbitrary $s, s' \in \mathcal{NN}_{\underline{t}}$, with $\mathcal{L}_{[\underline{t},\bar{t}]}(u) + \mathcal{L}_{[\underline{t},\bar{t}]}(v)$ replaced by $\mathcal{L}_{[\underline{t},\bar{t}]}(s) + \mathcal{L}_{[\underline{t},\bar{t}]}(s') + n^{-A_2}$, up to another additive n^{-A_2}. Choosing A_2 sufficiently large as a function of A and a_{poly} yields the same display with the covering number evaluated at $n^{-A_{\mathrm{cov}}}$.

For brevity in the remaining displays, set $\mathcal{V}_{[\underline{t},\bar{t}]} := (D \vee \log n)^2 \int_{\underline{t}}^{\bar{t}} \sigma_t^{-2} \mathrm{d}t$.

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,

$$\mathbb{E}_{\boldsymbol{X}_0}\ell_{[\underline{t},\bar{t}]}(\widehat{s},\boldsymbol{X}_0) - \mathbb{E}_{\boldsymbol{X}_0}\ell_{[\underline{t},\bar{t}]}(s_0,\boldsymbol{X}_0) \le (\mathbb{E}_{\boldsymbol{X}_0} - \mathbb{P}_n)\{\ell_{[\underline{t},\bar{t}]}(\widehat{s}) - \ell_{[\underline{t},\bar{t}]}(s_0)\} + n^{-A_2} + \mathcal{E}_{\mathrm{opt}},$$

because $\mathbb{P}_n\ell_{[\underline{t},\bar{t}]}(\widehat{s}) \le \mathbb{P}_n\ell_{[\underline{t},\bar{t}]}(s_0) + \mathcal{E}_{\mathrm{opt}}$. Therefore,

$$\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le \mathcal{L}_{[\underline{t},\bar{t}]}(s_0) + C\left[\sqrt{\frac{\mathcal{V}_{[\underline{t},\bar{t}]}\{\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) + \mathcal{L}_{[\underline{t},\bar{t}]}(s_0)\}r}{n}} + \frac{\mathcal{V}_{[\underline{t},\bar{t}]}r}{n}\right] + n^{-A_2} + \mathcal{E}_{\mathrm{opt}}.$$

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

$$\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le C\mathcal{L}_{[\underline{t},\bar{t}]}(s_0) + C\frac{\mathcal{V}_{[\underline{t},\bar{t}]}r}{n} + n^{-A_2} + C\mathcal{E}_{\mathrm{opt}}.$$

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 Lemma 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

$$\mathbb{E}\mathcal{L}_{[\underline{t},\bar{t}]}(\widehat{s}) \le C\inf_{s\in\mathcal{NN}_{\underline{t}}}\mathcal{L}_{[\underline{t},\bar{t}]}(s) + C\frac{\mathcal{V}_{[\underline{t},\bar{t}]}\log\big(\mathcal{N}(n^{-A_{\mathrm{cov}}},\mathcal{NN}_{\underline{t}},\|\cdot\|_\infty)\vee n\big)}{n} + n^{-A} + C\mathbb{E}\mathcal{E}_{\mathrm{opt}}.$$

which is the claimed oracle inequality. □

Proposition G.2 (Oracle inequality for a weighted time measure)**.** *Let ν be a deterministic finite positive Borel measure on $[\underline{t},\bar{t}]$, possibly atomic, such that*

$$\nu([\underline{t},\bar{t}]) \le n^{a_\nu}, \qquad A_\nu := \int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\nu(\mathrm{d}t) \le n^{a_\nu}$$

for a fixed $a_\nu \ge 1$. Define

$$\mathcal{L}_\nu(s) := \int \mathbb{E}_{\boldsymbol{X}_t}\|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\nu(\mathrm{d}t),$$
$$\ell_\nu(s,\boldsymbol{x}_0) := \int \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}_0}\Big\|s(\boldsymbol{X}_t,t) + \frac{\boldsymbol{X}_t - m_t\boldsymbol{x}_0}{\sigma_t^2}\Big\|_2^2\nu(\mathrm{d}t).$$

Under the remaining assumptions of Theorem G.1, let $\widehat{s}$ be a measurable approximate empirical minimizer of $\mathbb{P}_n\ell_\nu$ over $\mathcal{NN}_{\underline{t}}$, and define its empirical excess by

$$\mathcal{E}_{\mathrm{opt},\nu} := \mathbb{P}_n\ell_\nu(\widehat{s}) - \inf_{s\in\mathcal{NN}_{\underline{t}}}\mathbb{P}_n\ell_\nu(s) \ge 0.$$

Then, for every fixed $A > 0$, there exists a fixed $A_{\mathrm{cov},\nu} = A_{\mathrm{cov},\nu}(A, a_{\mathrm{poly}}, a_\nu) > 0$ such that

$$\mathbb{E}\mathcal{L}_\nu(\widehat{s}) \le C\inf_{s\in\mathcal{NN}_{\underline{t}}}\mathcal{L}_\nu(s) + C\frac{(D\vee\log n)^2 A_\nu}{n}\log\Big(\mathcal{N}(n^{-A_{\mathrm{cov},\nu}},\mathcal{NN}_{\underline{t}},\|\cdot\|_\infty)\vee n\Big) + n^{-A} + C\mathbb{E}\mathcal{E}_{\mathrm{opt},\nu}. \tag{253}$$

Proof. The denoising score-matching identity holds for each $t > 0$ and hence after integration against the positive measure ν. Thus $\mathcal{L}_\nu(s) = \mathbb{E}\{\ell_\nu(s,\boldsymbol{X}_0) - \ell_\nu(s^*,\boldsymbol{X}_0)\}$. The clipping envelope and the conditional Gaussian identity

$$\mathbb{E}\Big[\Big\|\frac{\boldsymbol{X}_t - m_t\boldsymbol{X}_0}{\sigma_t^2}\Big\|_2^2\Big|\boldsymbol{X}_0\Big] = D\sigma_t^{-2}$$

give, for $s, s' \in \mathcal{NN}_{\underline{t}}$,

$$\sup_{\boldsymbol{x}_0} |\ell_\nu(s, \boldsymbol{x}_0) - \ell_\nu(s', \boldsymbol{x}_0)| \le C(D \vee \log n)^2 A_\nu.$$

Applying Cauchy–Schwarz first in the conditional Gaussian expectation and then in $L^2(\nu)$, exactly as in Lemma I.12, yields

$$\mathbb{E}_{\boldsymbol{X}_0}\Big[\{\ell_\nu(s, \boldsymbol{X}_0) - \ell_\nu(s', \boldsymbol{X}_0)\}^2\Big] \le C(D \vee \log n)^2 A_\nu \{\mathcal{L}_\nu(s) + \mathcal{L}_\nu(s')\}.$$

These are precisely the envelope and Bernstein-moment inputs used in Steps 3–6 of the proof of Theorem G.1. The cover-transfer remainder is at most $C\eta\{D\nu([\underline{t}, \bar{t}])(D \vee \log n)^2 A_\nu\}^{1/2}$ for a supremum-norm η-cover, and is n^{-A} after taking $\eta = n^{-A_{\mathrm{cov},\nu}}$ with $A_{\mathrm{cov},\nu}$ sufficiently large as a function of A, a_{poly}, and a_ν. Repeating the Bernstein, basic-inequality, absorption, and exceptional-event steps proves (253). □

G.2 Time-Grid and Integral Conventions

For the OU schedule 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

$$\begin{gathered}[t_1, T] = \bigcup_{k=2}^{K} I_k, \qquad I_k = [t_{k-1}, t_k)\ (2 \le k < K), \quad I_K = [t_{K-1}, T], \\ t_k = 2t_{k-1} \text{ while } t_k \le 1, \qquad K \le C \log n,\end{gathered} \tag{254}$$

where the dyadic portion is followed by $\mathcal{O}(\log n)$ order-one slabs of uniformly bounded length from time 1 to the terminal cutoff in (3). 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 under (3),

$$\int_{t_0}^{t_1} \sigma_t^{-2} \mathrm{d}t \le C \log(t_1/t_0) \le C \log n. \tag{255}$$

For the auxiliary time-discretization analysis, choose a reverse-time grid $0 = u_0 < \cdots < u_M = T - t_0$ containing every time-interval boundary $T - t_k$. Write

$$\tau_j := T - u_j, \qquad \Delta_j := u_{j+1} - u_j = \tau_j - \tau_{j+1},$$

and let $k(j)$ be the unique slab satisfying $[\tau_{j+1}, \tau_j] \subseteq [t_{k(j)-1}, t_{k(j)}]$. For $\eta_{\mathrm{step}} \in (0, 1/2]$, require

$$\Delta_j \le \eta_{\mathrm{step}} \min\{1, \tau_{j+1}\}, \qquad 0 \le j < M. \tag{256}$$

For each slab, set

$$J_k := \{j : k(j) = k\}, \qquad \bar{I}_k := [t_{k-1}, t_k], \qquad \nu_{k,M} := \sum_{j \in J_k} \Delta_j \delta_{\tau_j}.$$

Since $t \mapsto \sigma_t^{-2}$ is decreasing,

$$\sum_{j \in J_k} \Delta_j \sigma_{\tau_j}^{-2} \le \int_{t_{k-1}}^{t_k} \sigma_t^{-2} \mathrm{d}t. \tag{257}$$

Consequently, the left side is at most $C \log n$ for $k = 1$ and at most a constant for every $k \ge 2$. Also $\nu_{k,M}(\bar{I}_k) = \sum_{j \in J_k} \Delta_j = t_k - t_{k-1}$.

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 class complexity, and stochastic term are checked for the same large-noise class.

Large-noise class and cell-radius cap. Let I_k be one of the half-open large-noise slabs in $[t_1, T]$ satisfying (254). Recall ε_n from (2) and $q_\star$ from (4), and define the tangent-cell radius cap

$$r_{\star,n} := c_{\mathrm{cap}} \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{-C_\star} B_0^{-C_\star} p_{\min}^{C_\star} (D \vee \log n)^{-2} \varepsilon_n^{1/q_\star},$$

where $C_\star$ is sufficiently large and c_{cap} sufficiently small, depending only on (β, d) and the fixed numerical constants in Theorem E.5. The exponent $C_\star$ controls only geometry and density factors; the ambient and logarithmic exponent is fixed at 2, as verified below. In particular,

$$r_{\star,n}^{-d} \le C \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C} p_{\min}^{-C} (D \vee \log n)^{2d} \varepsilon_n^{-d/q_\star}.$$

Let $\mathcal{F}_k$ be the postprocessed score class on I_k, $k \ge 2$, with the tangent-cell raw-network parameter bounds in (259) and coordinatewise clipping at $C\sigma_t^{-1}(D \vee \log n)^{1/2}$. Choose its architecture budgets as in (235); the looser bounds in (259) summarize those budgets, rather than allowing arbitrary enlargement of the estimator class. On the global grid used in the main theorem, $\sigma_{t_{k-1}}^{-1} \le C\sigma_{t_1}^{-1} \lesssim n^{1/(d+2\beta)}$; hence its logarithmic contribution is polylogarithmic in n.

Proposition G.3 (Score estimation on large-noise regime)**.** *Assume Assumptions 1 to 3 and $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$. Fix $A > 0$; the fixed localization constants may depend on A. Let $\widehat{s}_k$ be a measurable approximate empirical minimizer over $\mathcal{F}_k$ on I_k, with empirical excess $\mathcal{E}_{\mathrm{opt},k}$ as in (7). Then*

$$\begin{aligned} &\mathbb{E} \int_{I_k} \mathbb{E}_{\boldsymbol{X}_t} \|\widehat{s}_k(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t \\ &\le C \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C} p_{\min}^{-C} \Big[n^{-2(\beta+1)/(d+2\beta)} + \frac{\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}}{n} \Big] (D \vee \log n)^{\mathfrak{p}_{\beta,d}} + n^{-A} + C\mathbb{E}\mathcal{E}_{\mathrm{opt},k}. \end{aligned} \tag{258}$$

Corollary G.4 (Capacity of the large-noise score class)**.** *Under the assumptions and construction of Proposition G.3, 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*

$$\begin{aligned} L_k \vee \log B_k &\le C \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}, \\ \|\boldsymbol{W}_k\|_\infty \vee S_k &\le C \Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C} p_{\min}^{-C} \big(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\big) (D \vee \log n)^{\mathfrak{p}_{\beta,d}}. \end{aligned} \tag{259}$$

Proof of Proposition G.3 and Corollary G.4. Step 1: construct a comparison network in the slab-local class. We first record the common input-coordinate conversion. If g is a comparison network on the physical input box, define its raw representative on $[0,1]^{D+1}$ by

$$f(\boldsymbol{z}) = g(2B_x \boldsymbol{z}_{1:D} - B_x \mathbf{1},\ \underline{t} + (\bar{t} - \underline{t}) z_{D+1}). \tag{260}$$

Absorb this affine map into the first layer. Only existing first-layer edges and at most one bias per active first-layer neuron are needed, so $S_f \le 2S_g$, the depth and widths are unchanged, and $B_f \le C(D+1)(1 + B_x + \bar{t})(1 + B_g)$. Thus $\log B_f \le \log(1 + B_g) + C \log(e + D + B_x + \bar{t})$. The preprocessing map sends every point of A_t to its inverse image under this affine map. This establishes comparison-class membership, with the same polynomial size and logarithmic weight bounds, for both noise regimes. 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 $\varepsilon_{\mathrm{sc}} = \varepsilon_n / \sqrt{D}$. Take $c_0 \le 1$; monotonicity of the OU noise schedule gives $r_k \le \sigma_{t_{k-1}} \le \sigma_t$ throughout I_k, as required by the tangent-cell theorem independently of the generic Laplace thresholds. Write $H_n = D \vee \log n$ and $G_\star = C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C} B_0^{C}$, enlarging its fixed powers when necessary. The definitions of q_{lb} and R_{lb}, the bounds $\log p_{\min}^{-1} \le a_0 \log n$ and $\log \sigma_{t_{k-1}}^{-1} \le C \log n$, and $\varepsilon_{\mathrm{sc}} = \varepsilon_n / \sqrt{D}$ give

$$1 + R_{\mathrm{lb}} \le G_\star H_n^{1/2}, \qquad H_{\mathrm{lg}}^{\mathrm{lb}} \le G_\star H_n, \qquad \varepsilon_{\mathrm{rel}} \ge c G_\star^{-1} \varepsilon_n H_n^{-1}. \tag{261}$$

For example, $\log q_{\mathrm{lb}}^{-1} \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} H_n + a_0 \log n$, and all other factors in $H_{\mathrm{lg}}^{\mathrm{lb}}$ enter logarithmically. Choose $C_\star$ and c_{cap} to absorb the geometry and density factors in these inequalities. Since $q_\star = q_{\mathrm{geom}} - 2 \ge 1$, the cap then gives

$$\Gamma_{\mathcal{M},q_{\mathrm{geom}}} r_k (H_{\mathrm{lg}}^{\mathrm{lb}})^{1/2} \le c_{\mathrm{ph,cell}} H_n^{-3/2} \le c_{\mathrm{ph,cell}},$$
$$\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} r_k^{q_\star} H_{\mathrm{lg}}^{\mathrm{lb}} \le c G_\star^{-1} \varepsilon_n H_n^{1-2q_\star} \le c\varepsilon_{\mathrm{rel}}.$$

This verifies both conditions in (67), with no dimension-dependent ambient exponent in the cap. Bounded n are absorbed by increasing constants. Thus the geometry condition in (67) is satisfied, and (68) gives a network $s_{k,0}$ such that, on the tube A_t defined in (6),

$$\|s_{k,0}(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_2^2 \le C\sigma_t^{-2} n^{-2(\beta+1)/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

After adding coordinatewise clipping at $C\sigma_t^{-1}(D \vee \log n)^{1/2}$, and enlarging constants if needed, we obtain $s_k \in \mathcal{F}_k$ with the mean-square tube error controlled below and with

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|s_k(\boldsymbol{x},t)\|_\infty \le C\sigma_t^{-1}(D \vee \log n)^{1/2}.$$

Apply Lemma I.11 with $u = A' \log n$ and $R = C\sqrt{D \vee \log n}$, choosing $C^2 \ge 2\log 2 + 4A'$. At every time $t \in I_k$, the mean-square error of s_k on A_t is at most twice the displayed raw approximation bound plus $8\sigma_t^{-2} n^{-A'}$. This does not require a geometry-independent pointwise bound on the true score. Its size is bounded by (69). The identity

$$(r_k^\beta)^{-d/\beta} = r_k^{-d} \le C\big(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\big),$$

and the absorption of $H_{\mathrm{lg}}^{\mathrm{lb}}$ and R_{lb} into the displayed geometry, density, and $(D \vee \log n)^{\mathfrak{p}_{\beta,d}}$ factors shows that the resulting class may be chosen with the parameter bounds displayed in (259).

Step 2: integrate the comparison loss. On A_t^c, apply Lemma C.3 with $\ell_{\mathrm{off}} = D \vee \log n$; the clipping bound for s_k is then within the required coordinatewise envelope. Thus, for any prescribed fixed $A' > 0$, after increasing C_ϱ,

$$\int_{A_t^c} \|s_k(\boldsymbol{x},t) - s^*(\boldsymbol{x},t)\|_2^2 p_t(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \le C\sigma_t^{-2}(D \vee \log n)^{\mathfrak{p}_{\beta,d}} n^{-A'}.$$

Taking A' large enough and using the polynomial lower bound on σ_t, the off-tube contribution is absorbed into the tube rate. Therefore

$$\int_{I_k} \mathbb{E}_{\boldsymbol{X}_t} \|s_k(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2 \mathrm{d}t \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2(\beta+1)/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

where we used $\int_{I_k} \sigma_t^{-2} \mathrm{d}t \le C_{\mathrm{grid}}$. Hence

$$\inf_{s\in\mathcal{F}_k} \int_{I_k} \mathbb{E}_{\boldsymbol{X}_t} \|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2 \mathrm{d}t \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2(\beta+1)/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

Step 3: evaluate the slab-local covering number. Apply the postprocessed score-estimation oracle Theorem G.1 on the slab I_k. Let A_{cov} be the fixed covering exponent supplied by that theorem for the desired remainder n^{-A} and the polynomial regime implied by the present assumptions. Substitute the construction-level bounds (235) into (8), retaining the exponent margin proved in Lemma F.32. Here $L_k + \log B_k$ is polylogarithmic up to the displayed polynomial prefactors, while $S_k + \|\boldsymbol{W}_k\|_\infty$ has leading cell-count factor $\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}$. Therefore

$$\log\big(\mathcal{N}(n^{-A_{\mathrm{cov}}}, \mathcal{F}_k, \|\cdot\|_\infty) \vee n\big) \le C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \big(\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}\big)(D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

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

$$C\Gamma^{C}_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \frac{\sigma_{t_{k-1}}^{-d} + r_{\star,n}^{-d}}{n} (D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

with all manifold and smoothness prefactors displayed. Combining this with the comparison loss from Step 2 proves (258). □

Corollary G.5 (Postprocessed comparison score on a large-noise slab)**.** *Under the assumptions of Proposition G.3, the class $\mathcal{F}_k$ contains a deterministic comparison score s_k such that, uniformly for $t \in I_k$,*

$$\mathbb{E}_{\boldsymbol{X}_t}\|s_k(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2 \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C}\sigma_t^{-2}\varepsilon_n^2 (D\vee\log n)^{\mathfrak{p}_{\beta,d}}. \tag{262}$$

The underlying raw network may be chosen with the parameter bounds in (259).

Proof. The score s_k constructed in Steps 1–2 of the proof of Proposition G.3 belongs to $\mathcal{F}_k$ and satisfies the asserted raw-network bounds. On A_t, Lemma I.11 bounds its mean-square error by twice the raw pointwise approximation error in (68) plus $8\sigma_t^{-2}n^{-A'}$. On A_t^c, Lemma C.3 gives the bound in Step 2. Choose A' sufficiently large and absorb these polynomially small remainders into the leading term, as in that proof. □

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.6 (Score estimation on small-noise regime)**.** *Assume Assumptions 1 to 3 and $D\vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$. Assume also the small-noise reach and denominator-resolution conditions (236) and (237). Fix $A > 0$. Let $\mathcal{F}_1$ be the postprocessed projection–Laplace score class on $[t_0, t_1]$, with clipping radius $C\sigma_t^{-1}(D\vee\log n)^{1/2}$ and the size bounds in (242) and (243), and let $\widehat{s}_1$ be a measurable approximate empirical minimizer over $\mathcal{F}_1$, with empirical excess $\mathcal{E}_{\mathrm{opt},1}$ as in (7). As in the large-noise class, fix the architecture budgets using (235), before relaxing their bounds to the common exponent $\mathfrak{p}_{\beta,d}$. Then there is a fixed $A_{\mathrm{cov}} > 0$, supplied by Theorem G.1 for the resulting polynomial regime, such that*

$$\begin{aligned}
&\mathbb{E}\int_{t_0}^{t_1}\mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}_1(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \\
&\quad\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}} \\
&\qquad + C\frac{(D\vee\log n)^C\int_{t_0}^{t_1}\sigma_t^{-2}\mathrm{d}t\log\big(\mathcal{N}(n^{-A_{\mathrm{cov}}},\mathcal{F}_1,\|\cdot\|_\infty)\vee n\big)}{n} + n^{-A} + C\mathbb{E}\mathcal{E}_{\mathrm{opt},1}.
\end{aligned} \tag{263}$$

Corollary G.7 (Evaluated small-noise score-estimation rate)**.** *For every fixed $A > 0$, under the assumptions of Theorem G.6, the construction-level capacity and exponent margins in Lemma F.32, together with (8), give*

$$\begin{aligned}
&\mathbb{E}\int_{t_0}^{t_1}\mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}_1(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \\
&\le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}} + n^{-A} + C\mathbb{E}\mathcal{E}_{\mathrm{opt},1}.
\end{aligned} \tag{264}$$

Proof of Theorem G.6 and Corollary G.7. The comparison network s_1 belongs to $\mathcal{F}_1$, after the same coordinatewise clipping at radius $C\sigma_t^{-1}(D\vee\log n)^{1/2}$. Here its raw representative is the affine reparametrization in (260). Let A_t be the tube defined in (6). The raw tube bound (239) and the clipping lemma Lemma I.11 give, for any fixed sufficiently large $A' > 0$,

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_{A_t}(\boldsymbol{X}_t)\|s_1(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2] &\le 8n^{-A'}\sigma_t^{-2} \\
&\quad + C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}\sigma_t^{-2}.
\end{aligned}$$

The single-slab integral bound (255) 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

$$C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}.$$

On A_t^c, apply Lemma C.3 with $\ell_{\mathrm{off}} = D\vee\log n$; the clipping bound (241) satisfies the required envelope. After increasing C_ϱ, this gives n^{-A}. Thus

$$\inf_{s\in\mathcal{F}_1}\int_{t_0}^{t_1}\mathbb{E}_{\boldsymbol{X}_t}\|s(\boldsymbol{X}_t,t) - s^*(\boldsymbol{X}_t,t)\|_2^2\mathrm{d}t \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)}(D\vee\log n)^{\mathfrak{p}_{\beta,d}} + n^{-A}.$$

Applying the postprocessed oracle inequality Theorem G.1 to the class $\mathcal{F}_1$ gives (263). The covering-number bound follows from (8) and (235), with the margins in Lemma F.32; it is bounded by

$$C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{d/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

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 (264). □

For the grid measure $\nu_{k,M}$ above, define the grid-weighted squared score error and conditional DSM loss by

$$\mathcal{L}_{k,M}(s) := \sum_{j \in J_k} \Delta_j \mathbb{E} \| s(\boldsymbol{X}_{\tau_j}, \tau_j) - s^*(\boldsymbol{X}_{\tau_j}, \tau_j) \|_2^2, \tag{265}$$

$$\ell_{k,M}(s, \boldsymbol{x}_0) := \sum_{j \in J_k} \Delta_j \mathbb{E}_{\boldsymbol{X}_{\tau_j} | \boldsymbol{X}_0 = \boldsymbol{x}_0} \left\| s(\boldsymbol{X}_{\tau_j}, \tau_j) + \frac{\boldsymbol{X}_{\tau_j} - m_{\tau_j} \boldsymbol{x}_0}{\sigma_{\tau_j}^2} \right\|_2^2. \tag{266}$$

For this discrete objective, $\mathcal{F}_k$ denotes the same postprocessed network class evaluated on the closed slab $\overline{I}_k$; when $\tau_j = t_k$, the value is taken from the slab selected by $k(j)$. Thus every grid atom has a unique prescribed score network even at a time-interval boundary. The pointwise denoising score-matching identity, summed with weights Δ_j, shows that the population DSM excess equals $\mathcal{L}_{k,M}(s)$. Let $\widehat{s}_{k,M}$ be a measurable approximate empirical minimizer over the same postprocessed class $\mathcal{F}_k$, and define

$$\mathcal{E}^{\mathrm{grid}}_{\mathrm{opt},k} := \mathbb{P}_n \ell_{k,M}(\widehat{s}_{k,M}) - \inf_{s \in \mathcal{F}_k} \mathbb{P}_n \ell_{k,M}(s) \ge 0. \tag{267}$$

Thus grid weighting replaces only the diffusion-time integral by deterministic quadrature; each Gaussian-corruption expectation remains exact.

Corollary G.8 (Grid-weighted score-estimation rates)**.** *Use the aligned grid and the grid-weighted approximate ERM defined in (265) to (267). For every fixed $A > 0$, under the geometric, density, polynomial-growth, and small-noise assumptions of Proposition G.3 and Theorem G.6, the same postprocessed classes satisfy*

$$\mathbb{E}\mathcal{L}_{1,M}(\widehat{s}_{1,M}) \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} n^{-2\beta/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}} + n^{-A} + C\mathbb{E}\mathcal{E}^{\mathrm{grid}}_{\mathrm{opt},1}, \tag{268}$$

$$\begin{aligned} \mathbb{E}\mathcal{L}_{k,M}(\widehat{s}_{k,M}) \le{}& C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \left[\varepsilon_n^2 + \frac{\sigma_{t_{k-1}}^{-d} + \varepsilon_n^{-d/q_\star}}{n} \right] (D \vee \log n)^{\mathfrak{p}_{\beta,d}} \\ &+ n^{-A} + C\mathbb{E}\mathcal{E}^{\mathrm{grid}}_{\mathrm{opt},k}, \qquad 2 \le k \le K. \end{aligned} \tag{269}$$

Proof. The mean-square comparison bounds in the proofs of Proposition G.3 and Theorem G.6 hold at every time in the corresponding slab. They include the raw on-tube approximation error, the clipping bias from Lemma I.11, and the off-tube error, all with the same σ_t^{-2} time envelope. Summing these expected-error bounds with weights Δ_j and using (257) gives the same comparison risks as in the continuous-time proofs. Apply Proposition G.2 with $\nu = \nu_{k,M}$; the covering-number bounds are unchanged, while (257) gives the same A_ν bounds as the corresponding time integrals. This proves (268) and (269). □

H W_1 Distribution Estimation

This appendix section contains the formal W_1 distribution-estimation results. The simplified main theorem is stated in Theorem 2.2, and its four-term decomposition is explained 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 noise-regime-specific density-lower-bound learned-score inputs and the off-tube localization bounds.

All distribution-estimation statements in this section use the OU schedule

$$\begin{aligned} m_t &= e^{-t}, \qquad \sigma_t^2 = 1 - e^{-2t} \quad (t \ge 0), \\ c_\sigma t \le \sigma_t^2 &\le C_\sigma t, \qquad 1 - m_t \le t \qquad (0 \le t \le 1). \end{aligned} \tag{270}$$

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 Corollary H.1. Its proof combines early stopping, terminal Gaussian initialization, and the small- and large-noise score errors, including their optimization contributions. The only distribution-level input is the slabwise perturbation lemma Lemma I.18; all score-estimation inputs come from Proposition G.3 and Theorem G.6. The time-discretized counterpart is Corollary H.13; it combines the grid-weighted score rates in Corollary G.8 with the direct path-space comparison in Proposition H.12.

H.1 Conditions for the distributional rate

This subsection expands the non-asymptotic regime conditions invoked in Theorem 2.2. The support and density satisfy Assumptions 1 to 3, and $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$. We use the OU schedule (270), the cutoffs (3), and the measurable slabwise approximate ERM from Sec. 4 and G, followed by projection of the generated endpoint onto $[0,1]^D$. The simplified rate treats the displayed geometry and density factors as fixed; their allowed growth is quantified in Corollary H.6.

Two further inequalities control the small-noise construction: (236) keeps the forward tube within the region of unique nearest-point projection, and (237) ensures sufficient accuracy in approximating the rescaled denominator before taking its reciprocal.

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 disjoint slabwise partition throughout this subsection: after choosing t_0, t_1, T, write

$$[t_0, T] = \bigcup_{k=1}^{K} I_k, \qquad I_1 = [t_0, t_1), \quad I_k = [t_{k-1}, t_k) \ (2 \le k < K), \quad I_K = [t_{K-1}, T],$$

where the large-noise slabs satisfy (254); 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.

Rate aggregation. Corollary 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 $\varepsilon_n^{-d/q_\star}$, with ε_n and $q_\star$ from (2) and (4), is carried explicitly through the W_1 perturbation sum. The following bound retains the early-stopping, initialization, and optimization terms explicitly; its proof gives the small-noise conversion in (274) and the large-noise sum in (276).

Corollary H.1 (Formal W_1 distribution-estimation bound)**.** *Assume Assumptions 1 to 3, $D \vee p_{\min}^{-1} \le n^{a_0}$ for a fixed $a_0 > 0$, and the two small-noise conditions (236) and (237). Fix q_{geom} satisfying (4) and put $q_\star = q_{\mathrm{geom}} - 2$. Use the cutoff schedule (3), ε_n from (2), and the finite ambient exponent*

$\mathfrak{p}_{\beta,d}$ *from* (227). *Let* $\widehat{s}$ *be the piecewise-in-time score estimator that uses* $\widehat{s}_1$ *from Theorem G.6 on* I_1 *and* $\widehat{s}_k$ *from Proposition G.3 on* I_k, $k = 2, \dots, K$, *with empirical excesses* $\mathcal{E}_{\mathrm{opt},k}$ *defined in* (7). *Let* $\widehat{P}_{t_0}$ *be the conditional generated law driven by* $\widehat{s}$, *and set* $\widehat{P}^{\mathrm{box}}_{t_0} := (\mathrm{Proj}_{\mathcal{C}})_{\#}\widehat{P}_{t_0}$, $\mathcal{C} = [0,1]^D$. *Then there are constants* $C, C_{\mathrm{opt}} < \infty$ *and finite nonnegative exponents* c_Γ, c_B, c_p, *all independent of* (n, D), *such that*

$$\begin{aligned}\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{box}}_{t_0}) \le{}& C\sqrt{Dt_0} + CDe^{-T} \\ &+ C\Gamma^{c_\Gamma}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p}\Big(\varepsilon_n + n^{-1/2} + n^{-1/2}\varepsilon_n^{-d/(2q_\star)}\Big)(D \vee \log n)^{\mathfrak{p}_{\beta,d}} \\ &+ C_{\mathrm{opt}}(D \vee \log n)\sum_{k=1}^{K}\sqrt{t_k \log n}\sqrt{\mathbb{E}\mathcal{E}_{\mathrm{opt},k}}.\end{aligned} \tag{271}$$

The expectation is over the data and optimizer randomness, and C_{opt} *depends only on fixed OU and clipping constants.*

Corollary H.2 (Formal simplified intrinsic-rate bound)**.** *Under the assumptions and construction of Corollary H.1, suppose the uniform empirical-excess condition*

$$\max_{1\le k\le K}\sqrt{\mathbb{E}\mathcal{E}_{\mathrm{opt},k}} \lesssim_{\log} \varepsilon_n. \tag{272}$$

holds. Since $K \lesssim \log n$ *and* $t_k \le T \lesssim \log n$, *the last line of equation 271 is* $\lesssim_{\log} (D \vee \log n)\varepsilon_n$ *and is absorbed by the displayed statistical term. For fixed geometry and density factors, the uniform guarantee* $\mathbb{E}\mathcal{E}_{\mathrm{opt},k} \le n^{-A_{\mathrm{ERM}}}$ *is sufficient whenever* $A_{\mathrm{ERM}} > 2(\beta+1)/(d+2\beta)$ *by a fixed margin. If the constant* c_T *in* (3) *exceeds* $a_0 + (\beta+1)/(d+2\beta)$ *by a fixed margin, the terminal term is absorbed by the third term in equation 271. Since* $t_0 \asymp \varepsilon_n^2$ *and* $\mathfrak{p}_{\beta,d} \ge 1$, *the early-stopping term is also absorbed by that term. If* $d > 2$ *and* $q_\star \ge d(\beta+1)/(d-2)$, *then* $n^{-1/2} + n^{-1/2}\varepsilon_n^{-d/(2q_\star)} \le C\varepsilon_n$, *and hence*

$$\mathbb{E}\,\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{box}}_{t_0}) \lesssim_{\log} \Gamma^{c_\Gamma}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p}\varepsilon_n (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

Proposition H.3 (Aggregate complexity of the piecewise-in-time score estimator)**.** *Under the assumptions and construction of Corollary H.1, write* $L_{\max} := \max_k L_k$, $\log B_{\max} := \max_k \log B_k$, $W_{\mathrm{agg}} := \sum_k \|\boldsymbol{W}_k\|_\infty$, *and* $S_{\mathrm{agg}} := \sum_k S_k$. *These quantities are, respectively, the maximum depth, maximum log-weight bound, aggregate width, and aggregate sparsity of the interval-specific score networks. The prescribed time-interval boundaries are not learned parameters. Then*

$$\begin{aligned}L_{\max} \vee \log B_{\max} &\le C\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p}(D \vee \log n)^{\mathfrak{p}_{\beta,d}}, \\ W_{\mathrm{agg}} \vee S_{\mathrm{agg}} &\le C\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p} n^{d/(d+2\beta)}(D \vee \log n)^{\mathfrak{p}_{\beta,d}},\end{aligned}$$

where

$$\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} := (1 \vee S_{\mathcal{M}})^{\mathfrak{a}_{\mathrm{geom}}}(1 \vee \kappa^{-1})^{\mathfrak{a}_{\mathrm{geom}}}\mathcal{K}^{\mathfrak{a}_{\mathrm{geom}}}_{\mathcal{M},q_{\mathrm{geom}}}.$$

One may take

$$\mathfrak{p}_{\beta,d} = c_{\mathrm{amb}}(q_{\mathrm{geom}} + 2)^2\Big(1 + d + \frac{(m_\beta+1)d}{\beta - m_\beta}\Big).$$

The finite exponent constant $c_{\mathrm{amb}} \ge 1$ *is universal but is not assigned an explicit numerical value; it is independent of* $\beta, d, D, n, p_{\min}$, *and the numerical geometry constants. Likewise,* c_Γ, c_B, c_p *are finite but not numerically specified, so the theorem records polynomial dependence without claiming optimal polynomial exponents or constants. Under the above* $d > 2$ *finite-order choice,* $\mathfrak{p}_{\beta,d} = \mathcal{O}_\beta(d)$.

Proof of Corollaries H.1 and H.2 and Proposition H.3. Step 1: decompose the cube-projected distribution error. Use the half-open partition from Sec. 4. For $0 \le k \le K$, let $\overline{P}^{(k)}_{t_0}$ be the exact-terminal hybrid from Lemma I.18; thus $\overline{P}^{(0)}_{t_0} = P_{t_0}$, whereas $\overline{P}^{(K)}_{t_0}$ uses every interval-specific learned score network and is initialized from P_T. The generated law $\widehat{P}_{t_0}$ uses the same score networks but is initialized from $\mathcal{N}(0, I_D)$. Since P_0 is fixed by $\mathrm{Proj}_{\mathcal{C}}$, the triangle inequality gives

$$\begin{aligned}\mathbb{E}\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{box}}_{t_0}) \le{}& \mathsf{W}_1\big(P_0, (\mathrm{Proj}_{\mathcal{C}})_{\#}P_{t_0}\big) + \sum_{k=1}^{K}\mathbb{E}\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(k-1)}_{t_0}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(k)}_{t_0}\big) \\ &+ \mathbb{E}\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(K)}_{t_0}, \widehat{P}^{\mathrm{box}}_{t_0}\big).\end{aligned} \tag{273}$$

The cube-projected early-stopping and terminal-initialization lemmas Lemmas I.16 and I.17 bound the first and last terms by

$$C\sqrt{Dt_0} \qquad \text{and} \qquad CDe^{-T},$$

respectively. These bounds hold conditionally on the learned score networks where applicable and therefore also after averaging over the data and optimizer randomness.

Step 2: bound the small-noise contribution. Each interval-specific score network is clipped coordinatewise, so the piecewise estimator satisfies

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|\widehat{s}(\boldsymbol{x},t)\|_\infty \le C\sigma_t^{-1}(D\vee\log n)^{1/2}, \qquad t\in[t_0,T].$$

Apply Lemma I.18 with $k=1$, then use (264) and $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$. After enlarging the finite ambient and geometry exponents and choosing the negligible exponent A sufficiently large,

$$\begin{aligned}&\mathbb{E}\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(0)},(\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(1)}\big)\\&\quad\le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{c_\Gamma}B_0^{c_B}p_{\min}^{-c_p}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}\varepsilon_n + C_{\mathrm{opt}}(D\vee\log n)\sqrt{t_1\log n}\sqrt{\mathbb{E}\mathcal{E}_{\mathrm{opt},1}}.\end{aligned}\tag{274}$$

Indeed, the statistical part uses $\sqrt{t_1}\,n^{-\beta/(d+2\beta)}=\varepsilon_n$.

Step 3: bound and sum the large-noise contributions. For $k\ge 2$, Lemma I.18 and (258) yield

$$\begin{aligned}&\mathbb{E}\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k-1)},(\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k)}\big)\\&\quad\le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{c_\Gamma}B_0^{c_B}p_{\min}^{-c_p}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}\sqrt{t_k\log n}\Big[\varepsilon_n+n^{-1/2}\sigma_{t_{k-1}}^{-d/2}+n^{-1/2}\varepsilon_n^{-d/(2q_\star)}\Big]\\&\qquad+C_{\mathrm{opt}}(D\vee\log n)\sqrt{t_k\log n}\sqrt{\mathbb{E}\mathcal{E}_{\mathrm{opt},k}}+n^{-A}.\end{aligned}\tag{275}$$

All powers arising from the transfer factor and square roots of the oracle prefactors are included in the displayed finite exponents. Because $K\lesssim\log n$ and $t_k\le T\lesssim\log n$, summing the ε_n part changes only the displayed ambient/logarithmic power. For the stochastic cell-count term, the dyadic grid and $\sigma_t^2\asymp t$ for $t\le 1$ give

$$n^{-1/2}\sum_{k\ge2:\,t_k\le1}\sqrt{t_k}\,\sigma_{t_{k-1}}^{-d/2}\le Cn^{-1/2}\begin{cases}1, & d<2,\\ 1+\log(1/t_1), & d=2,\\ \sigma_{t_1}^{1-d/2}, & d>2.\end{cases}$$

Indeed, the summands are comparable to $t_{k-1}^{(2-d)/4}$, giving a geometric series except at $d=2$. For $d>2$, substitution of $\sigma_{t_1}\asymp n^{-1/(d+2\beta)}$ gives $n^{-1/2}\sigma_{t_1}^{1-d/2}\asymp\varepsilon_n$. For $d=2$, the additional factor is at most $C\log n$; for $d<2$, the bound is $Cn^{-1/2}$. Thus the sum is at most $C(\varepsilon_n+n^{-1/2})\log n$ for every $d\ge1$, and the logarithm is included in the exponent margin of (227). The order-one-time slabs contribute $n^{-1/2}$ up to logarithmic factors, and the accuracy-cap terms sum to $n^{-1/2}\varepsilon_n^{-d/(2q_\star)}$ up to the same factors. Choosing A large enough therefore gives

$$\begin{aligned}&\sum_{k=2}^{K}\mathbb{E}\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k-1)},(\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k)}\big)\\&\quad\le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{c_\Gamma}B_0^{c_B}p_{\min}^{-c_p}(D\vee\log n)^{\mathfrak{p}_{\beta,d}}\big(\varepsilon_n+n^{-1/2}+n^{-1/2}\varepsilon_n^{-d/(2q_\star)}\big)\\&\qquad+C_{\mathrm{opt}}(D\vee\log n)\sum_{k=2}^{K}\sqrt{t_k\log n}\sqrt{\mathbb{E}\mathcal{E}_{\mathrm{opt},k}}.\end{aligned}\tag{276}$$

Combining (273), (274) and (276) proves (271). Since $D\le n^{a_0}$, choosing c_T in (3) larger than $a_0+(\beta+1)/(d+2\beta)$ by a fixed margin makes $De^{-T}=o(\varepsilon_n)$. Also $\sqrt{Dt_0}=\sqrt{D}\,\varepsilon_n$ is absorbed because $\mathfrak{p}_{\beta,d}\ge1/2$ and $D\le D\vee\log n$.

Step 4: record the aggregate sizes of the interval-specific score networks. The small-noise learned class from Theorem G.6 has separate bounds

$$L_{\mathrm{sm}}\vee\log B_{\mathrm{sm}}\le C\Gamma_{\mathcal{M},q_{\mathrm{geom}}}^{C}B_0^{C}p_{\min}^{-C}(D\vee\log n)^{\mathfrak{p}_{\beta,d}},$$

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, (259) gives

$$L_k \vee \log B_k \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C (D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$
$$\|\boldsymbol{W}_k\|_\infty \vee S_k \le C\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}} B_0^C p_{\min}^{-C} \big(\sigma_{t_{k-1}}^{-d} + \varepsilon_n^{-d/q_\star}\big)(D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

There are $K \lesssim \log n$ large-noise slabs. The deterministic half-open time-selection rule adds no learned weights; the relevant quantities are therefore the maximum depth and weight magnitude and the sums of interval-specific network widths and sparsities. 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

$$K\varepsilon_n^{-d/q_\star} \le C n^{d/(d+2\beta)} (D \vee \log n)^C,$$

because $q_\star > \beta + 1$ by (4). Using the preceding two sums and enlarging the finite exponents c_Γ, c_B, c_p if necessary gives

$$L_{\max} \vee \log B_{\max} \le C\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p} (D \vee \log n)^{\mathfrak{p}_{\beta,d}},$$

and

$$W_{\mathrm{agg}} \vee S_{\mathrm{agg}} \le C\mathfrak{G}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p} n^{d/(d+2\beta)} (D \vee \log n)^{\mathfrak{p}_{\beta,d}}.$$

These are aggregate bounds for separately parameterized score networks, not a claim that their discontinuous half-open time selection is representable by one finite ReLU network. □

Remark H.4 (Disconnected supports and boundaries of the assumptions). Let $\mathcal{M} = \bigsqcup_{j=1}^J \mathcal{M}_j$ be a fixed finite disjoint union of compact, boundaryless, d-dimensional smooth submanifolds with $\delta := \min_{i\neq j} \operatorname{dist}(\mathcal{M}_i, \mathcal{M}_j) > 0$. Then $\mathcal{M}$ is a possibly disconnected manifold with

$$\operatorname{reach}(\mathcal{M}) \ge \min\Big\{\min_j \operatorname{reach}(\mathcal{M}_j), \frac{\delta}{2}\Big\}.$$

If $P_0 = \sum_{j=1}^J \pi_j P_j$, where P_j has density p_j with respect to $\mathrm{vol}_{\mathcal{M}_j}$, $\pi_j > 0$, and $p_j \ge p_{j,\min} > 0$, then the density on the union satisfies $p_0 \ge \min_j \pi_j p_{j,\min} > 0$. With uniformly controlled component atlases and fixed J, this case is therefore already covered by the stated theorem. Intersections, vanishing separation, mixed intrinsic dimensions, manifolds with boundary, singular strata, densities vanishing on their support, and ambient observational noise are not covered. Although the geometric projection module may remain meaningful in some of these cases, the present denominator lower bound and reciprocal argument do not extend to them without a new proof.

Corollary H.5 (Fixed-geometry sufficient ambient-growth regime)**.** *Fix $\beta > 0$ and $d \ge 1$. Work under the OU schedule convention equation 270. Suppose Assumptions 1 to 3 holds and that, for the fixed order q_{geom}, the quantities $\underline{m}$, $\eta_\Pi/c_\star$, $\Gamma_{\mathcal{M},q_{\mathrm{geom}}}$, B_0, and $p_{\min}$ are bounded above or below by positive constants independent of (D, n). Let*

$$a_{\beta,d} := \min\Big\{\frac{2}{d+2\beta}, \frac{2\beta}{(d+2\beta)(\mathfrak{p}_{\beta,d}-1)}\Big\}.$$

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 (236) and (237). Consequently, under this fixed-geometry growth regime, the bound in Corollary H.1 applies.

Proof. Write $H_n = D \vee \log n$. By (3) and (270), $\sigma_{t_1} \asymp n^{-1/(d+2\beta)}$. Hence the small-noise reach condition (236) follows whenever

$$H_n \le c_1 n^{2/(d+2\beta)}$$

for a constant c_1 depending only on $\underline{m}$, $\eta_\Pi/c_\star$, and the schedule constants. Likewise, after absorbing the fixed quantities $p_{\min}^{-1}\Gamma^C_{\mathcal{M},q_{\mathrm{geom}}}B_0^C$ into a constant, the denominator condition (237) 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. □

Corollary H.6 (Sufficient regime with growing ambient, geometry, and density parameters)**.** *Fix $(d, \beta, q_{\mathrm{geom}})$, let $\mathfrak{p} := \mathfrak{p}_{\beta,d}$, and write $c_{\Gamma,\mathrm{den}}, c_{B,\mathrm{den}} < \infty$ for the fixed powers denoted by C in (237). Suppose, along a sequence indexed by n,*

$$D \le Cn^{a_D}, \qquad p_{\min}^{-1} \le Cn^{a_p}, \qquad \Gamma_{\mathcal{M},q_{\mathrm{geom}}} \le Cn^{a_\Gamma},$$
$$B_0 \le Cn^{a_B}, \qquad c_\star\eta_\Pi^{-1} \le Cn^{a_\eta},$$

where all exponents are nonnegative and the OU lower bound $\underline{m}$ is uniform. The reach exponent a_η controls the full ratio $c_\star/\eta_\Pi$, since the localization multiplier $c_\star$ may vary with the geometry and density. Assume

$$a_\eta + \frac{a_D}{2} < \frac{1}{d+2\beta}, \tag{277}$$

$$a_p + c_{\Gamma,\mathrm{den}}a_\Gamma + c_{B,\mathrm{den}}a_B + \frac{\mathfrak{p}-1}{2}a_D < \frac{\beta}{d+2\beta}. \tag{278}$$

Then both small-noise conditions hold for all sufficiently large n. If also $d > 2$, $q_\star \ge d(\beta+1)/(d-2)$, exact ERM is used (or (272) holds), and

$$c_\Gamma a_\Gamma + c_B a_B + c_p a_p + \mathfrak{p}a_D < \frac{\beta+1}{d+2\beta}, \tag{279}$$

then the simplified upper bound in Corollary H.2 converges to zero. The standing polynomial premise holds after choosing any $a_0 > \max\{a_D, a_p\}$ and then choosing c_T in (3) larger than $a_0 + (\beta+1)/(d+2\beta)$; this absorbs the terminal term.

Proof. Since $H_n = D \vee \log n \le n^{a_D+o(1)}$, the left side of the reach condition has exponent $a_\eta + a_D/2 - 1/(d+2\beta) + o(1)$, and the left side of the denominator condition has exponent

$$a_p + c_{\Gamma,\mathrm{den}}a_\Gamma + c_{B,\mathrm{den}}a_B + \frac{\mathfrak{p}-1}{2}a_D - \frac{\beta}{d+2\beta} + o(1).$$

The strict inequalities make both quantities tend to zero. In the $d > 2$ regime, the simplified statistical bound has exponent $c_\Gamma a_\Gamma + c_B a_B + c_p a_p + \mathfrak{p}a_D - (\beta+1)/(d+2\beta) + o(1)$, which is negative by (279). The final two assertions follow directly from the polynomial premise and $De^{-T} \le Cn^{a_D - c_T}$. □

Corollary H.7 (Explicit growing-ambient affine-sphere benchmark)**.** *Set $\beta = 2$, $d = 3$, $q_{\mathrm{geom}} = q_{\mathrm{opt}}(2) = 16$, and, for sufficiently large n, let*

$$D_n := \max\{4, \lfloor(\log n)^2\rfloor\},$$
$$\mathcal{M}_{D_n} := \left\{\left(\tfrac{1}{2}(1,1,1,1) + \tfrac{1}{4}\boldsymbol{u}, \mathbf{0}_{D_n-4}\right) : \boldsymbol{u} \in \mathbb{R}^4,\ \|\boldsymbol{u}\|_2 = 1\right\} \subset [0,1]^{D_n}.$$

Let $P_{0,n}$ be uniform with respect to Riemannian volume on $\mathcal{M}_{D_n}$, and use exact slabwise ERM in the piecewise-in-time estimator. Then the assumptions and both small-noise conditions of Corollary H.1 hold for all sufficiently large n, with geometry and density constants uniform in D_n. Consequently,

$$\mathbb{E}\mathsf{W}_1(P_{0,n}, \widehat{P}_{t_0}^{\mathrm{box}}) \le C(\log n)^{2\mathfrak{p}_{2,3}} n^{-3/7}, \qquad \mathfrak{p}_{2,3} = 3240c_{\mathrm{amb}},$$

where C is independent of (n, D_n).

Proof. The manifold is a radius-$1/4$ round three-sphere contained in the first four coordinates. Hence its reach is $1/4$, its volume is $2\pi^2(1/4)^3 = \pi^2/32$, and its uniform density is $32/\pi^2$. A fixed atlas of the round sphere, padded by zero coordinates, has finite-order geometry constants uniform in D_n; the constant density has a uniform $\mathcal{H}^2$ norm and admits, for example, the uniform lower bound $p_{\min} = 1$. Thus all non-ambient factors in Corollary H.1 are fixed.

Here $d+2\beta = 7$, $t_1 \asymp n^{-2/7}$, and $H_n = D_n \vee \log n \asymp (\log n)^2$. The two left sides in the small-noise checks are, up to fixed constants,

$$n^{-1/7}\log n \longrightarrow 0, \qquad n^{-2/7}(\log n)^{\mathfrak{p}_{2,3}-1} \longrightarrow 0.$$

Moreover $D_n \le n^{a_0}$ for every fixed $a_0 > 0$ eventually. Finally, $q_{\mathrm{geom}} = 16$ and $m_2 = \lceil 2 \rceil - 1 = 1$ give $\mathfrak{p}_{2,3} = 3240 c_{\mathrm{amb}}$ in (227). Since $\varepsilon_n = n^{-(\beta+1)/(d+2\beta)} = n^{-3/7}$ and exact ERM makes every $\mathcal{E}_{\mathrm{opt},k}$ zero, the claimed bound follows from Corollary H.1. □

H.3 Aligned time discretization and direct path-space control

Use the aligned grid from (256) and the grid-weighted score networks $\widehat{s}_{k,M}$ from (265) and (266). The same grid replaces the DSM diffusion-time integral by deterministic quadrature and defines the exponential-integrator time steps; each conditional Gaussian-corruption expectation remains exact. Let $Z_0, \xi_0, \dots, \xi_{M-1}$ be mutually independent $\mathcal{N}(\mathbf{0}, I_D)$ random vectors, jointly independent of the training sample and optimizer randomness, and define

$$Z_{j+1} = e^{\Delta_j} Z_j + 2(e^{\Delta_j} - 1)\widehat{s}_{k(j),M}(Z_j, \tau_j) + \sqrt{e^{2\Delta_j} - 1}\,\xi_j. \tag{280}$$

This exponential integrator freezes only the score and integrates the linear OU drift exactly (Chen et al., 2023a; Benton et al., 2024). Let $\widehat{P}^{\mathrm{disc,box}}_{t_0,M}$ be the conditional law of $\mathrm{Proj}_{[0,1]^D}(Z_M)$. This is an auxiliary appendix result; the main theorem concerns the continuous-time reverse SDE. Unless an outer expectation over the learning randomness is displayed, all sampler, hybrid, path, endpoint laws, and probability kernels in this subsection are conditional on the training sample and optimizer randomness. Throughout this subsection, $n \ge 3$, $0 < t_0 < 1 \le T$, $0 < \eta_{\mathrm{step}} \le 1/2$, $\mathcal{C} = [0,1]^D$, and $H := D \vee \log n$. Assume $P_0(\mathcal{C}) = 1$, the aligned reverse grid satisfies (256), and every grid-evaluated score network is Borel and obeys the closed-slab clipping bound

$$\sup_{\boldsymbol{x} \in \mathbb{R}^D} \|\widehat{s}_{k,M}(\boldsymbol{x}, t)\|_\infty \le C_{\mathrm{clip}} \sigma_t^{-1} H^{1/2}, \qquad t \in \overline{I}_k. \tag{281}$$

The discrete recursion and its continuous interpolation below are therefore well defined and nonexplosive. These conditions hold for the postprocessed grid-weighted score networks constructed in Corollary G.8.

Lemma H.8 (Existence of an aligned reverse grid)**.** *There is a grid containing every time-interval boundary and satisfying* (256) *with*

$$M \le C\eta_{\mathrm{step}}^{-1}\{T + \log(1/t_0)\} + K. \tag{282}$$

Proof. In forward time, partition $[1, T]$ into intervals of length at most η_{step} and partition $[t_0, 1]$ geometrically so that consecutive endpoints satisfy $\tau_j \le (1 + \eta_{\mathrm{step}})\tau_{j+1}$. The first part uses at most $C\eta_{\mathrm{step}}^{-1} T$ intervals and the second at most $C\eta_{\mathrm{step}}^{-1}\log(1/t_0)$. Add the $K - 1$ time-interval boundaries. Adding points only decreases step lengths, so the required inequality remains valid. □

Let $(Y_u)_{0 \le u \le T - t_0}$ be the exact reverse process initialized from P_T, so $Y_u \sim P_{T-u}$. For each slab define the true-score increment

$$\mathcal{D}_{k,M} := \sum_{j \in J_k} \int_{u_j}^{u_{j+1}} \mathbb{E}\Big\| s^*(Y_u, T-u) - s^*(Y_{u_j}, \tau_j)\Big\|_2^2 \mathrm{d}u. \tag{283}$$

Lemma H.9 (True-score discretization along the exact reverse path)**.** *Under Assumption 1,*

$$\sum_{k=1}^{K} \mathcal{D}_{k,M} \le C\{\eta_{\mathrm{step}}^2 DM + \eta_{\mathrm{step}} DT\}. \tag{284}$$

Proof. This is the OU true-score increment estimate of Benton et al. (2024, Lemma 2 and Appendix C), specialized to ambient dimension D, early-stopping time t_0, and the grid condition (256). That general-second-moment form bounds the left side by

$$C\{\eta_{\mathrm{step}}^2 DM + \eta_{\mathrm{step}} DT + \eta_{\mathrm{step}}\mathfrak{m}_2\}, \qquad \mathfrak{m}_2 := \mathbb{E}\|\boldsymbol{X}_0\|_2^2.$$

Here $\boldsymbol{X}_0 \in [0,1]^D$ almost surely, so $\mathfrak{m}_2 \le D$; because $T \ge 1$, the last term is absorbed by $\eta_{\mathrm{step}} DT$. No score-Lipschitz assumption is used. □

For $0 \le k \le K$, define a hybrid reverse process as follows. It is initialized from P_T, uses the exact continuous-time reverse dynamics on $I_{k+1}, \dots, I_K$, and uses the learned exponential-integrator steps on $I_1, \dots, I_k$. Let $\overline{P}^{(k)}_{t_0,M}$ be its endpoint law. Thus $\overline{P}^{(0)}_{t_0,M} = P_{t_0}$ and $\overline{P}^{(K)}_{t_0,M}$ is the learned discrete sampler initialized from P_T.

For $1 \le k \le K$ and an entrance state $\boldsymbol{z}$ at forward time t_k, let $\mu^{\mathrm{true}}_{k,\boldsymbol{z}}$ and $\mu^{\mathrm{disc}}_{k,\boldsymbol{z}}$ be the conditional endpoint laws of the $(k-1)$st and kth hybrids, respectively. These are Borel probability kernels and

$$\overline{P}^{(k-1)}_{t_0,M} = \int \mu^{\mathrm{true}}_{k,\boldsymbol{z}} P_{t_k}(\mathrm{d}\boldsymbol{z}), \qquad \overline{P}^{(k)}_{t_0,M} = \int \mu^{\mathrm{disc}}_{k,\boldsymbol{z}} P_{t_k}(\mathrm{d}\boldsymbol{z}).$$

Lemma H.10 (Averaged endpoint localization)**.** *Suppose $t_k \le 1$. For every fixed $B > 0$, there is a constant $C_B < \infty$, depending only on B, C_{clip}, and the fixed OU constants, such that, with $R_{k,B} := C_B H\sqrt{t_k \log n}$ and $B(\boldsymbol{z}, R)$ denoting the Euclidean ball of radius R centered at $\boldsymbol{z}$,*

$$\begin{aligned}\int \mu^{\mathrm{true}}_{k,\boldsymbol{z}}\big(B(\boldsymbol{z}, R_{k,B})^c\big) P_{t_k}(\mathrm{d}\boldsymbol{z}) &\le C_B n^{-B},\\ \int \mu^{\mathrm{disc}}_{k,\boldsymbol{z}}\big(B(\boldsymbol{z}, R_{k,B})^c\big) P_{t_k}(\mathrm{d}\boldsymbol{z}) &\le C_B n^{-B}.\end{aligned} \tag{285}$$

Proof. First consider a learned discrete continuation from a grid or time-interval boundary $s \le 1$ to t_0, starting at V. Since $\sigma_t^2 \asymp t$ on $(0,1]$, (281) gives $\|\widehat{s}_{j,M}(\cdot,t)\|_2 \le CHt^{-1/2}$ for every score network used by this continuation. Moreover,

$$\sum_{\ell:\, \tau_\ell \le s} (e^{\Delta_\ell} - 1)\tau_\ell^{-1/2} \le C \sum_{\ell:\, \tau_\ell \le s} \Delta_\ell \tau_\ell^{-1/2} \le C\sqrt{s}. \tag{286}$$

Unrolling the recursion therefore gives

$$Z_{\mathrm{end}} - V = (e^{s-t_0} - 1)V + R_s + G_s, \qquad \|R_s\|_2 \le CH\sqrt{s}, \qquad G_s \sim \mathcal{N}\big(0, (e^{2(s-t_0)} - 1)I_D\big). \tag{287}$$

No independence between R_s and G_s is needed. If $V \sim P_s$, then $V = e^{-s}\boldsymbol{X}_0 + \sigma_s \boldsymbol{Z}$ with $\|\boldsymbol{X}_0\|_2 \le \sqrt{D}$. Standard Gaussian norm concentration (Vershynin, 2018), $e^{s-t_0} - 1 \le Cs$, and $e^{2(s-t_0)} - 1 \le Cs$ imply, for every fixed $B > 0$,

$$\Pr\Big(\|Z_{\mathrm{end}} - V\|_2 > C_B H\sqrt{s \log n}\Big) \le C_B n^{-B}. \tag{288}$$

Taking $s = t_k$ proves the second inequality in (285). For the first, let W be the exit state after the exact reverse segment on I_k. Averaged over the entrance $\boldsymbol{z} \sim P_{t_k}$, the pair $(\boldsymbol{z}, W)$ has the forward-OU joint law $(\boldsymbol{X}_{t_k}, \boldsymbol{X}_{t_{k-1}})$. With $h_k := t_k - t_{k-1}$, the forward transition can be written as

$$\boldsymbol{X}_{t_k} = e^{-h_k}\boldsymbol{X}_{t_{k-1}} + \sqrt{1 - e^{-2h_k}}\,\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, I_D),$$

where $\boldsymbol{\xi}$ is independent of $\boldsymbol{X}_{t_{k-1}}$. The same Gaussian norm bound, $h_k \le t_k \le 1$, and $\boldsymbol{X}_{t_{k-1}} \sim P_{t_{k-1}}$ give

$$\Pr\big(\|W - \boldsymbol{z}\|_2 > C_B\sqrt{t_k H}\big) \le C_B n^{-B}.$$

Conditional on W, the remaining learned continuation starts at forward time t_{k-1}; its averaged entrance law is $P_{t_{k-1}}$, so (288) applies with $s = t_{k-1}$. The triangle and union bounds prove the first inequality. When $k = 1$, this remaining continuation has length zero, and the same argument applies. □

Lemma H.11 (Conditional total variation to projected Wasserstein distance)**.** *Let ρ be a probability measure on $\mathbb{R}^D$, and let $(\mu_{\boldsymbol{z}})_{\boldsymbol{z}}$ and $(\nu_{\boldsymbol{z}})_{\boldsymbol{z}}$ be Borel probability kernels on $\mathbb{R}^D$. Write $\mu = \int \mu_{\boldsymbol{z}}\rho(\mathrm{d}\boldsymbol{z})$, $\nu = \int \nu_{\boldsymbol{z}}\rho(\mathrm{d}\boldsymbol{z})$, and $\delta(\boldsymbol{z}) = \mathsf{TV}(\mu_{\boldsymbol{z}}, \nu_{\boldsymbol{z}})$. For any $R > 0$, let $B(\boldsymbol{z}, R)$ be the Euclidean ball of radius R centered at $\boldsymbol{z}$. Then*

$$\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\mu, (\mathrm{Proj}_{\mathcal{C}})_{\#}\nu\big) \le 2R \int \delta(\boldsymbol{z})\rho(\mathrm{d}\boldsymbol{z}) + \sqrt{D}\int\Big\{\mu_{\boldsymbol{z}}(B(\boldsymbol{z},R)^c) + \nu_{\boldsymbol{z}}(B(\boldsymbol{z},R)^c)\Big\}\rho(\mathrm{d}\boldsymbol{z}). \tag{289}$$

Proof. Because $\mathbb{R}^D$ is standard Borel, jointly measurable Radon–Nikodym versions exist for probability kernels. With $\kappa_{\boldsymbol{z}} := \mu_{\boldsymbol{z}} + \nu_{\boldsymbol{z}}$, set

$$\alpha_{\boldsymbol{z}}(\mathrm{d}\boldsymbol{x}) := \min\Big\{\frac{\mathrm{d}\mu_{\boldsymbol{z}}}{\mathrm{d}\kappa_{\boldsymbol{z}}}(\boldsymbol{x}), \frac{\mathrm{d}\nu_{\boldsymbol{z}}}{\mathrm{d}\kappa_{\boldsymbol{z}}}(\boldsymbol{x})\Big\}\kappa_{\boldsymbol{z}}(\mathrm{d}\boldsymbol{x}).$$

The common subkernel has mass $1-\delta(\boldsymbol{z})$. Couple it on the diagonal and, when $\delta(\boldsymbol{z}) > 0$, couple the two normalized residual kernels independently; use any fixed measurable convention for the residual part when $\delta(\boldsymbol{z}) = 0$. The two residual densities are the positive and negative parts, respectively, of $\mathrm{d}\mu_{\boldsymbol{z}}/\mathrm{d}\kappa_{\boldsymbol{z}} - \mathrm{d}\nu_{\boldsymbol{z}}/\mathrm{d}\kappa_{\boldsymbol{z}}$. Hence the residual kernels are mutually singular, and their product coupling assigns zero mass to the diagonal. This gives a measurable maximal-coupling kernel $\pi_{\boldsymbol{z}}$ satisfying $\pi_{\boldsymbol{z}}\{\boldsymbol{x} \ne \boldsymbol{y}\} = \delta(\boldsymbol{z})$. Integrate this kernel against ρ. On coupling failure with both coordinates in $B(\boldsymbol{z}, R)$, nonexpansiveness of $\mathrm{Proj}_{\mathcal{C}}$ bounds their projected distance by $2R$; if either coordinate is outside this ball, the projected diameter is at most $\sqrt{D}$. The union bound for the latter event gives (289). □

Proposition H.12 (Direct slabwise path-space transfer)**.** *Under the standing support, grid, and clipping conditions above, condition on the training sample and optimizer randomness. For every fixed $A > 0$ and $1 \le k \le K$,*

$$\mathsf{W}_1\Big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(k-1)}_{t_0,M}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(k)}_{t_0,M}\Big) \le C_A H\Big\{\sqrt{t_k \log n}\,\big(\mathcal{L}_{k,M}(\widehat{s}_{k,M}) + \mathcal{D}_{k,M}\big)^{1/2} + n^{-A}\Big\}. \tag{290}$$

Consequently,

$$\mathbb{E}\mathsf{W}_1\Big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(k-1)}_{t_0,M}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}^{(k)}_{t_0,M}\Big) \le C_A H\Big\{\sqrt{t_k \log n}\,\big(\mathbb{E}\mathcal{L}_{k,M}(\widehat{s}_{k,M}) + \mathcal{D}_{k,M}\big)^{1/2} + n^{-A}\Big\}. \tag{291}$$

Proof. Because the higher-noise slabs use the exact reverse dynamics in both adjacent hybrids, their common entrance law to slab I_k is P_{t_k}. Conditional on an entrance state $\boldsymbol{z}$, compare the exact path on I_k with the continuous interpolation of the exponential integrator,

$$\mathrm{d}\overline{Z}_u = \{\overline{Z}_u + 2\widehat{s}_{k,M}(\overline{Z}_{u_j}, \tau_j)\}\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_u, \qquad u \in [u_j, u_{j+1}].$$

This interpolation has the recursion in (280) at the grid points.

We justify the path-space comparison by localization. Set $a_k := T - t_k$ and $b_k := T - t_{k-1}$. Let $\mathbb{Q}^{\mathrm{true,k}}$ and $\mathbb{Q}^{\mathrm{disc,k}}$ be the two path laws on $C([a_k, b_k]; \mathbb{R}^D)$ with common initial law P_{t_k}, and let $\mathbb{Q}^{\mathrm{true,k}}_{\boldsymbol{z}}$ and $\mathbb{Q}^{\mathrm{disc,k}}_{\boldsymbol{z}}$ be regular conditional versions given the initial coordinate $\boldsymbol{z}$. On canonical path space, define

$$\zeta_R(\omega) := \inf\{u \in [a_k, b_k] : \|\omega_u\|_2 \ge R\} \wedge b_k.$$

For $u \in [u_j, u_{j+1}]$, write the two progressively measurable drift functionals as

$$b^{\mathrm{true}}(u, \omega) := \omega_u + 2s^*(\omega_u, T-u), \qquad b^{\mathrm{disc}}(u, \omega) := \omega_u + 2\widehat{s}_{k,M}(\omega_{u_j}, \tau_j).$$

Let $\mathbb{Q}^{\mathrm{true,k}}_R$ and $\mathbb{Q}^{\mathrm{disc,k}}_R$ denote the laws obtained by replacing b^{true} and b^{disc} with $\mathbf{1}_{\{u<\zeta_R\}}b^{\mathrm{true}}$ and $\mathbf{1}_{\{u<\zeta_R\}}b^{\mathrm{disc}}$, respectively. On $\{u < \zeta_R\}$, the linear term is bounded by R, while the positive-time OU score is bounded on the compact set $\overline{B}(\mathbf{0}, R) \times [t_{k-1}, t_k]$ by smoothness. Hence the exact drift

is bounded there, and the learned drift is bounded by the linear-term bound and (281). Thus the difference of the stopped drifts is bounded. The exact stopped equation is pathwise unique by local smoothness, and the learned stopped equation is constructed uniquely interval by interval because its score value is frozen within each step. The localized Girsanov argument used in diffusion-model sampling analyses (Chen et al., 2023c, Section 5.2) first gives the following bound with expectation under $\mathbb{Q}_R^{\mathrm{true,k}}$. Under the synchronous coupling of the stopped and full exact equations, their paths agree on $[a_k, \zeta_R)$; consequently, the stopped-law energy equals the displayed expectation under the full exact law of Y:

$$\mathrm{KL}(\mathbb{Q}_R^{\mathrm{true,k}} \| \mathbb{Q}_R^{\mathrm{disc,k}}) \le \sum_{j \in J_k} \int_{u_j}^{u_{j+1}} \mathbb{E}\Big[\mathbf{1}_{\{u<\zeta_R(Y)\}} \big\| s^*(Y_u, T-u) - \widehat{s}_{k,M}(Y_{u_j}, \tau_j)\big\|_2^2\Big] \mathrm{d}u. \quad (292)$$

Here the coefficient is one because the diffusion coefficient is $\sqrt{2} I_D$ and the drift difference is twice the score difference.

Both full processes are nonexplosive: the exact process is the time reversal of the continuous forward OU path on the finite slab, while clipping gives linear growth for the discrete interpolation. Coupling each drift-stopped process to its full counterpart with the same initial state and Brownian motion shows that the two paths agree unless the full path exits the radius-R ball. Continuity on the compact interval $[a_k, b_k]$ makes this exit probability vanish as $R \to \infty$. Hence $\mathbb{Q}_R^{\mathrm{true,k}} \to \mathbb{Q}^{\mathrm{true,k}}$ and $\mathbb{Q}_R^{\mathrm{disc,k}} \to \mathbb{Q}^{\mathrm{disc,k}}$ in total variation. Joint lower semicontinuity of relative entropy, followed by monotone convergence on the right side of (292), gives the same inequality for the full path laws. Since their initial marginal is the common law P_{t_k}, the entropy chain rule gives

$$\begin{aligned}
&\int P_{t_k}(\mathrm{d}\boldsymbol{z})\, \mathrm{KL}(\mathbb{Q}_{\boldsymbol{z}}^{\mathrm{true,k}} \| \mathbb{Q}_{\boldsymbol{z}}^{\mathrm{disc,k}}) \\
&\quad \le C \sum_{j \in J_k} \int_{u_j}^{u_{j+1}} \mathbb{E}\Big\| s^*(Y_u, T-u) - \widehat{s}_{k,M}(Y_{u_j}, \tau_j)\Big\|_2^2 \mathrm{d}u \\
&\quad \le C\{\mathcal{L}_{k,M}(\widehat{s}_{k,M}) + \mathcal{D}_{k,M}\}. \qquad (293)
\end{aligned}$$

Indeed, under the exact reverse law $Y_{u_j} \sim P_{\tau_j}$, so the frozen learned-score term is exactly the grid risk. The second inequality is $\|a+b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$. The true-score increment estimate in Lemma H.9 is the direct discretization ingredient; the stopping argument is the spatial-localization version of the nonsmooth Girsanov method in Chen et al. (2023c, Section 5.2), and the path-measure step follows Chen et al. (2023c); Benton et al. (2024).

After leaving I_k, both hybrids use the same learned discrete Markov kernel. Data processing and conditional Pinsker's inequality (Tsybakov, 2009) therefore give, for the conditional endpoint total-variation distance $\delta_k(\boldsymbol{z})$,

$$\int P_{t_k}(\mathrm{d}\boldsymbol{z})\, \delta_k(\boldsymbol{z}) \le C\{\mathcal{L}_{k,M}(\widehat{s}_{k,M}) + \mathcal{D}_{k,M}\}^{1/2}. \quad (294)$$

It remains to convert this averaged total-variation bound to projected Wasserstein distance without losing the low-noise factor. If $t_k \le 1$, apply Lemma H.11 with $\rho = P_{t_k}$, $(\mu_{\boldsymbol{z}}, \nu_{\boldsymbol{z}}) = (\mu_{k,\boldsymbol{z}}^{\mathrm{true}}, \mu_{k,\boldsymbol{z}}^{\mathrm{disc}})$, and $R = R_{k,B}$. Combining (285) and (294) gives

$$\begin{aligned}
&\mathsf{W}_1\Big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0,M}^{(k-1)}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0,M}^{(k)}\Big) \\
&\quad \le C_B H \sqrt{t_k \log n}\, \{\mathcal{L}_{k,M}(\widehat{s}_{k,M}) + \mathcal{D}_{k,M}\}^{1/2} + C_B \sqrt{D}\, n^{-B}.
\end{aligned}$$

Choose $B > A + 2$ and use $\sqrt{D} \le H$ to obtain (290). If $t_k \ge 1$, total variation contracts under mixing and projection, so (294) and the projected diameter give

$$\mathsf{W}_1\Big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0,M}^{(k-1)}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0,M}^{(k)}\Big) \le C\sqrt{D}\, \{\mathcal{L}_{k,M}(\widehat{s}_{k,M}) + \mathcal{D}_{k,M}\}^{1/2},$$

which is no larger than the first term in (290) because $n \ge 3$ and $t_k \ge 1$. Finally, average over the learned score networks and apply Jensen's inequality to obtain (291). □

Corollary H.13 (Time-discretized distribution bound)**.** *Under the assumptions of Corollary H.1, let* $0 < \eta_{\mathrm{step}} \le 1/2$*, let the aligned grid satisfy* (256)*, and use the corresponding grid-weighted score networks and the sampler in* (280)*. Set* $H := D \vee \log n$*. Then*

$$\begin{aligned}\mathbb{E}\mathsf{W}_1(P_0, \widehat{P}^{\mathrm{disc,box}}_{t_0,M}) \le{}& C\sqrt{Dt_0} + CDe^{-T} \\ &+ C\Gamma^{\mathrm{cr}}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p}\big(\varepsilon_n + n^{-1/2} + n^{-1/2}\varepsilon_n^{-d/(2q_\star)}\big) H^{\mathfrak{p}_{\beta,d}} \\ &+ C_{\mathrm{opt}} H \sum_{k=1}^{K} \sqrt{t_k \log n}\, \sqrt{\mathbb{E}\mathcal{E}^{\mathrm{grid}}_{\mathrm{opt},k}} \\ &+ CH(\log n)^{3/2}\{\eta_{\mathrm{step}}^2 DM + \eta_{\mathrm{step}} DT\}^{1/2}. \end{aligned} \tag{295}$$

Proof. Telescope the hybrids from $\overline{P}^{(0)}_{t_0,M} = P_{t_0}$ to $\overline{P}^{(K)}_{t_0,M}$, use Proposition H.12, and compare initialization from P_T with initialization from $\mathcal{N}(0, I_D)$ through the common learned discrete kernel. The early-stopping and terminal terms are bounded by Lemmas I.16 and I.17; the proof of the latter uses only contraction through a common Markov kernel and therefore applies verbatim to the discrete learned kernel. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and substituting (268) and (269) into the score part of (291) gives exactly the dyadic sums in Steps 2–3 of the proof of Corollary H.1; this yields the first three lines of (295), after choosing A large enough to absorb the sum of the $K \lesssim \log n$ negligible remainders.

For the true-score increments, Cauchy–Schwarz, $\sum_{k=1}^{K} t_k \le C(\log n)^2$, and (284) give

$$\begin{aligned}\sum_{k=1}^{K} \sqrt{t_k \log n}\, \mathcal{D}_{k,M}^{1/2} &\le \Big(\log n \sum_{k=1}^{K} t_k\Big)^{1/2} \Big(\sum_{k=1}^{K} \mathcal{D}_{k,M}\Big)^{1/2} \\ &\le C(\log n)^{3/2}\{\eta_{\mathrm{step}}^2 DM + \eta_{\mathrm{step}} DT\}^{1/2},\end{aligned}$$

which proves the last line of (295). □

Corollary H.14 (Simplified time-discretization rate)**.** *Under the assumptions and parameter choices of Corollary H.2, set* $H := D \vee \log n$ *and suppose*

$$\max_{1\le k\le K} \sqrt{\mathbb{E}\mathcal{E}^{\mathrm{grid}}_{\mathrm{opt},k}} \lesssim_{\log} \varepsilon_n.$$

For all sufficiently large n*, choose a fixed sufficiently small numerical constant* $c > 0$ *and set*

$$\eta_{\mathrm{step}} = c\frac{\varepsilon_n^2}{DH^2(\log n)^4} \tag{296}$$

and construct an aligned grid with

$$M \le CDH^2\varepsilon_n^{-2}(\log n)^5. \tag{297}$$

The sampler in (280) *then satisfies*

$$\mathbb{E}\,\mathsf{W}_1\big(P_0, \widehat{P}^{\mathrm{disc,box}}_{t_0,M}\big) \lesssim_{\log} \Gamma^{\mathrm{cr}}_{\mathcal{M},q_{\mathrm{geom}}} B_0^{c_B} p_{\min}^{-c_p} D^{\mathfrak{p}_{\beta,d}} \varepsilon_n.$$

This controls the number of score-network evaluations, not the time needed to compute the slabwise approximate empirical minimizers.

Proof. By $T + \log(1/t_0) \le C\log n$ and Lemma H.8, $M \le C\eta_{\mathrm{step}}^{-1}\log n$ and $\eta_{\mathrm{step}}^2 DM + \eta_{\mathrm{step}} DT \le C\eta_{\mathrm{step}} D\log n$. The choice in (296) gives (297) and makes the last line of (295) at most $C\varepsilon_n$. The remaining terms simplify exactly as in Corollary H.2. □

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 tools are grouped into neural-network approximation, generalization bounds, and distribution-estimation lemmas.

I.1 NEURAL NETWORK APPROXIMATION TOOLKIT

I.1.1 IMPORTED NEURAL NETWORK APPROXIMATION LEMMAS

Arithmetic constructions below use tolerances in $(0, 1/2]$. Whenever a subsequent construction requests a larger additive tolerance δ, use $\delta \wedge (1/2)$ instead; its logarithmic cost is $\log(2 \vee \delta^{-1})$ and is absorbed by the displayed logarithmic scales, which are at least one. For a reciprocal network, use the capped parameter also in the input-sensitivity factor.

Theorem I.1 (ReLU approximation of Hölder functions, after (Schmidt-Hieber, 2020, Theorem 5))**.** *For any function $f \in \mathcal{H}^\beta([0,1]^D, B)$, put $B_\Sigma := 2(D+1)^{\lceil\beta\rceil+2}B$ to convert the maximum-based Hölder radius to the sum-based radius in the cited theorem. For any integers $m \geq 1$ and $N \geq (\beta+1)^D \vee (B_\Sigma+1)e^D$, there exists a network*

$$\widetilde{f} \in \mathcal{NN}(L, \boldsymbol{W}, S, 1)$$

with

$$\begin{aligned} L &= 9 + (m+5)(1 + \lceil \log_2(D \vee \beta) \rceil), \\ \|\boldsymbol{W}\|_\infty &\leq 6(D + \lceil\beta\rceil)N, \\ S &\leq 141(D+\beta+1)^{D+3} N(m+6) \end{aligned}$$

such that

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^D)} \leq (2B_\Sigma + 1)(1 + D^2 + \beta^2)6^D N 2^{-m} + B_\Sigma 3^\beta N^{-\frac{\beta}{D}}.$$

The additional layer in L converts the source's hidden-layer convention to the affine-layer convention in Definition B.1. The radius conversion follows by summing the finitely many derivative seminorms and using $\|u\|_2 \leq \sqrt{D}\|u\|_\infty$. In the applications below the approximation domain is d-dimensional, so this conversion changes only the displayed intrinsic and geometry constants.

Lemma I.2 (ReLU approximation of the OU schedules, after Lemma B.1 of (Oko et al., 2023))**.** *Let $m_t = e^{-t}$ and $\sigma_t = (1 - e^{-2t})^{1/2}$ 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 $\bar{\varepsilon} := \varepsilon \wedge (\underline{t}/2)$, $L \leq \mathcal{O}(\log^2 \bar{\varepsilon}^{-1})$, $\|\boldsymbol{W}\|_\infty = \mathcal{O}(\log^3 \bar{\varepsilon}^{-1})$, $S = \mathcal{O}(\log^4 \bar{\varepsilon}^{-1})$, and $B = \exp(\mathcal{O}(\log^2 \bar{\varepsilon}^{-1}))$. Applying the cited lemma at $\bar{\varepsilon}$ enforces its lower-time condition.

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 $0 < \varepsilon \leq 1/2$, there exists a ReLU network $\phi_{\mathrm{multi}}(x_1, \ldots, x_d) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$\begin{aligned} L &\lesssim \log d\big(\log(d/\varepsilon) + d \log C\big), \\ \|\boldsymbol{W}\|_\infty &= 48d, \\ S &\lesssim d\big(\log(d/\varepsilon) + d \log C\big), \\ B &= C^d, \end{aligned}$$

such that

$$\Big|\phi_{\mathrm{multi}}(x'_1, \ldots, x'_d) - \prod_{k=1}^d x_k\Big| \leq \varepsilon + dC^{d-1}\varepsilon',$$

$$\text{for all } \boldsymbol{x} \in [-C, C]^d,\ \boldsymbol{x}' \in \mathbb{R}^d \text{ with } \|\boldsymbol{x} - \boldsymbol{x}'\|_\infty \leq \varepsilon'.$$

Moreover, $|\phi_{\mathrm{multi}}(\boldsymbol{x})| \le C^d$ *for all* $\boldsymbol{x} \in \mathbb{R}^d$*, and* $\phi_{\mathrm{multi}}(x'_1, \dots, x'_d) = 0$ *if at least one of* $x'_k = 0$*. We note that some of* x_i, x_j $(i \ne j)$ *can be shared. For the product* $\prod_{k=1}^{I} x_k^{\alpha_k}$ *with* $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_I) \in \mathbb{N}^I$ *and* $\sum_{k=1}^{I} \alpha_k = d$*, there exists a neural network satisfying the same bounds as above, denoted by* $\phi_{\mathrm{multi}}(\boldsymbol{x}; \boldsymbol{\alpha})$*.*

Proof. Use the zero-preserving two-input multiplication construction in the proof of Oko et al. (2023, Lemma F.6), with its output clipped to $[-1, 1]$. For accuracy $0 < \delta \le 1/2$, this block has constant width, depth and sparsity $\mathcal{O}(\log \delta^{-1})$, and weights bounded by one. Clipping is exact ReLU arithmetic, preserves zero, and cannot increase the error relative to a product in $[-1, 1]$. Arrange these blocks in a balanced binary tree with $p = 2^{\lceil \log_2 d \rceil} < 2d$ leaves, padding unused leaves by one. All intermediate outputs lie in $[-1, 1]$. The inequality $|ab - a'b'| \le |a - a'| + |b - b'|$ on this square shows inductively that the tree error is at most $(p - 1)\delta$. Choose $\delta = \varepsilon/(2dC^d)$, clip each input to $[-C, C]$, divide it by C, and multiply the final output by C^d. The exact-input error is at most ε. There are $\mathcal{O}(d)$ blocks and $\mathcal{O}(\log d)$ levels; the construction in the cited proof gives the displayed width, and counting its blocks gives the stated depth and sparsity with $\log \delta^{-1} \lesssim \log(d/\varepsilon) + d \log C$. The input clipping and final rescaling have weights at most C^d. For perturbed inputs, let $\bar{x}'_k = \mathrm{clip}_{[-C,C]}(x'_k)$; then $|\bar{x}'_k - x_k| \le |x'_k - x_k|$, and telescoping the product gives an additional error at most $dC^{d-1}\varepsilon'$. The output bound and exact zero preservation follow from the clipped tree. For repeated coordinates, an initial signed-copy layer copies the d inputs using $\mathcal{O}(d)$ unit-weight connections, proving the monomial assertion without increasing the weight bound. □

The bounds retain the degree and range factors arising in the construction of Oko et al. (2023, Lemma F.6).

Lemma I.4 (ReLU reciprocal network, after Lemma F.7 of (Oko et al., 2023))**.** *For any* $\varepsilon \in (0, 1/2]$*, 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

$$\left|\phi_{\mathrm{rec}}(x') - \frac{1}{x}\right| \le \varepsilon + \frac{|x' - x|}{\varepsilon^2}, \quad \text{for all } x \in [\varepsilon, \varepsilon^{-1}] \text{ and } x' \in \mathbb{R}.$$

In particular, on any interval $[a, b] \subset (0, \infty)$, choosing the lemma's parameter no larger than $\eta \wedge a \wedge b^{-1} \wedge (1/2)$ gives additive error η at exact inputs. For perturbed inputs, the preceding squared-denominator factor must also be included in the internal error allocation.

Lemma I.5 (Pointwise ReLU approximation of e^{-x}, after Lemma F.12 in the arXiv v1 of (Oko et al., 2023))**.** *For any* $0 < \varepsilon \le 1/2$*, there exists ReLU network* $\phi_{\exp} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ *such that*

$$\left|e^{-x'} - \phi_{\exp}(x)\right| \le \varepsilon + |x - x'| \qquad \text{for all } x, x' \ge 0$$

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 Lemma I.2 and the algebraic network lemmas.

Lemma I.6 (ReLU approximation of the inverse OU 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

$$\varepsilon_m := \frac{\underline{m}^2 \varepsilon^3}{8}.$$

By Lemma 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 (Lemma 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

$$\left|\phi_{\mathrm{rec}}(\widetilde{\phi}_m(t)) - \frac{1}{m_t}\right| \le \varepsilon_{\mathrm{rec}} + \frac{|\widetilde{\phi}_m(t) - m_t|}{\varepsilon_{\mathrm{rec}}^2} \le \frac{\underline{m}\varepsilon}{2} + \frac{\varepsilon_m}{(\underline{m}\varepsilon/2)^2} \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

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 $M_x \ge 1$ and assume $\|\boldsymbol{x}\|_\infty \le M_x$. For every $\varepsilon \in (0, 1/2]$ there exists a vector-valued ReLU network*

$$\widetilde{\phi}(\boldsymbol{x}, t) = \big(\widetilde{\phi}_1(\boldsymbol{x}, t), \dots, \widetilde{\phi}_D(\boldsymbol{x}, t)\big) \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$$

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

$$\left\|\widetilde{\phi}(\boldsymbol{x}, t) - \frac{\boldsymbol{x}}{m_t}\right\|_{L^\infty(\{\|\boldsymbol{x}\|_\infty \le M_x\} \times [\underline{t},\bar{t}])} \le \varepsilon.$$

Proof. Let $\widetilde{\phi}_{1/m}$ be the network from Lemma 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

$$\|\widetilde{\phi}_{1/m} - m_t^{-1}\|_{L^\infty([\underline{t},\bar{t}])} \le \varepsilon_0.$$

For each coordinate $j = 1, \dots, D$, apply the multiplication lemma (Lemma 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

$$\left|\phi_{\mathrm{mult},j}\big(x_j, \widetilde{\phi}_{1/m}(t)\big) - x_j m_t^{-1}\right| \le \frac{\varepsilon}{2} + 2(1 \vee M_x \vee \underline{m}^{-1})\,\varepsilon_0 \le \varepsilon.$$

Define

$$\widetilde{\phi}_j(\boldsymbol{x}, t) := \phi_{\mathrm{mult},j}\big(x_j, \widetilde{\phi}_{1/m}(t)\big), \qquad j = 1, \dots, D.$$

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 $\varepsilon \le \underline{t} < \bar{t} < \infty$, there exists a ReLU network $\phi_{\sigma^k} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with*

$$L \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + k,$$
$$S \lesssim \log^4 \varepsilon^{-1} + \log^4 k + k \log(k/\varepsilon), \quad \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 k,$$

such that

$$\left\|\phi_{\sigma^k}(t) - \sigma_t^k\right\|_{L^\infty([\underline{t},\bar{t}])} \le 3\varepsilon.$$

Proof. By Lemma 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

$$\begin{aligned} L &\lesssim \log^2 \varepsilon^{-1} + \log^2 k, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + \log^3 k, \\ S &\lesssim \log^4 \varepsilon^{-1} + \log^4 k, \quad \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \end{aligned} \tag{298}$$

such that

$$|\phi_\sigma(t) - \sigma_t| \le \varepsilon/k, \quad \text{for all } t \in [\underline{t}, \bar{t}].$$

Notice that $\sigma_t \in [0, 1]$ for all $t \in [\underline{t}, \bar{t}]$. If $k = 1$, take $\phi_{\sigma^k} := \phi_\sigma$, and the claim follows. Assume now that $k \ge 2$. Apply Lemma 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 Lemma I.3. Thus there exists a ReLU network $\phi_{\text{multi}} \in \mathcal{NN}(L_1, \boldsymbol{W}_1, S_1, B_1)$ with

$$L_1 \lesssim \log k \, \log(k/\varepsilon), \quad \|\boldsymbol{W}_1\|_\infty = 48k, \quad S_1 \lesssim k \log(k/\varepsilon), \quad B_1 = 1 \tag{299}$$

such that

$$\left|\phi_{\text{multi}}\big(\underbrace{\phi_\sigma(t), \ldots, \phi_\sigma(t)}_{k \text{ copies}}\big) - \sigma_t^k\right| \le \varepsilon + k\frac{\varepsilon}{k} \le 2\varepsilon. \tag{300}$$

For all $t \in [\underline{t}, \bar{t}]$, we define

$$\phi_{\sigma^k}(t) \coloneqq \phi_{\text{multi}}\big(\underbrace{\phi_\sigma(t), \ldots, \phi_\sigma(t)}_{k \text{ copies}}\big).$$

Then by the configurations given in (298) and (299) and the bound in (300), we obtain that there exists a ReLU network $\phi_{\sigma^k} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with

$$L \lesssim \log^2 \varepsilon^{-1} + \log^2 k, \quad \|\boldsymbol{W}\|_\infty \lesssim \log^3 \varepsilon^{-1} + k,$$
$$S \lesssim \log^4 \varepsilon^{-1} + \log^4 k + k \log(k/\varepsilon), \quad \log B \lesssim \log^2 \varepsilon^{-1} + \log^2 k,$$

such that

$$\left\|\phi_{\sigma^k}(t) - \sigma_t^k\right\|_{L^\infty([\underline{t},\bar{t}])} \le 3\varepsilon.$$

□

Lemma I.9 (ReLU approximation of inverse powers of σ_t)**.** *For $0 < \underline{t} < \bar{t} < \infty$, $k \in \mathbb{N}_+$ and $0 < \varepsilon \le 1$, put*

$$\Lambda_k := \log(e/\varepsilon) + k \log(e/\sigma_{\underline{t}}) + \log(1 + k) + \log(e + \underline{t}^{-1}).$$

There is a ReLU network $\phi_{\sigma^{-k}} \in \mathcal{NN}(L, \boldsymbol{W}, S, B)$ with

$$\|\phi_{\sigma^{-k}} - \sigma^{-k}\|_{L^\infty([\underline{t},\bar{t}])} \le \varepsilon, \quad L \lesssim \Lambda_k^2, \quad \|\boldsymbol{W}\|_\infty \lesssim \Lambda_k^3 + k, \quad S \lesssim \Lambda_k^4 + k\Lambda_k, \quad \log B \lesssim \Lambda_k^2.$$

Proof. Set $\varepsilon_{\text{rec}} = \sigma_{\underline{t}}^k \varepsilon/4$ and

$$\varepsilon_{\sigma,k} := \frac{\varepsilon \varepsilon_{\text{rec}}^2}{12} \wedge \frac{\underline{t}}{2} \wedge \frac{1}{4}.$$

By Lemma I.8, choose f_k such that $\|f_k - \sigma^k\|_\infty \le 3\varepsilon_{\sigma,k}$. Since $\sigma_t^k \in [\varepsilon_{\text{rec}}, \varepsilon_{\text{rec}}^{-1}]$, the corrected reciprocal estimate gives

$$\left|\phi_{\text{rec}}(f_k(t)) - \sigma_t^{-k}\right| \le \varepsilon_{\text{rec}} + \frac{3\varepsilon_{\sigma,k}}{\varepsilon_{\text{rec}}^2} \le \varepsilon/2.$$

Both logarithmic internal accuracies are bounded by $C\Lambda_k$. Since $\log k \le \Lambda_k$, the positive-power multiplication cost $k \log(k/\varepsilon_{\sigma,k})$ is at most $Ck\Lambda_k$. Adding the positive-power and reciprocal sizes proves the stated bounds, including the linear $k\Lambda_k$ product cost. □

I.2 Support lemmas for generalization analysis

Lemma I.10 (Unconditional Fisher bound for the smoothed data law)**.** *For the OU process (1), 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 (Efron, 2011) gives

$$\nabla \log p_t(\boldsymbol{y}) = \mathbb{E}\Big[-\frac{\boldsymbol{Z}}{\sigma_t} \;\Big|\; \boldsymbol{X}_t = \boldsymbol{y}\Big].$$

Therefore Jensen's inequality and the tower property imply

$$\mathbb{E}_{\boldsymbol{X}_t}\|\nabla \log p_t(\boldsymbol{X}_t)\|_2^2 \le \mathbb{E}\Big\|\frac{\boldsymbol{Z}}{\sigma_t}\Big\|_2^2 = D\sigma_t^{-2}.$$

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 (Score clipping in mean square)**.** *Let $\boldsymbol{X}_t = m_t \boldsymbol{X}_0 + \sigma_t \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_D)$ is independent of $\boldsymbol{X}_0$, and let $t > 0$. For $u \ge 0$ and $R > 0$ with $R^2 \ge 2D \log 2 + 4u$, let $\mathcal{C}_{R,t}$ denote coordinatewise clipping to $[-R/\sigma_t, R/\sigma_t]^D$. For every Borel score f and Borel set $A \subset \mathbb{R}^D$,*

$$\mathbb{E}\big[\mathbf{1}_A(\boldsymbol{X}_t)\|\mathcal{C}_{R,t} f(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2\big] \le 2\mathbb{E}\big[\mathbf{1}_A(\boldsymbol{X}_t)\|f(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2\big] + 8\sigma_t^{-2} e^{-u}. \tag{301}$$

The inequality also holds after integration against any finite positive time measure, with the last term replaced by $8e^{-u}\int \sigma_t^{-2}\mathrm{d}\nu(t)$, provided the functions and sets are jointly measurable.

Proof. The Gaussian score identity gives $\boldsymbol{V}_t := \sigma_t s^*(\boldsymbol{X}_t, t) = -\mathbb{E}[\boldsymbol{Z} \mid \boldsymbol{X}_t]$. Conditional Jensen's inequality applied to the convex function $\boldsymbol{v} \mapsto \exp(\|\boldsymbol{v}\|_2^2/4)$ gives

$$\mathbb{E}e^{\|\boldsymbol{V}_t\|_2^2/4} \le \mathbb{E}e^{\|\boldsymbol{Z}\|_2^2/4} = 2^{D/2}.$$

Hence, by the tail integral formula,

$$\mathbb{E}(\|\boldsymbol{V}_t\|_2 - R)_+^2 \le \int_{R^2}^{\infty} \Pr(\|\boldsymbol{V}_t\|_2^2 > v)\mathrm{d}v \le 4\,2^{D/2} e^{-R^2/4} \le 4e^{-u}.$$

The cube $[-R, R]^D$ contains the Euclidean ball of radius R, so its projection residual at $\boldsymbol{V}_t$ is at most $(\|\boldsymbol{V}_t\|_2 - R)_+$. Nonexpansiveness of coordinatewise clipping, the triangle inequality, and $(a+b)^2 \le 2a^2 + 2b^2$ now prove (301). Tonelli's theorem proves the time-integrated assertion, including atomic time measures. □

Lemma I.12 (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. For every $\boldsymbol{x} \in \mathbb{R}^D$ such that $\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}[\|s(\boldsymbol{X}_t, t)\|_2^2 + \|s'(\boldsymbol{X}_t, t)\|_2^2]\mathrm{d}t < \infty$, the two losses are finite and it holds that*

$$\begin{aligned}
&\big(\ell_{\underline{t},\bar{t}}(s, \boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s', \boldsymbol{x})\big)^2 \\
&\le 3\Big(\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t, t)\|_2^2 + \|s'(\boldsymbol{X}_t, t)\|_2^2\big]\mathrm{d}t + 4D\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t\Big) \\
&\quad \times \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\big\|s(\boldsymbol{X}_t, t) - s'(\boldsymbol{X}_t, t)\big\|_2^2\Big]\mathrm{d}t.
\end{aligned}$$

Proof. For brevity, set $\Delta_t := s(\boldsymbol{X}_t, t) - s'(\boldsymbol{X}_t, t)$ and $\Sigma_t := s(\boldsymbol{X}_t, t) + s'(\boldsymbol{X}_t, t) + 2(\boldsymbol{X}_t - m_t\boldsymbol{x})/\sigma_t^2$.

$$\begin{aligned}
&\ell_{\underline{t},\bar{t}}(s, \boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s', \boldsymbol{x}) \\
&= \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t, t) + \frac{\boldsymbol{X}_t - m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2 - \Big\|s'(\boldsymbol{X}_t, t) + \frac{\boldsymbol{X}_t - m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t \\
&= \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\Delta_t^\top \Sigma_t\big]\mathrm{d}t.
\end{aligned}$$

Applying the Cauchy–Schwarz inequality twice, we obtain that

$$\begin{aligned}
&|\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})| \\
&\le \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|\Delta_t\|_2\|\Sigma_t\|_2\big]\mathrm{d}t \qquad \text{(by Cauchy-Schwarz inequality)} \\
&\le \sqrt{\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|\Delta_t\|_2^2\big]\mathrm{d}t} \cdot \sqrt{\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|\Sigma_t\|_2^2\big]\mathrm{d}t}. \qquad \text{(by Cauchy–Schwarz)}
\end{aligned}$$

For the OU 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

$$\begin{aligned}
&\sqrt{\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t) + s'(\boldsymbol{X}_t,t) + \frac{2(\boldsymbol{X}_t - m_t\boldsymbol{x})}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t} \\
&\le \sqrt{3\int_{\underline{t}}^{\bar{t}} \Big(\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2 + \|s'(\boldsymbol{X}_t,t)\|_2^2\big] + 4\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|\frac{\boldsymbol{X}_t - m_t\boldsymbol{x}}{\sigma_t^2}\Big\|_2^2\Big]\Big)\mathrm{d}t} \\
&\le \sqrt{3\int_{\underline{t}}^{\bar{t}} \Big(\mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2 + \|s'(\boldsymbol{X}_t,t)\|_2^2\big] + 4D\sigma_t^{-2}\Big)\mathrm{d}t}. \qquad (302)
\end{aligned}$$

Hence we obtain

$$\begin{aligned}
&\sqrt{\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\Big[\Big\|s(\boldsymbol{X}_t,t) + s'(\boldsymbol{X}_t,t) + \frac{2(\boldsymbol{X}_t - m_t\boldsymbol{x})}{\sigma_t^2}\Big\|_2^2\Big]\mathrm{d}t} \\
&\le \sqrt{3\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2 + \|s'(\boldsymbol{X}_t,t)\|_2^2\big]\mathrm{d}t + 12D\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t}. \qquad (303)
\end{aligned}$$

Thus, again by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\big(\ell_{\underline{t},\bar{t}}(s,\boldsymbol{x}) - \ell_{\underline{t},\bar{t}}(s',\boldsymbol{x})\big)^2 \\
&\le \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|\Delta_t\|_2^2\big]\mathrm{d}t \cdot \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|\Sigma_t\|_2^2\big]\mathrm{d}t \\
&\le 3\Big(\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|s(\boldsymbol{X}_t,t)\|_2^2 + \|s'(\boldsymbol{X}_t,t)\|_2^2\big]\mathrm{d}t + 4D\int_{\underline{t}}^{\bar{t}} \sigma_t^{-2}\mathrm{d}t\Big) \\
&\qquad \times \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{\boldsymbol{X}_t|\boldsymbol{X}_0=\boldsymbol{x}}\big[\|\Delta_t\|_2^2\big]\mathrm{d}t.
\end{aligned}$$

□

Lemma I.13 (Uniform boundedness of clipped DSM 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 OU 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

$$\|s(\boldsymbol{y}, t)\|_2^2 \le DC_{\text{clip}}^2 \sigma_t^{-2}(D \vee \log n) \le CC_{\text{clip}}^2 (D \vee \log n)^2 \sigma_t^{-2}.$$

Also

$$\mathbb{E}_{\boldsymbol{X}_t \mid \boldsymbol{X}_0 = \boldsymbol{x}} \Big\| \frac{\boldsymbol{X}_t - m_t\boldsymbol{x}}{\sigma_t^2} \Big\|_2^2 = D\sigma_t^{-2} \le (D \vee \log n)^2 \sigma_t^{-2}.$$

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.14 (Bernstein's inequality for bounded random variables, after 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*

$$\Pr\Big[\Big|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\Big| \ge t\Big] \le 2\exp\Big(-\frac{t^2/2}{\sum_{i=1}^n \mathbb{E}[X_i^2] + 2Kt/3}\Big).$$

In other words, with probability at least $1-\delta$, it holds that

$$\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\Big| \le \frac{4K\log(2/\delta)}{3n} + \sqrt{\frac{\frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i^2]\log(2/\delta)}{n}}.$$

To obtain these forms, apply Bernstein's inequality to $X_i - \mathbb{E}X_i$, whose absolute value is at most $2K$ and whose variance is at most $\mathbb{E}X_i^2$, and solve the resulting quadratic tail bound.

I.2.2 Covering number bounds

Lemma I.15 (Covering number of sparse ReLU networks, after Suzuki (2019), Lemma 3)**.** *For any input dimension D_{in} and any $0 < \tau \le 1$, the covering number of $\mathcal{NN}(L, \boldsymbol{W}, S, B)$ on $[0,1]^{D_{\text{in}}}$ can be bounded by*

$$\log \mathcal{N}(\tau, \mathcal{NN}(L, \boldsymbol{W}, S, B), \|\cdot\|_{L^\infty([0,1]^{D_{\text{in}}})}) \lesssim SL\log\big(2+\tau^{-1}L(\|\boldsymbol{W}\|_\infty + 1)(B \vee 1)\big).$$

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.

Proof. Put $W = \max_{0 \le \ell \le L} W_\ell$, including the input and output widths, and suppose $L, S \ge 1$. There are at most $P = L(W+1)^2$ parameter locations and at most $(P+1)^S$ support patterns of size at most S. On a fixed support, perturb each parameter by at most η. Induction through the affine layers and the 1-Lipschitz ReLU gives output perturbation at most $L[(W+1)(B \vee 1)]^L \eta$ in the coordinatewise supremum norm. Take $\eta = \tau / \{2L[(W+1)(B \vee 1)]^L\}$ and cover each active parameter interval by at most $1 + 2B/\eta$ grid points. Taking logarithms of the product of the support and parameter-grid counts yields the displayed bound. This argument applies directly to vector outputs and nonuniform hidden widths. The class with $S = 0$ is a singleton and has zero entropy. □

I.3 Auxiliary lemmas

The following lemmas are stated and proved for the cube-projected OU endpoint used by the main theorem. Projection onto $[0,1]^D$ is essential: it ensures a bounded diameter after total-variation comparison, while leaving the target law fixed.

Lemma I.16 (Cube-projected early stopping for the OU forward law)**.** *Let $\mathcal{C} = [0,1]^D$, let $P_t^{\text{box}} := (\text{Proj}_{\mathcal{C}})_\# P_t$, and assume $P_0(\mathcal{C}) = 1$. Then*

$$\mathsf{W}_1(P_0, P_{t_0}^{\text{box}}) \le \sqrt{D}\,\{1 - e^{-t_0} + \sqrt{1 - e^{-2t_0}}\} \le C\sqrt{Dt_0}, \qquad 0 < t_0 \le 1. \tag{304}$$

Proof. Couple $\boldsymbol{X}_0 \sim P_0$ with $\boldsymbol{X}_{t_0} = e^{-t_0}\boldsymbol{X}_0 + \sqrt{1-e^{-2t_0}}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(0, I_D)$ is independent. Since $\mathrm{Proj}_{\mathcal{C}}$ fixes $\boldsymbol{X}_0$ and is nonexpansive,

$$\mathsf{W}_1(P_0, P_{t_0}^{\mathrm{box}}) \le \mathbb{E}\|\mathrm{Proj}_{\mathcal{C}}(\boldsymbol{X}_{t_0}) - \mathrm{Proj}_{\mathcal{C}}(\boldsymbol{X}_0)\|_2 \le \mathbb{E}\|\boldsymbol{X}_{t_0} - \boldsymbol{X}_0\|_2.$$

Now $\|\boldsymbol{X}_0\|_2 \le \sqrt{D}$ and $\mathbb{E}\|\boldsymbol{Z}\|_2 \le \sqrt{D}$, which gives the first bound. The inequalities $1 - e^{-t_0} \le t_0$ and $1 - e^{-2t_0} \le 2t_0$ give the second. □

Lemma I.17 (Cube-projected terminal Gaussian initialization)**.** *Condition on the training sample and any optimizer randomness, and fix the resulting reverse OU Markov kernel from time T to t_0. Let $\overline{P}_{t_0}$ and $\widehat{P}_{t_0}$ be its output laws when initialized from P_T and $\widehat{P}_T = \mathcal{N}(0, I_D)$, respectively. Then, almost surely,*

$$\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\widehat{P}_{t_0}\big) \le \sqrt{D}\,\mathsf{TV}(P_T, \widehat{P}_T) \le CDe^{-T}. \tag{305}$$

The same inequality holds after averaging over the learned score.

Proof. Total variation contracts through a common Markov kernel and through measurable post-processing. Both cube-projected output laws are supported on $\mathcal{C}$, whose Euclidean diameter is $\sqrt{D}$; hence $\mathsf{W}_1 \le \sqrt{D}\,\mathsf{TV}$ for these laws.

For the second inequality, condition on $\boldsymbol{X}_0 = \boldsymbol{y}$. The OU terminal conditional law is $\mathcal{N}(e^{-T}\boldsymbol{y}, (1 - e^{-2T})I_D)$. If $e^{-T} \le 1/2$, the Gaussian KL formula and $\|\boldsymbol{y}\|_2^2 \le D$ give

$$\mathrm{KL}\big(\mathcal{N}(e^{-T}\boldsymbol{y}, (1-e^{-2T})I_D) \,\|\, \mathcal{N}(0, I_D)\big) \le CDe^{-2T}.$$

Convexity of KL over the mixture P_T, followed by Pinsker's inequality (Tsybakov, 2009), yields $\mathsf{TV}(P_T, \widehat{P}_T) \le C\sqrt{D}e^{-T}$. If $e^{-T} > 1/2$, the same inequality follows after increasing C from $\mathsf{TV} \le 1$. Multiplication by the diameter $\sqrt{D}$ of the projected state space proves the claim. □

Lemma I.18 (Cube-projected slabwise OU perturbation)**.** *Let $n \ge 3$, $H := D \vee \log n$, and use the disjoint partition*

$$I_1 = [t_0, t_1), \qquad I_k = [t_{k-1}, t_k)\ (2 \le k < K), \qquad I_K = [t_{K-1}, T].$$

Condition on the training sample and optimizer randomness. Let $\widehat{s}_j$ be the learned score network on I_j, and assume the coordinatewise clipping bound

$$\sup_{\boldsymbol{x}\in\mathbb{R}^D} \|\widehat{s}_j(\boldsymbol{x}, t)\|_\infty \le C_{\mathrm{clip}}\sigma_t^{-1}H^{1/2}, \qquad t \in I_j. \tag{306}$$

For $0 \le k \le K$, let $\overline{P}_{t_0}^{(k)}$ be the endpoint law of the reverse OU process initialized from P_T, using the learned score networks on $I_1, \dots, I_k$ and the true score on $I_{k+1}, \dots, I_K$. Set

$$\mathcal{L}_{I_k}(\widehat{s}_k) := \int_{I_k} \mathbb{E}_{\boldsymbol{X}_t}\|\widehat{s}_k(\boldsymbol{X}_t, t) - s^*(\boldsymbol{X}_t, t)\|_2^2 \mathrm{d}t.$$

Then, for every fixed $A > 0$ and every $1 \le k \le K$, almost surely,

$$\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k-1)}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k)}\big) \le C_A H\Big\{\sqrt{t_k \log n}\,\mathcal{L}_{I_k}(\widehat{s}_k)^{1/2} + n^{-A}\Big\}. \tag{307}$$

Here C_A depends only on A, C_{clip}, and fixed OU constants, not on (n, D, k) or $p_{\min}$.

Taking expectations and applying Jensen's inequality gives

$$\begin{aligned}&\mathbb{E}\,\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k-1)}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k)}\big)\\ &\qquad \le C_A H\Big\{\sqrt{t_k \log n}\,\big(\mathbb{E}\mathcal{L}_{I_k}(\widehat{s}_k)\big)^{1/2} + n^{-A}\Big\}.\end{aligned} \tag{308}$$

Proof. Fix the learned score networks. The two adjacent hybrids use the true score above t_k; hence, when the reverse dynamics enter the only differing slab I_k, their common entrance law is exactly P_{t_k}. Disintegrate both path laws conditionally on this entrance state $\boldsymbol{z}$. On I_k their drift difference is $2\{\widehat{s}_k - s^*\}$ and their diffusion matrix is $\sqrt{2}I_D$.

Apply the localized Girsanov argument used in diffusion-model sampling analyses (Chen et al., 2023c, Section 5.2) first up to the exit time from a ball of radius R, then let $R \to \infty$. The entropy chain rule, lower semicontinuity of relative entropy, and the fact that the true reverse marginal at forward time t is P_t give

$$\int P_{t_k}(\mathrm{d}\boldsymbol{z})\,\mathrm{KL}(\mathbb{Q}^{\mathrm{true}}_{\boldsymbol{z}} \| \mathbb{Q}^{\mathrm{learned}}_{\boldsymbol{z}}) \le C\mathcal{L}_{I_k}(\widehat{s}_k). \tag{309}$$

More precisely, let $\mathbb{Q}^R$ be the joint law of the entrance state and the process that follows the true drift until exiting the radius-R ball and follows the learned drift afterwards. The stopped drift difference is bounded, so the exponential martingale satisfies Novikov's condition and

$$\mathrm{KL}(\mathbb{Q}^R \| \mathbb{Q}^{\mathrm{learned}}) = \mathbb{E}_{\mathbb{Q}^{\mathrm{true}}} \int_{\text{slab before exit}} \|\widehat{s}_k - s^*\|_2^2 \mathrm{d}t \le \mathcal{L}_{I_k}(\widehat{s}_k).$$

The equality uses the noise coefficient $\sqrt{2}I_D$ and drift difference $2(\widehat{s}_k - s^*)$. Couple $\mathbb{Q}^R$ and the true process up to that exit time. The true process is nonexplosive on the positive-time slab, so its exit probability tends to zero and $\mathbb{Q}^R$ converges to $\mathbb{Q}^{\mathrm{true}}$ in total variation. Lower semicontinuity of relative entropy and the entropy chain rule for their common entrance law prove (309). The learned process is also nonexplosive: its drift is a linear OU term plus a bounded, spatially Lipschitz term on each fixed positive-time slab. After I_k, both hybrids use the same learned continuation. Data processing through that common continuation and conditional Pinsker's inequality (Tsybakov, 2009) therefore imply

$$\int P_{t_k}(\mathrm{d}\boldsymbol{z})\,\delta_k(\boldsymbol{z}) \le C\mathcal{L}_{I_k}(\widehat{s}_k)^{1/2}, \tag{310}$$

where $\delta_k(\boldsymbol{z})$ is the total variation distance between the two conditional endpoint laws given the common entrance state $\boldsymbol{z}$.

We next record the averaged localization needed to convert conditional total variation to Wasserstein distance. Suppose first that $t_k \le 1$. From equation 306, $\sigma_t^2 \asymp t$, and $D \le H$, every learned continuation satisfies

$$\|\widehat{s}(\cdot,t)\|_2 \le CHt^{-1/2}, \qquad \int_{t_0}^{t_k} \|\widehat{s}(\cdot,t)\|_2 \mathrm{d}t \le CH\sqrt{t_k}.$$

Gronwall's inequality for the linear OU drift and the Brownian maximal inequality then show that, for every fixed $B > 0$, the learned continuation remains within

$$R_{k,B} := C_B H \sqrt{t_k \log n}$$

of its entrance state outside an event whose probability, averaged over $\boldsymbol{z} \sim P_{t_k}$, is at most n^{-B}. The required entrance-state tail follows directly from $\boldsymbol{X}_{t_k} = e^{-t_k}\boldsymbol{X}_0 + \sigma_{t_k}\boldsymbol{Z}$, with $\|\boldsymbol{X}_0\|_2 \le \sqrt{D}$, and Gaussian concentration (Vershynin, 2018). For the hybrid using the true score on I_k, time reversal identifies that segment with the corresponding forward OU segment; the same Gaussian maximal bound localizes it, after which the identical clipped continuation gives the same radius. Thus both conditional endpoint laws are, on average over P_{t_k}, supported inside the ball $B(\boldsymbol{z}, R_{k,B})$ up to total mass Cn^{-B}.

For each $\boldsymbol{z}$, maximally couple the two conditional endpoint laws and then project both coordinates onto $\mathcal{C}$. On the successful part of the maximal coupling the coordinates agree. On its failure, but on the two localization events, nonexpansiveness of $\mathrm{Proj}_{\mathcal{C}}$ bounds their distance by $2R_{k,B}$ (use $\mathrm{Proj}_{\mathcal{C}}(\boldsymbol{z})$ as the common center). On the exceptional part, the diameter of $\mathcal{C}$ bounds the distance by $\sqrt{D}$. Averaging this coupling and using equation 310 yields

$$\mathsf{W}_1\big((\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k-1)}, (\mathrm{Proj}_{\mathcal{C}})_{\#}\overline{P}_{t_0}^{(k)}\big) \le C_B H \sqrt{t_k \log n}\,\mathcal{L}_{I_k}(\widehat{s}_k)^{1/2} + C\sqrt{D}\,n^{-B}.$$

Choose $B > A + 2$ and use $\sqrt{D} \le H$ to obtain equation 307.

If $t_k \ge 1$, bounded diameter and equation 310 directly give $\mathsf{W}_1 \le C\sqrt{D}\,\mathcal{L}_{I_k}^{1/2}$, which is no larger than the first term in equation 307. Finally, average over the learned score networks and apply Jensen's inequality to $\mathcal{L}_{I_k}^{1/2}$ to obtain equation 308. □