GEOMETRY-AWARE DIFFUSION THROUGH ANISOTROPIC PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based generative models perform remarkably well on data supported on lower-dimensional manifolds, despite relying on ambient forward processes that inject isotropic Gaussian noise in all directions. However, as the noise level vanishes, the score develops a singular normal component near the manifold, destabilizing score matching and hence complicating distribution estimation. We study a geometry-aware procedure that first perturbs data with Gaussian noise in the normal directions and then applies a standard ambient Ornstein-Uhlenbeck forward process, which regularizes the normal score without obscuring intrinsic density information. We establish non-asymptotic score estimation and distributional guarantees under γ-Hölder integral probability metrics for $0 < \gamma \leq 1$. Under suitable geometric and smoothness regularity conditions, the resulting rates depend on the intrinsic dimension and match the corresponding minimax rate up to logarithmic factors, yielding better rates in the small γ regime compared to the usual isotropic diffusion model. Empirical evaluations on synthetic manifolds and molecular dynamics validate our theoretical framework, demonstrating the practical advantages of geometry-aware perturbations.

1 INTRODUCTION

Generative modeling aims to learn a data distribution given a finite sample, with applications ranging from image and audio synthesis to molecular conformation generation and single-cell transcriptomics (Ho et al., 2020; Kong et al., 2020; Xu et al., 2022; Luo et al., 2024). Modern high-dimensional data often lies on or near a lower-dimensional manifold, and it has been shown that generative models such as diffusion models and flow matching can adapt to these intrinsic structures (Tang & Yang, 2024; Azangulov et al., 2024; Kumar et al., 2026). Specifically, recent theoretical work establishes that the underlying distribution estimators yield rates that are dependent on the intrinsic dimension of the data, including minimax-optimal guarantees under suitable smoothness assumptions and probability metrics (Oko et al., 2023; Tang & Yang, 2024; Azangulov et al., 2024). Chakraborty et al. (2026) generalize statistical properties of diffusion models to the (p, q)-Wasserstein dimension without requiring strict topological constraints, and Martinez et al. (2026) generalize generative modeling to learn structures of varying dimensions (stratified spaces).

Standard score-matching for data supported on a lower-dimensional manifold encounters practical and theoretical issues, particularly in the magnitude of the score. Although the diffusion process creates a smooth ambient density at every positive time, the associated score develops singular behavior in directions normal to the manifold as the diffusion time vanishes (Pidstrigach, 2022; Stanczuk et al., 2024). These insights motivate the question of how diffusion models, or more specifically the score function, can learn the geometry of the manifold. For example, Stanczuk et al. (2024) use the score to estimate the intrinsic dimension of the manifold, while Li et al. (2025); Shen et al. (2026) argue that learning the geometric support can be statistically easier than recovering the full target distribution. For $0 < \gamma < 1$, existing diffusion-based upper bounds can fail to attain the minimax rates in certain smoothness regimes under the γ-Hölder IPMs, motivating a closer examination of how geometric information can improve density estimation. Tang & Yang (2024) hypothesize that the gap is related to isotropic Gaussian smoothing in the forward process. This raises the question: *Can utilizing or learning the geometry of the data manifold before learning the score lead to a minimax-optimal diffusion-based estimator of the distribution?*

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.

To address this, we utilize the tangent and normal spaces of a manifold to anisotropically perturb the data. We perturb the observations in the (potentially estimated) normal directions, train an ambient OU diffusion model on the perturbed distribution, and project the generated samples onto the corresponding manifold. When the geometry is known, Keegan & Ruthotto (2026) establish (intrinsic) total variation bounds between the original distribution and the projection of the ideal perturbed distribution. Liu et al. (2025) provide a theoretical analysis of the singularity structure of the score via a tangent and normal decomposition and propose two approaches for training: one with anisotropic noise and one focusing only on the tangent component of the score, again in the known manifold case. Our focus is on the finite-sample statistical performance of this approach: we analyze the score approximation and estimation, propagate these errors to the generated distribution, and extend guarantees to the estimated manifold setting. Within a tubular neighborhood nearest-point projection reverses the normal perturbation, so we obtain ambient smoothing while maintaining the manifold-supported distribution after projection.

Our main contributions can be summarized as follows:

- We develop a nonasymptotic neural approximation and statistical learning theory for the anisotropically smoothed score.
- For each fixed $\gamma \in (0, 1]$, we establish distribution estimation guarantees under the γ-Hölder IPM for any α-Hölder density supported on a d-dimensional manifold under stated regularity assumptions. With known geometry, the rate is $n^{-(\alpha+\gamma)/(2\alpha+d)} + n^{-1/2}$ up to logarithmic factors. For an estimated β-smooth manifold with integer β and suitable reconstruction assumptions, our bound includes an additional geometric term, yielding the rate of $n^{-(\alpha+\gamma)/(2\alpha+d)} + n^{-1/2} + n^{-\beta\gamma/d}$ up to logarithmic factors, matching the minimax rate.
- We evaluate the procedure on a variety of manifolds and molecular dynamics trajectories. We compare standard diffusion models, projected samples, and the normally perturbed model under Hölder transport costs.

Related Work Empirically, projection and noise schedule designs have demonstrated significant benefits for generative modeling. For instance, Jing et al. (2022) restrict the diffusion via projections as the data distribution evolves, and Christopher et al. (2024) modify the sampling process by incorporating projections to satisfy constraints. Moreover, anisotropic stochastic PDEs have been utilized for diffusion in order to preserve geometric structure in data for longer (Holl et al., 2026).

Our focus on the γ-Hölder IPM for $0 < \gamma \leq 1$ provides a natural connection to the theoretical foundations of adversarial learning. Distribution estimation using generative adversarial networks (GANs) and Wasserstein-GANs has been extensively studied under these metrics (Liang, 2021; Singh et al., 2018). Tang & Yang (2023) establish minimax rates of learning distributions on unknown submanifolds under adversarial losses for $0 < \gamma \leq 1$. For $\gamma \geq 1$, Stéphanovitch (2025) develops a generative model that is a minimax-optimal distribution estimator.

Another line of work involves an intrinsic geometry perspective, where one can define SDEs on a manifold and establish corresponding distribution estimation guarantees (De Bortoli et al., 2022; Huang et al., 2022; Fu et al., 2026). These approaches formulate diffusion directly on the manifold and require manifold-specific geometric operations. Our process uses ambient diffusion with geometric preprocessing and projection, which accommodates estimated geometry.

2 Preliminaries

Let D and d be fixed integers such that $1 \leq d < D$. For a noninteger $s = k + \vartheta$ with $k \in \mathbb{N}_0$ and $\vartheta \in (0, 1)$, we use C^s to mean the Hölder space $C^{k,\vartheta}$. For integer values $s = k$, C^k denotes the class of k-times continuously differentiable functions (Giné & Nickl, 2021).

2.1 Geometric and distributional setup

Let $\mathcal{M} \subset \mathbb{R}^D$ be a d-dimensional compact embedded submanifold without boundary. We let p_* denote the target probability measure and f denote its density with respect to the Riemannian volume measure $\mathrm{vol}_{\mathcal{M}}$.

Assumption 2.1 (Regularity of data manifold)**.** *The manifold $\mathcal{M}$ is β-Hölder smooth for some $\beta \geq 2$. Moreover, there exists $R_0 > 0$ such that $\mathcal{M} \subset B(0, R_0)$ and $\mathcal{M}$ has reach $\rho > 0$. Its atlas has uniform chart, inverse-chart, and C^β bounds as specified in Assumption A.1.*

Assumption 2.2 (Regularity of density)**.** *The target distribution p_* satisfies $p_* = f \, d\,\mathrm{vol}_{\mathcal{M}}$, where $f \in C^\alpha$ for some $\alpha \in (0, \beta - 1]$, and f is uniformly bounded away from zero on $\mathcal{M}$.*

These geometric and density regularity assumptions are standard in theoretical analyses of distributions supported on unknown submanifolds (Tang & Yang, 2023; 2024). The parameter β controls the regularity of the charts and tangent fields, while α controls intrinsic regularity of the density. The lower bound on f provides uniform local mass and supports lower bounds on the smoothed density near the manifold. The positive-reach condition guarantees that the nearest-point projection map

$$\Pi_{\mathcal{M}}(x) = \arg\min_{z \in \mathcal{M}} \|x - z\|$$

is uniquely defined within a tubular neighborhood of the manifold.

For each $z \in \mathcal{M}$, let $T_z\mathcal{M}$ and $N_z\mathcal{M} = (T_z\mathcal{M})^\perp$ denote the tangent and normal spaces of $\mathcal{M}$ at z. We write

$$P_T(z) : \mathbb{R}^D \to T_z\mathcal{M}, \qquad P_N(z) : \mathbb{R}^D \to N_z\mathcal{M}$$

for the corresponding orthogonal projection matrices. These satisfy the relations

$$P_N(z) = I_D - P_T(z), \quad P_T(z)^\top = P_T(z), \quad P_T(z)^2 = P_T(z),$$

and analogous identities for $P_N(z)$. Equivalently, for any fixed z, if $U(z) \in \mathbb{R}^{D \times d}$ is any orthonormal frame for $T_z\mathcal{M}$, then

$$P_T(z) = U(z)U(z)^\top, \qquad P_N(z) = I_D - U(z)U(z)^\top.$$

These definitions are independent of the choice of frame.

2.2 Hölder integral probability metrics

To quantify the discrepancy between the generated and target distributions, we measure distributional convergence via integral probability metrics (IPMs). For any $\gamma \in (0, 1]$, the γ-Hölder IPM between distributions μ and ν is defined as

$$d_\gamma(\mu, \nu) = \sup_{g : \|g(x) - g(y)\| \leq \|x - y\|^\gamma} \left| \int g \, d\mu - \int g \, d\nu \right|.$$

For $\gamma = 1$, this reduces to W_1, the 1-Wasserstein distance. The choice of γ allows for flexibility in the desired model: smaller values of γ make the metric more sensitive to support displacements Tang & Yang (2023). This sensitivity is reflected in the geometric component of the minimax rate in the regime where $\frac{\beta\gamma}{d} < \min\{\frac{1}{2}, \frac{\alpha+\gamma}{2\alpha+d}\}$. Thus, sufficiently small discriminator smoothness γ can make the manifold reconstruction term the statistical bottleneck.

2.3 Normal perturbation and Ornstein-Uhlenbeck Processes

Fix a normal perturbation level $\eta > 0$ and a field of rank-d orthogonal projections $Q_T(z)$, $z \in \mathcal{M}$, with $Q_N = I_D - Q_T$. We view $Q_N(z)$ as the (possibly estimated) normal projection at z. The canonical choice is $Q_N = P_N$, and Section 3.1 gives the choice built from a learned manifold. Given $Z \sim p_*$, define

$$X_0^Q = Z + \eta \, Q_N(Z)\xi, \qquad \xi \sim \mathcal{N}(0, I_D),$$

where ξ is independent of Z. Equivalently, $X_0^Q \mid Z = z \sim \mathcal{N}\big(z, \eta^2 Q_N(z)\big)$. We denote the law of X_0^Q by μ_η^Q.

Starting from $X_0^Q \sim \mu_\eta^Q$, consider the ambient Ornstein-Uhlenbeck (OU) process

$$dX_t = -b_t X_t dt + \sqrt{2b_t} dB_t$$

where $\{B_t\}_{t \geq 0}$ is Brownian motion in $\mathbb{R}^D$ and the drift satisfies $0 < b_- \leq b \leq b_+ < \infty$. Define

$$m_t = \exp\left(-\int_0^t b_s ds\right), \qquad \sigma_t^2 = 1 - m_t^2$$

The OU process admits the closed-form solutions $X_t = m_t X_0^Q + \int_0^t \frac{m_t}{m_s}\sqrt{2b_s}dB_s$ with conditional marginals $p_t(x|X_0^Q) = \mathcal{N}(m_t X_0^Q, \sigma_t^2 I_D)$. Conditioned on $Z = z$,

$$X_t = m_t X_0^Q + \sigma_t \epsilon = m_t z + m_t \eta\, Q_N(z)\xi + \sigma_t \epsilon.$$

Therefore $\mathbb{E}[X_t \mid Z = z] = m_t z$ and, using $Q_N(z)^2 = Q_N(z)$,

$$\mathrm{Cov}(X_t \mid Z = z) = m_t^2\eta^2 Q_N(z) + \sigma_t^2 I_D = \sigma_t^2 Q_T(z) + (\sigma_t^2 + m_t^2\eta^2)Q_N(z).$$

Set

$$\nu_t^2 = \sigma_t^2 + m_t^2\eta^2, \qquad \Sigma_t(z) = \sigma_t^2 Q_T(z) + \nu_t^2 Q_N(z).$$

The conditional variance is σ_t^2 along the range of $Q_T(z)$ and ν_t^2 along the range of $Q_N(z)$, with no cross-covariance.

Let p_t^Q denote the density of X_t and define its score by

$$s_t^Q(x) = \nabla_x \log p_t^Q(x).$$

Although the perturbation applies additional variance in the normal directions, the OU process remains the standard isotropic diffusion. Standard time-reversal results (Song et al., 2020; Haussmann & Pardoux, 1986) give us a reverse process

$$d\overleftarrow{X}_t = [b_{T-t}\overleftarrow{X}_t + 2b_{T-t}s_{T-t}^Q(\overleftarrow{X}_t)]dt + \sqrt{2b_{T-t}}dB_t, \qquad \overleftarrow{X}_0 \sim p_T^Q$$

with the distribution of $\overleftarrow{X}_t$ being p_{T-t}^Q. Its marginal law at reverse time $t < T$ is p_{T-t}^Q, and its terminal law at $t = T$ is μ_η^Q. In practice, a common choice is taking b_t constant so that p_t converges exponentially quick to the limiting distribution $\mathcal{N}(0, I_D)$, and one generates samples by taking $\overleftarrow{X}_0 \sim \mathcal{N}(0, I_D)$.

2.4 Score Matching and the Reverse Process

Because the exact score of the forward diffusion process is generally intractable, we estimate it in practice via denoising score matching (Song et al., 2020; Vincent, 2011). For standard isotropic diffusion, this involves minimizing a weighted squared-error loss against the exact conditional score $\nabla_{x_t} \log p_t(x_t|x_i)$.

We observe i.i.d. samples $Z_1, \ldots, Z_n \sim p_*$ and form augmented observations $X_i^Q = Z_i + \eta\, Q_N(Z_i)\xi$ as in Section 2.3. When Q_N is learned (Section 3.1), it is built from a geometry sample independent of the score-training sample. The forward diffusion then yields $X_{t,i} = m_t X_i^Q + \sigma_t \epsilon$, where $\xi, \epsilon \sim \mathcal{N}(0, I_D)$ are independent of each other and of Z_i. Because the transition distribution conditional on $X_{0,i}$ remains purely isotropic Gaussian, the exact conditional score is simply $-\epsilon/\sigma_t$. Thus, the empirical denoising loss for our perturbed distribution is:

$$\widehat{\mathcal{L}}(S_\theta) = \frac{1}{n}\sum_{i=1}^{n}\int_\tau^T \mathbb{E}_{\xi,\epsilon}\left[\left\| S_\theta(X_{t,i}, t) + \frac{\epsilon}{\sigma_t}\right\|^2\right]\lambda(t)\,dt, \tag{1}$$

For the theoretical estimator, we take $\lambda(t) = t$ and minimize the loss separately over the blockwise network classes defined below, as specified in Section C.2.

Given the empirical minimizer $\widehat{S}^Q$, we simulate the generative reverse SDE:

$$d\overleftarrow{Y}_t = b_{T-t}\big[\overleftarrow{Y}_t + 2\widehat{S}^Q(\overleftarrow{Y}_t, T-t)\big]\,dt + \sqrt{2b_{T-t}}\,d\bar{B}_t, \qquad \overleftarrow{Y}_0 \sim \mathcal{N}(0, I_D),$$

for $t \in [0, T-\tau]$. Let $\widehat{\mu}_\tau^{\widehat{S}^Q}$ denote the conditional law of the terminal state $\overleftarrow{Y}_{T-\tau}$ given the training data. Our final distribution estimator is constructed by projecting this state onto the manifold:

$$\widehat{p} = (\mathsf{R}_{\mathcal{M}_{\mathrm{out}}})_{\#}\widehat{\mu}_\tau^{\widehat{S}^Q},$$

where $\mathcal{M}_{\mathrm{out}} = \mathcal{M}$ for the known-geometry case and $\mathcal{M}_{\mathrm{out}} = \widehat{\mathcal{M}}$ for the learned-geometry case (Section 3.1). The operator $\mathsf{R}_{\mathcal{M}_{\mathrm{out}}}$ agrees with nearest-point projection inside a fixed tubular neighborhood of the manifold and evaluates to an arbitrary fixed point on $\mathcal{M}_{\mathrm{out}}$ outside that neighborhood.

To analyze the statistical capacity of this procedure, we formulate the hypothesis class of score estimators S_θ using deep ReLU networks, where $\mathrm{ReLU}(x) = \max(0, x)$ is the rectified linear unit.

Definition 2.3. *The class of neural networks $\Phi(L, W, R, B, V)$ with depth L, width vector $W = (W_0, \ldots, W_L)$, sparsity R, parameter norm constraint B, and uniform output constraint V consists of all functions $f = T_L \circ \mathrm{ReLU} \circ \cdots \circ \mathrm{ReLU} \circ T_1$, where $T_i(x) = A^{(i)}x + b^{(i)}$ with $A^{(i)} \in \mathbb{R}^{W_i \times W_{i-1}}$ and $b^{(i)} \in \mathbb{R}^{W_i}$, such that $\sum_{i=1}^{L}(\|A^{(i)}\|_0 + \|b^{(i)}\|_0) \le R$, $\max_{1 \le i \le L}(\|A^{(i)}\|_{\max} \vee \|b^{(i)}\|_\infty) \le B$, and $\|f\|_\infty \le V$.*

3 Main results

In this section, we show that diffusion models can achieve (near) minimax-optimality in distribution estimation under adversarial losses for $0 < \gamma \le 1$ using normal perturbations, both in the known and learned geometry setting. For a sequence $\{a_n : n = 1, 2, ...\}$, we use $O(a_n)$ to say of order at most a_n up to a multiplicative constant for sufficiently large n. Similarly, we use $\widetilde{O}$ to say of order at most a_n up to a multiplicative constant and logarithmic factors for sufficiently large n.

3.1 Assumptions

Our analysis allows the field Q_T of Section 2.3 to come from either the true manifold or a learned one. In the known-geometry setting, $Q_T = P_T$. In the learned-geometry setting, we take

$$Q_T(z) = P_{T,\widehat{\mathcal{M}}}\big(\Pi_{\widehat{\mathcal{M}}}(z)\big), \qquad z \in \mathcal{M},$$

where $\widehat{\mathcal{M}}$ is a manifold estimate constructed from an independent geometry sample. On a high-probability event, $\widehat{\mathcal{M}}$ is a compact embedded d-dimensional C^β manifold without boundary, contained in a fixed bounded set, with uniformly positive reach and uniformly controlled C^β charts, and $d_H(\widehat{\mathcal{M}}, \mathcal{M})$ is sufficiently small. The precise reconstruction and failure-probability conditions, and the extension of Q_T off this event, are given in Assumption D.1 and Appendix D.

Assumption 3.1. *$Q_T(y)$ is a rank-d orthogonal projection on $\mathcal{M}$ with $Q_N = I_D - Q_T$ and admits a local orthonormal frame U_j^* on each uniformly regular chart G_j^* with*

$$Q_T(G_j^*(z)) = U_j^*(z)U_j^*(z)^\top, \quad \|U_j^*\|_{C^{\beta-1}} \le C_Q, \quad \varepsilon_P := \sup_{y \in \mathcal{M}} \|Q_T(y) - P_T(y)\|_{\mathrm{op}} \le c_Q \le \frac{1}{2}. \tag{2}$$

The regularity condition of Assumption 3.1 controls the variation of the perturbation directions across charts, while the spectral norm bound $\varepsilon_P < 1$ guarantees non-degeneracy.

We also impose the following condition on the drift:

Assumption 3.2. *There exist constants $C_b, A_b > 0$ such that, for every integer $j \ge 0$,*

$$\sup_{t \ge 0} \left|\nabla^j b(t)\right| \le C_b A_b^j j!.$$

This derivative condition is only used to approximate the scalar noise schedule functions by polylogarithmic-size ReLU networks. A constant schedule b_t is often used in practice and satisfies this condition.

3.2 Score Approximation

Normal perturbation changes the conditional variance to $\nu_t^2 = m_t^2\eta^2 + \sigma_t^2$, which remains bounded below by $m_t^2\eta^2$. This resolves part of the difficulty in current theoretical analyses of isotropic diffusion models by regularizing the score while preserving the intrinsic density-estimation problem. The main approximation difficulty then occurs at small diffusion times, where the tangential variance vanishes and the score must unravel the local geometry. The following theorem establishes the convergence rates in both the small-time and high-time regime. We first suppose we have a partition $\tau = t_0 < t_1 < \cdots < t_K = T$ with $1 < \frac{t_{k+1}}{t_k} \le 2$ and contains $n^{-\frac{2}{2\alpha+d}}$ and 1, and we set $\eta \asymp n^{-1/(2\alpha+d)}$, $\tau = n^{-2(\alpha+\gamma)/[\gamma(2\alpha+d)]}$ and $T \asymp \log n$.

Theorem 3.3 (Score approximation)**.** *Suppose Assumptions 2.1,2.2, 3.1, and the drift condition 3.2 hold. For $V_k = C_V\sqrt{\log n}/\sigma_{t_k}$ with C_V sufficiently large, each block admits a network $\varphi_k \in \Phi(L_k, W_k, R_k, B_k, V_k)$ with $L_k = O((\log n)^4)$ and $B_k = \exp(O((\log n)^4))$ satisfying*

$$\int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \|\varphi_k(x,t) - \nabla \log p_t^Q(x)\|^2 p_t^Q(x)\, dx\, dt = \begin{cases} \widetilde{O}(n^{-\frac{2\alpha}{2\alpha+d}}), & t_{k+1} \le h^2, \\ \widetilde{O}(n^{-1}), & t_k \ge h^2. \end{cases}$$

Here $\|W_k\|_\infty + R_k = \widetilde{O}(n^{\frac{d}{2\alpha+d}})$ on small-time blocks and $\widetilde{O}(1 + ((t_k \wedge 1)\log n)^{-d/2})$ on higher-time blocks. The bounds are uniform over fields Q_T satisfying Assumption 3.1.

The proof of the theorem is given in Appendix B. This theorem has the following consequences: first, the network complexity depends polynomially on the intrinsic resolution $n^{d/(2\alpha+d)}$. Second, the bound is uniform over the fields Q_T, which covers both the cases $Q_T = P_T$ for the known geometry and $Q_T = P_{T,\widehat{M}}(\Pi_{\widehat{M}})$ for the learned geometry.

3.3 Distribution Estimation

In standard diffusion model theory, one employs Girsanov's theorem to translate the score approximation bounds to KL divergence guarantees. However, to establish guarantees under γ-Hölder IPMs, we must account for the anisotropic geometry of the perturbed data. The recovery rate ultimately depends on how the reverse process propagates through this perturbation and how the final projection map collapses the residual ambient error.

Theorem 3.4 (Distribution Recovery)**.** *Under the assumptions and scales above, take C_T sufficiently large and let $\widehat{S}$ be the blockwise empirical minimizer of equation C.5, with network budgets containing the approximators of Theorem A.4. Then:*

1. *If the geometry is known and $Q_T = P_T$, then*

$$\mathbb{E}d_\gamma\left((\mathsf{R}_{\mathcal{M}})_{\#}\widehat{\mu}_\tau^{\widehat{S}}, p_*\right) = \widetilde{O}(n^{-1/2} + n^{-(\alpha+\gamma)/(2\alpha+d)}).$$

2. *For integer $\beta \ge 2$, connected $\mathcal{M}$, and the regular reconstruction of Corollary D.4, use independent geometry and score samples of sizes proportional to n. With $Q_T = P_{T,\widehat{\mathcal{M}}} \circ \Pi_{\widehat{\mathcal{M}}}$ on the regular reconstruction event,*

$$\mathbb{E}d_\gamma\left((\mathsf{R}_{\widehat{\mathcal{M}}})_{\#}\widehat{\mu}_\tau^{\widehat{S}^Q}, p_*\right) = \widetilde{O}(n^{-1/2} + n^{-(\alpha+\gamma)/(2\alpha+d)} + n^{-\beta\gamma/d}).$$

Appendices C.2 and D prove the two claims. More generally, any reconstruction satisfying Assumption D.1 gives the same bound with $n^{-\beta\gamma/d}$ replaced by ϵ_M^γ, where $\epsilon_M = \sup_{z\in\mathcal{M}} \mathrm{dist}(z, \widehat{\mathcal{M}})$ (Theorem D.3). The error terms in the final upper bound of Theorem 3.4 come from different sources. The term $n^{-1/2}$ comes from the parametric component of distribution estimation, $n^{-(\alpha+\gamma)/(2\alpha+d)}$ is the cost of learning an α-smooth density on a d-dimensional manifold, and $n^{-\beta\gamma/d}$ is the rate of recovering the unknown β-smooth support (Tang & Yang, 2023; Aamari & Levrard, 2019). This closes the gap left by the isotropic diffusion analysis, where the suboptimal rate $\widetilde{O}(n^{-1/2} + n^{-(\alpha+\gamma)/(2\alpha+d)} + n^{-\beta\gamma/(2\alpha+d)})$. A consequence of injecting Gaussian noise exclusively in the normal directions is that the uncorrupted distribution can be recovered directly via nearest-point projection. Perturbation with other forms of noise can slow down distribution recovery, with some classical examples in (Fan, 1991). We leave the choice of more general anisotropic perturbations as future work.

Remark 3.5. *If Q_N is estimated separately from $\widehat{\mathcal{M}}$, for instance by local PCA, then projection may not completely cancel the perturbation. Comparing $\Pi_{\widehat{\mathcal{M}}}$ with $\Pi_{\mathcal{M}}$ on the η-tube instead bounds the geometric bias by $\delta_0^\gamma + (\eta\delta_1)^\gamma + (\eta\epsilon_Q)^\gamma$ up to logarithmic factors, where δ_0, δ_1 are the chart and chart-derivative errors of $\widehat{\mathcal{M}}$ and $\epsilon_Q = \|Q_N - P_N\|_\infty$. With $\delta_1, \epsilon_Q = O(n^{-(\beta-1)/d})$ and $\eta \asymp n^{-1/(2\alpha+d)}$, this is $n^{-\gamma\{(\beta-1)/d+1/(2\alpha+d)\}}$, which is slower than $n^{-\beta\gamma/d}$. The compatible choice $Q_N = P_{N,\widehat{\mathcal{M}}} \circ \Pi_{\widehat{\mathcal{M}}}$ removes the derivative terms.*

To our knowledge, this is the first work to establish finite-sample guarantees for both score estimation and distribution estimation for diffusion with normal-direction perturbation and projection, including when the geometry is estimated. Liu et al. (2025) analyze the singular structure of the score through a tangent–normal decomposition, and Keegan & Ruthotto (2026) bound the intrinsic total variation distance between the target and the projection of the ideal perturbed distribution. Both assume known geometry and do not analyze estimation from finite samples.

4 EXPERIMENTS

We compare three diffusion configurations to separate the effects of final projection and normal perturbation:

- **Model A (Baseline):** a variance-preserving diffusion model trained on the original observations and sampled to the forward-time cutoff τ, without a final geometric correction.
- **Model B (Final Projection):** the same trained model and reverse trajectories as Model A, followed by a learned local affine projection.
- **Model C (Normal Perturbation and Final Projection):** a separate diffusion model trained on normally perturbed observations, using the same architecture, reverse sampler, and final projection procedure as Model B.

The comparison between A and B isolates the effect of the final projection, while the comparison between B and C assesses the additional effect of normal perturbation. All reverse processes use isotropic diffusion. Note that Model B is the special case of $\eta = 0$ in Model C.

4.1 SYNTHETIC DATASETS

We evaluate the performance of the proposed method on a suite of synthetic manifolds covering a range of ambient dimensions, curvatures and topological complexities in a simulation study. These include the Klein Bottle, Swiss Roll, Torus and Sphere. On each manifold we sample points from its natural parameterization and map their coordinates into ambient dimension $D \in \{3, 4, 5, 6\}$ using a random isometric linear map. We use $N = 3000$ training points (6000 for the Klein bottle), a disjoint set of 3000 points to estimate the local manifold geometry, and a further 2000 held-out points for evaluation.

Geometry estimation and perturbation. At each geometry point a_j, we form the matrix of offsets to its 30 nearest neighbors and use its leading d right singular vectors to estimate the tangent space. We use the known intrinsic dimension for the synthetic datasets. Let $\widehat{P}_{N,j}$ denote the complementary orthogonal projector and $j(x) = \arg\min_j \|x - a_j\|$. Model C uses

$$Z_i^\eta = Z_i + \eta \widehat{P}_{N,j(Z_i)} \xi_i, \qquad \xi_i \sim \mathcal{N}(0, I_D)$$

with one independent perturbation drawn per observation and retained throughout training. Models B and C apply the same final map

$$\widehat{R}(x) = x - \widehat{P}_{N,j(x)}(x - a_{j(x)}).$$

Training setting. All three models use the same MLP architecture with four hidden layers of width 512 and SiLU activations, trained with the standard noise-prediction objective for 30,000 steps with batch size 64 using AdamW (learning rate 3×10^{-4}). We use a variance-preserving SDE with $\beta(t)$ linear from 0.1 to 20, and sample diffusion times $t \sim \mathrm{Uniform}(0.01, 1)$ to avoid the singularity as $t \to 0$. Model C uses $\eta = 0.1$.

Evaluation. We compare 2000 generated samples with 2000 held-out observations using the empirical d_γ for $\gamma \in \{0.25, 0.5, 0.75, 1\}$, that is, exact optimal transport with cost $\|x - z\|^\gamma$, computed with the unregularized earth mover's distance solver in POT; $\gamma = 1$ gives the empirical W_1 distance. All distances are computed in the normalized ambient coordinates, not in the three-dimensional projections used for visualization. Results are the mean and standard deviation over three seeds.

Table 1: d_γ between generated samples from Models A, B, and C, and a held-out sample of p_*, means over three seeds, with standard deviations shown as subscripts; lower is better. D represents ambient dimension.

	$\gamma = 0.25$			$\gamma = 0.5$			$\gamma = 0.75$			$\gamma = 1$ (W_1)		
D	A	B	C	A	B	C	A	B	C	A	B	C
Klein Bottle												
3	$.613_{.011}$	$.555_{.011}$	$\mathbf{.549}_{.006}$	$.389_{.017}$	$.329_{.017}$	$\mathbf{.318}_{.008}$	$.254_{.019}$	$.208_{.021}$	$\mathbf{.193}_{.008}$	$.169_{.018}$	$.136_{.020}$	$\mathbf{.120}_{.007}$
4	$.625_{.007}$	$.556_{.008}$	$\mathbf{.553}_{.001}$	$.401_{.012}$	$.330_{.012}$	$\mathbf{.324}_{.001}$	$.264_{.013}$	$.208_{.015}$	$\mathbf{.200}_{.001}$	$.176_{.013}$	$.136_{.015}$	$\mathbf{.127}_{.001}$
5	$.631_{.003}$	$\mathbf{.552}_{.005}$	$.560_{.007}$	$.407_{.005}$	$\mathbf{.325}_{.008}$	$.336_{.014}$	$.267_{.006}$	$\mathbf{.203}_{.010}$	$.216_{.019}$	$.177_{.006}$	$\mathbf{.131}_{.010}$	$.144_{.021}$
6	$.642_{.010}$	$.553_{.012}$	$\mathbf{.552}_{.007}$	$.420_{.016}$	$.327_{.021}$	$\mathbf{.323}_{.010}$	$.280_{.019}$	$.206_{.027}$	$\mathbf{.198}_{.010}$	$.188_{.018}$	$.135_{.028}$	$\mathbf{.126}_{.009}$
Swiss Roll												
3	$.561_{.018}$	$.505_{.018}$	$\mathbf{.501}_{.008}$	$.343_{.031}$	$.293_{.032}$	$\mathbf{.285}_{.016}$	$.231_{.041}$	$.198_{.043}$	$\mathbf{.187}_{.023}$	$.165_{.046}$	$.150_{.051}$	$\mathbf{.137}_{.029}$
4	$.573_{.006}$	$\mathbf{.495}_{.005}$	$.501_{.003}$	$.348_{.009}$	$\mathbf{.277}_{.008}$	$.285_{.001}$	$.226_{.011}$	$\mathbf{.178}_{.010}$	$.185_{.005}$	$.153_{.011}$	$\mathbf{.127}_{.011}$	$.133_{.009}$
5	$.594_{.002}$	$.497_{.004}$	$\mathbf{.495}_{.005}$	$.370_{.005}$	$.279_{.007}$	$\mathbf{.276}_{.011}$	$.243_{.008}$	$.179_{.011}$	$\mathbf{.174}_{.016}$	$.165_{.010}$	$.128_{.013}$	$\mathbf{.120}_{.019}$
6	$.609_{.008}$	$.501_{.011}$	$\mathbf{.492}_{.011}$	$.386_{.014}$	$.284_{.017}$	$\mathbf{.271}_{.019}$	$.256_{.020}$	$.183_{.024}$	$\mathbf{.169}_{.025}$	$.174_{.023}$	$.131_{.029}$	$\mathbf{.117}_{.029}$
Torus												
3	$.556_{.006}$	$.496_{.008}$	$\mathbf{.495}_{.005}$	$.324_{.011}$	$.268_{.013}$	$\mathbf{.266}_{.008}$	$.196_{.014}$	$.157_{.017}$	$\mathbf{.155}_{.011}$	$.122_{.014}$	$.097_{.018}$	$\mathbf{.095}_{.012}$
4	$.573_{.005}$	$.493_{.003}$	$\mathbf{.491}_{.003}$	$.339_{.005}$	$.263_{.003}$	$\mathbf{.260}_{.005}$	$.206_{.003}$	$.152_{.003}$	$\mathbf{.148}_{.006}$	$.127_{.002}$	$.092_{.003}$	$\mathbf{.088}_{.005}$
5	$.590_{.009}$	$.495_{.011}$	$\mathbf{.490}_{.004}$	$.359_{.015}$	$.268_{.018}$	$\mathbf{.260}_{.007}$	$.224_{.018}$	$.160_{.022}$	$\mathbf{.149}_{.010}$	$.142_{.017}$	$.102_{.023}$	$\mathbf{.091}_{.012}$
6	$.606_{.003}$	$.495_{.009}$	$\mathbf{.493}_{.006}$	$.377_{.008}$	$.269_{.016}$	$\mathbf{.266}_{.009}$	$.240_{.012}$	$.160_{.021}$	$\mathbf{.158}_{.012}$	$.155_{.012}$	$.102_{.022}$	$\mathbf{.099}_{.012}$
Sphere												
3	$.544_{.004}$	$.501_{.004}$	$\mathbf{.498}_{.012}$	$.316_{.007}$	$.277_{.008}$	$\mathbf{.273}_{.018}$	$.196_{.009}$	$.171_{.010}$	$\mathbf{.165}_{.022}$	$.126_{.009}$	$.112_{.011}$	$\mathbf{.106}_{.024}$
4	$.562_{.006}$	$\mathbf{.493}_{.005}$	$.498_{.011}$	$.327_{.009}$	$\mathbf{.264}_{.008}$	$.274_{.016}$	$.197_{.011}$	$\mathbf{.153}_{.010}$	$.167_{.019}$	$.120_{.010}$	$\mathbf{.092}_{.010}$	$.108_{.021}$
5	$.588_{.006}$	$.504_{.007}$	$\mathbf{.499}_{.008}$	$.360_{.009}$	$.282_{.011}$	$\mathbf{.275}_{.013}$	$.230_{.011}$	$.176_{.013}$	$\mathbf{.168}_{.016}$	$.149_{.012}$	$.117_{.014}$	$\mathbf{.110}_{.018}$
6	$.600_{.013}$	$.500_{.013}$	$\mathbf{.493}_{.005}$	$.372_{.022}$	$.276_{.022}$	$\mathbf{.266}_{.007}$	$.238_{.028}$	$.169_{.030}$	$\mathbf{.157}_{.009}$	$.155_{.030}$	$.112_{.035}$	$\mathbf{.098}_{.010}$

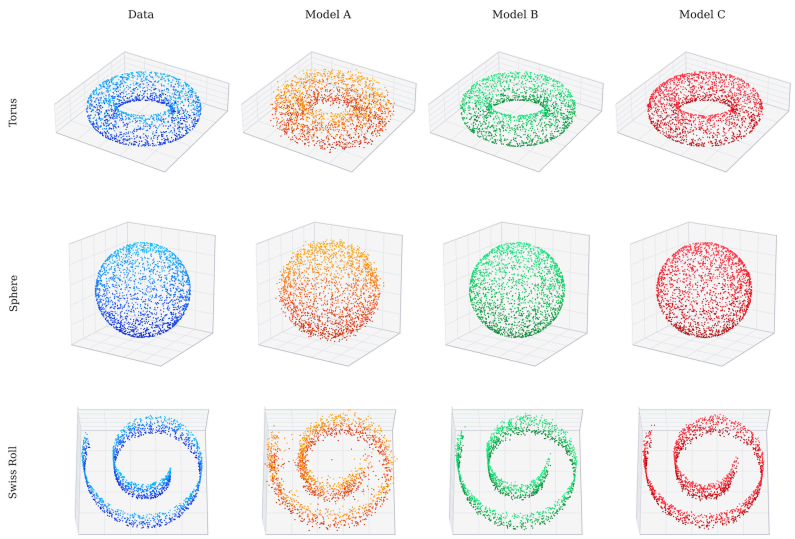

Figure 1: Samples generated by Models A, B and C for three manifolds embedded in ambient dimension $D = 6$, alongside the training data, visualized using PCA fitted on the training data. Color shading encodes depth along the third principal component.

Both projection-based methods reduce the empirical distributional error relative to Model A across all reported synthetic settings. Model C achieves a lower mean error than Model B in 52 of the 64 settings, although the improvements are generally modest. Baseline errors tend to increase with ambient dimension, with exceptions, and the projected models show less systematic variation.

4.2 Alanine Dipeptide Trajectories

We evaluate whether normal perturbation and final projection improve the generation of molecular conformations from an alanine dipeptide trajectory. The dataset contains $250{,}000$ conformations of 22 atoms, represented by Cartesian coordinates in $\mathbb{R}^{66}$. For computational efficiency, we subsample $8{,}000$ frames. We train Models A–C to generate conformations and compare their distributions with held-out trajectory frames using d_γ. For Model C, we vary the perturbation scale over $\eta \in \{0.02, 0.05, 0.10, 0.20\}$ to assess its effect on distributional accuracy. Motivated by the two dominant backbone dihedral angles ϕ and ψ, we use $d = 2$ as the working dimension for local tangent-space estimation, giving a normal-space dimension of 64. This is a modeling choice rather than an assumption that the Cartesian trajectory lies exactly on a two-dimensional manifold.

Table 2: Alanine dipeptide ($D = 66$, $d = 2$, codimension 64). d_γ between $m = 2000$ generated samples and a held-out sample of the trajectory, mean $\pm$ s.d. over three seeds; lower is better. $\gamma = 1$ recovers W_1.

Model	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 0.75$	$\gamma = 1$ (W_1)
Model A	0.938 $\pm$0.003	0.882 $\pm$0.005	0.831 $\pm$0.007	0.785 $\pm$0.009
Model B	0.812 $\pm$0.003	0.670 $\pm$0.007	0.562 $\pm$0.010	0.479 $\pm$0.014
Model C ($\eta = 0.02$)	0.804 $\pm$0.002	0.656 $\pm$0.004	0.543 $\pm$0.006	0.456 $\pm$0.007
Model C ($\eta = 0.05$)	0.802 $\pm$0.002	0.652 $\pm$0.004	0.536 $\pm$0.005	0.447 $\pm$0.005
Model C ($\eta = 0.10$)	**0.801** $\pm$0.007	**0.649** $\pm$0.013	**0.531** $\pm$0.017	**0.440** $\pm$0.021
Model C ($\eta = 0.20$)	0.813 $\pm$0.005	0.666 $\pm$0.008	0.551 $\pm$0.010	0.459 $\pm$0.011

On alanine dipeptide, all tested perturbation levels improve the mean error over Model B, with $\eta = 0.10$ giving the lowest reported means across all four metrics. These comparisons are descriptive, based on three seeds.

5 Discussion and Future Work

We show that perturbing along normal directions and projecting onto a compatible reconstruction attains the minimax rate for distribution estimation on an unknown manifold in the γ-Hölder IPM (Tang & Yang, 2023), up to logarithmic factors, for integer $\beta \geq 2$ under the stated regularity and reconstruction assumptions. Explicit geometric reconstruction replaces the term $n^{-\beta\gamma/(2\alpha+d)}$ in the upper bound of Tang & Yang (2024) by $n^{-\beta\gamma/d}$. Their geometric term exceeds their density term exactly when $\gamma(\beta - 1) < \alpha$, highlighting the relevance of this improvement for small γ. For W_1, their geometric term is already no larger than the density term under $\alpha \leq \beta - 1$.

In our experiments, both diffusion followed by projection and normal-direction perturbation followed by projection reduce distributional error relative to standard diffusion. Their similar performance suggests that projection accounts for much of the observed improvement. The experimental procedures also differ from the theoretical construction: the perturbation scale is fixed rather than scaled with sample size, and the normal field is estimated by local PCA rather than induced by the reconstruction used for projection. These differences may affect the relative performance of the methods. Evaluation at a single sample size also cannot establish their convergence rates. Sample-size sweeps with $\eta \asymp n^{-1/(2\alpha+d)}$, together with controlled variations in manifold smoothness and curvature, would help determine when normal perturbation provides an additional advantage.

A natural next step is to analyze computationally practical geometric estimators in place of the minimum-distance reconstruction used here. Other important extensions include observational noise and discretization of the reverse process. Extending the analysis to stratified supports, studied under isotropic Gaussian noise by Martinez et al. (2026), would require handling changes in dimension and the absence of a single smooth normal bundle.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, to provide critical ingredients for proving mathematical claims, to refine the formulation of mathematical claims, and to provide feedback on research methodology and experiments. Specifically, an LLM: (1) checked the appendix for logical gaps and for consistency of rates, logarithmic factors, and network complexity bounds, (2) proposed repairs to identified gaps, including proof steps that the authors verified before adopting, and (3) reviewed the experimental code and evaluation protocol for consistency with the theory. We have not used generative AI tools to generate synthetic data sets, develop the theoretical model or conceptual framework, propose or refine hypotheses, implement methods, or interpret results, and assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft or rewrite parts of the paper from the authors' drafts and derivations, edit it for readability and notational consistency, and summarize and analyze existing literature, and identify relevant literature. We have reviewed all AI-assisted work. Every mathematical statement and proof in the paper was checked line by line by the authors, and the main arguments were independently checked by a coauthor. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.

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

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.

Jacob K Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in neural information processing systems*, 35:2406–2422, 2022.

DLMF. NIST Digital Library of Mathematical Functions. `https://dlmf.nist.gov/`, 2026. Release 1.2.8 of 2026-09-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pp. 1257–1272, 1991.

Guoji Fu, Taiji Suzuki, Wee Sun Lee, and Atsushi Nitanda. Intrinsic wasserstein rates for score-based generative models on smooth manifolds. *arXiv preprint arXiv:2605.15822*, 2026.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pp. 1188–1205, 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sascha Holl, Jente Vandersanden, Gurprit Singh, and Hans-Peter Seidel. Score-based generative modeling through anisotropic stochastic partial differential equations. *arXiv preprint arXiv:2605.08976*, 2026.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022.

Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. Subspace diffusion generative models. In *European conference on computer vision*, pp. 274–289. Springer, 2022.

Katherine Keegan and Lars Ruthotto. Manifold-aware perturbations for constrained generative modeling. *arXiv preprint arXiv:2601.23151*, 2026.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Shivam Kumar, Yixin Wang, and Lizhen Lin. Flow matching is adaptive to manifold structures. *arXiv preprint arXiv:2602.22486*, 2026.

Gunther Leobacher and Alexander Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of global analysis and geometry*, 60(3):559–587, 2021.

Xiang Li, Zebang Shen, Ya-Ping Hsieh, and Niao He. When scores learn geometry: Rate separations under the manifold hypothesis. *arXiv preprint arXiv:2509.24912*, 2025.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

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

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024.

Randy Martinez, Rong Tang, and Lizhen Lin. A deep generative approach to stratified learning. *arXiv preprint arXiv:2604.10650*, 2026.

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

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

Zebang Shen, Ya-Ping Hsieh, and Niao He. Manifold generalization provably proceeds memorization in diffusion models. *arXiv preprint arXiv:2603.23792*, 2026.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *Advances in Neural Information Processing Systems*, 31, 2018.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

Arthur Stéphanovitch. Generative model for optimal density estimation on unknown manifold. *arXiv preprint arXiv:2506.19587*, 2025.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282–1308, 2023.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International conference on artificial intelligence and statistics*, pp. 1648–1656. PMLR, 2024.

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

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

APPENDIX

Overview. Section A states the model and the two main results. Section B proves Theorem 3.3 in the main text (restated in the Appendix as Theorem A.4): the higher-time regime $t \geq h^2$ is treated by localized moment expansions, following Tang & Yang (2024), and the small-time regime $t \leq h^2$ by a change of variables to coordinates matched to the field Q_T, which keeps the geometric residual on the normal scale ν_t. Section C.1 converts the score risk of the learned estimator into distribution error. Section C.2 bounds the score risk of the empirical minimizers and completes the proof of the first statement of Theorem 3.4 in the main text (restated in the Appendix as Theorem A.5). Section D replaces the exact normal field by one pulled back from a reconstructed manifold and proves the second statement of Theorem 3.4 in the main text.

A ASSUMPTIONS AND MAIN RESULTS

Throughout, $1 \leq d < D$ and D are fixed. For a scalar function, ∇ denotes its gradient; for a vector-valued map $G : \mathbb{R}^d \to \mathbb{R}^D$, ∇G is its $D \times d$ Jacobian, and $\nabla_b G$ denotes its bth partial derivative. Higher derivatives are understood as multilinear maps. Vector norms are Euclidean unless a subscript specifies otherwise, and matrix norms without a subscript are taken to be operator norms. The notation $T_y\mathcal{M}$ denotes the tangent space, and $(T_y\mathcal{M})^\perp$ its normal space. For noninteger $s = k + \theta$, C^s means $C^{k,\theta}$; at integer $s = k$, it means C^k with bounded derivatives through order k. For a multiindex $\iota \in \mathbb{N}_0^m$, write $|\iota| = \sum_{j=1}^m \iota_j$ and $x^\iota = \prod_{j=1}^m x_j^{\iota_j}$. The symbol $\mathrm{N}(a, \Sigma)$ denotes a Gaussian with mean a and covariance Σ.

Assumption A.1. *The data manifold $\mathcal{M} \subset \mathbb{R}^D$ is a d-dimensional compact β-Hölder smooth boundaryless manifold for some $\beta \geq 2$. We assume $\mathcal{M} \subset B(0, R_0)$ for some $R_0 > 0$ and has reach $\rho_0 > 0$. Suppose $\mathcal{M}$ has an atlas with uniform chart, inverse-chart, and C^β bounds. Assume we observe $Z_1, ..., Z_n$ i.i.d. with distribution $p_* = f \, d\,\mathrm{vol}_\mathcal{M}$, where $0 < f_{\min} \leq f \leq f_{\max}$, $f \in C^\alpha$ with $\|f\|_{C^\alpha(\mathcal{M})} \leq K_f$ and $0 < \alpha \leq \beta - 1$.*

Let $P_T(y)$ and $P_N(y) = I_D - P_T(y)$ denote the tangent and normal projectors of $\mathcal{M}$. When the true projection operators are unavailable, we assume that we have a suitable approximation $Q_T(y)$, such as $P_{T,\widehat{\mathcal{M}}}(\pi(z))$, where $P_{T,\widehat{\mathcal{M}}}$ is the projection map onto the tangent space of $\widehat{\mathcal{M}}$ and $\pi(z) = \Pi_{\widehat{\mathcal{M}}}(z)$ is the projection onto the learned manifold $\hat{\mathcal{M}}$.

Assumption A.2. *$Q_T(y)$ is a rank-d orthogonal projection on $\mathcal{M}$ with $Q_N = I_D - Q_T$ and admits a local orthonormal frame U_j^* on each uniformly regular chart G_j^* with*

$$Q_T(G_j^*(z)) = U_j^*(z)U_j^*(z)^\top, \qquad \|U_j^*\|_{C^{\beta-1}} \leq C_Q, \qquad \epsilon_P := \sup_{y\in\mathcal{M}} \|Q_T(y) - P_T(y)\|_{\mathrm{op}} \leq c_Q \leq \tfrac{1}{2}. \tag{A.1}$$

In the special case of $Q_T = P_T$, the chart derivatives $U^* = \nabla G^*((\nabla G^*)^\top \nabla G^*)^{-1/2}$ supply these frames with $\epsilon_P = 0$. Our first lemma ensures that local projections between the true and estimated geometry remain well-conditioned.

Lemma A.3. *Under Assumption A.2, on every chart,*

$$\sigma_{\min}\big(U^{*\top} \nabla G^*\big) \;\geq\; \sqrt{1 - \epsilon_P^2}\, \sigma_{\min}(\nabla G^*) \;\geq\; \tfrac{\sqrt{3}}{2}\, \sigma_{\min}(\nabla G^*) \;\geq\; c > 0. \tag{A.2}$$

Proof. For $v = \nabla G^* w \in T_y\mathcal{M}$, orthogonality gives

$$\|U^{*\top} v\|^2 = \|Q_T v\|^2 = \|v\|^2 - \|(P_T - Q_T)v\|^2 \geq (1 - \epsilon_P^2)\|v\|^2.$$

Minimizing over unit vectors w and using the uniform atlas conditioning gives us the claim. □

Now fix a radius $r_{\text{geo}} > 0$ depending only on the uniform atlas and the reach ρ_0 such that the following property holds: for every $y \in \mathcal{M}$, the set $\mathcal{M} \cap B(y, 4r_{\text{geo}})$ is contained in a parametrized neighborhood on which uniform chart regularity and inverse-Lipschitz bounds hold. Note that such a radius exists by compactness and Assumption A.1. On the tubular neighborhood $\{x : \mathrm{dist}(x, \mathcal{M}) < \rho_0\}$, we define the nearest-point projection map

$$\Pi_\mathcal{M}(x) = \arg\min_{y\in\mathcal{M}} \|x - y\|.$$

A.1 Anisotropic Perturbation and Forward Process

For a particular projector Q_T, $y \in \mathcal{M}$ and $\eta > 0$, we perturb the data via

$$X_0 = Z + \eta Q_N(Z)\xi, \qquad \xi \sim \mathrm{N}(0, I_D) \text{ independently of } Z, \qquad X_0 \mid Z = y \sim \mathrm{N}(y, \eta^2 Q_N(y)). \tag{A.3}$$

Write $Q = (Q_T, Q_N)$ for the chosen projector pair and $P = (P_T, P_N)$ for the exact pair. Superscripts Q and P label these models. We denote $\mu_\eta^Q = \mathcal{L}(X_0)$. At each fixed t, the forward marginal has the representation $X_t = m_t X_0 + \sigma_t \xi_{\text{iso}}$, where $\xi_{\text{iso}} \sim \mathrm{N}(0, I_D)$ is independent of X_0. Let b_t be a drift coefficient that is infinitely differentiable with $0 < b_- \le b_t \le b_+ < \infty$. For the network-complexity claims, assume in addition

$$\sup_{t \ge 0} |\nabla^j b_t| \le C_b A_b^j j!, \qquad j = 0, 1, 2, \ldots, \tag{A.4}$$

for fixed $C_b, A_b > 0$. We consider the OU process

$$dX_t = -b_t X_t \, dt + \sqrt{2b_t} \, dB_t, \qquad m_t = \exp\Big(-\int_0^t b_s \, ds\Big), \qquad \sigma_t^2 = 1 - m_t^2. \tag{A.5}$$

Here B_t is standard Brownian motion in $\mathbb{R}^D$. Note that $1 - m_t \asymp t \wedge 1$ and $\sigma_t^2 \asymp t \wedge 1$. The conditional marginals of the diffusion process then follow

$$X_t \mid Z = y \sim \mathrm{N}(m_t y, \Sigma_t(y)), \quad \Sigma_t(y) = \sigma_t^2 Q_T(y) + \nu_t^2 Q_N(y), \quad \nu_t^2 = \sigma_t^2 + m_t^2 \eta^2. \tag{A.6}$$

For the fixed pair Q, write $X_t = X_t^Q$, $p_t = p_t^Q$, and $s_t = s_t^Q = \nabla_x \log p_t^Q$. We suppress this model label throughout the score and transfer proofs. A superscript P denotes the special case $Q_T = P_T$. The law μ_η^Q retains its label to distinguish the augmented reference laws.

A.2 Time grid and Drifts

Fix $0 < \gamma \le 1$ and set

$$h = n^{-1/(2\alpha + d)}, \qquad c_\eta h \le \eta \le C_\eta h, \qquad \tau = h^{2(\alpha+\gamma)/\gamma}, \qquad T = C_T \log n. \tag{A.7}$$

Choose a partition of $[\tau, T]$ that splits at h^2:

$$\tau = t_0 < t_1 < \cdots < t_K = T, \quad 1 < \frac{t_{k+1}}{t_k} \le 2$$

with $K = O(\log n)$. Use a dyadic grid with h^2 and 1 inserted and the endpoints retained. We write the blocks as $I_k = [t_k, t_{k+1}]$. For this fixed model, we denote the score risk on each interval $[t_k, t_{k+1}]$ by

$$\mathcal{E}_k(S) = \int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \|S(x,t) - s_t(x)\|^2 p_t(x) \, dx \, dt. \tag{A.8}$$

We write $a_n \lesssim b_n$ if $a_n \le C b_n$ for some $C > 0$ for sufficiently large n. Similarly, we write $a_n \lesssim_{\log} b_n$ if $a_n \le C(\log n)^c b_n$ for some constants $c, C > 0$ for sufficiently large n.

For sufficiently large n, we have that $\eta \le 1$ and

$$\eta^2 \;\le\; \nu_t^2 = 1 - m_t^2(1 - \eta^2) \;\le\; 1, \qquad \sigma_t \ge c n^{-\kappa}, \qquad \kappa = \frac{\alpha + \gamma}{\gamma(2\alpha + d)}. \tag{A.9}$$

Write $\Phi(L, W, R, B)$ for feedforward ReLU networks with at most L affine layers, maximum hidden width W, at most R nonzero weights and biases, and weights and biases bounded in absolute value by B; the final layer is affine. The input and output dimensions are determined by context. The subclass $\Phi(L, W, R, B, V)$ has every output coordinate globally bounded by V. Clipping is coordinatewise and is exactly realizable by ReLU networks. We reserve Φ for network classes, ϕ_{symbol} for modules, and φ_k for a complete block score network. Scalar–vector multiplication means coordinatewise use of the scalar multiplication module.

A.3 Main Results

Theorem A.4 (Score approximation: Theorem 3.3 of Main text)**.** *Suppose Assumptions A.1, A.2, and the drift condition equation A.4 hold. Set* $L_k = \lceil C(\log n)^4 \rceil$, $\log B_k = C(\log n)^4$, *and* $V_k = C_V\sqrt{\log n}/\sigma_{t_k}$, *with* C, C_V *sufficiently large; in particular* V_k *dominates the score envelope equation B.4. Let* $J_k = 1 + ((t_k \wedge 1)\log n)^{-d/2}$. *Then every interval* $[t_k, t_{k+1}]$ *with* $t_{k+1} \le h^2$ *admits a network* $\varphi_k \in \Phi(L_k, W_k, R_k, B_k, V_k)$ *with* $W_k + R_k \lesssim_{\log} n^{d/(2\alpha+d)}$ *such that*

$$\int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \|\varphi_k(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\, dx\, dt \lesssim_{\log} n^{-2\alpha/(2\alpha+d)}.$$

For $t_k \ge h^2$, *the interval* $[t_k, t_{k+1}]$ *admits a network* $\varphi_k \in \Phi(L_k, W_k, R_k, B_k, V_k)$ *with* $W_k + R_k \lesssim_{\log} J_k$ *such that*

$$\int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \|\varphi_k(x,t) - \nabla \log p_t(x)\|^2 p_t(x)\, dx\, dt \lesssim_{\log} n^{-1}.$$

The bounds are uniform over fields Q_T *satisfying Assumption A.2.*

For laws with finite γ-moments, use the Hölder IPM

$$\begin{aligned} d_\gamma(\mu,\nu) &= \sup_{[g]_\gamma \le 1} \left| \int g\, d\mu - \int g\, d\nu \right| = \inf_{\pi \in \Pi(\mu,\nu)} \int \|x-y\|^\gamma\, d\pi(x,y), \\ [g]_\gamma &= \sup_{x \ne y} \frac{|g(x)-g(y)|}{\|x-y\|^\gamma}. \end{aligned} \tag{A.10}$$

Here $\Pi(\mu,\nu)$ is the set of couplings of μ and ν, and the second equality follows from Kantorovich duality for the metric $\|x-y\|^\gamma$. We also use the following bounded Hölder IPM

$$\bar{d}_\gamma(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int (1 \wedge \|x-y\|^\gamma)\, d\pi(x,y) \le d_{\mathrm{TV}}(\mu,\nu),$$

where $d_{\mathrm{TV}}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$. The cost $1 \wedge \|x-y\|^\gamma$ is itself a metric, so $\bar{d}_\gamma$ obeys the triangle inequality. Fix $R_{\mathrm{out}} > R_0 + 2$ and let Π_{out} be Euclidean projection onto $\overline{B}(0, R_{\mathrm{out}})$. For the blockwise score estimator $\widehat{S}$, defined in Theorem A.5 below and in Section C.2, we use the reverse diffusion

$$dY_t = b_{T-t}\{Y_t + 2\widehat{S}(Y_t, T-t)\}\, dt + \sqrt{2b_{T-t}}\, d\overline{B}_t, \qquad Y_0 \sim \mathrm{N}(0, I_D), \tag{A.11}$$

where $\overline{B}$ is standard Brownian motion in reverse time and $(Y_0, \overline{B})$ is independent of the training sample. We write $\widehat{\mu}_\tau^{\widehat{S}}$ for the conditional law of $Y_{T-\tau}$ given the training sample, and distinguish this raw endpoint from the bounded output

$$\widehat{\mu}_{\tau,\mathrm{out}}^{\widehat{S}} = (\Pi_{\mathrm{out}})_{\#} \widehat{\mu}_\tau^{\widehat{S}}.$$

When the training field must be explicit, we write $\widehat{S}^Q$, or $\widehat{S}^P$ for the exact field.

Choose h, η, τ, T as in equation A.7, with C_T sufficiently large, and use the exact field $Q_T = P_T$. For each block I_k, let $\widehat{S}_k^P$ be a measurable empirical minimizer of equation C.5 over $\Phi_k = \Phi(L_k, W_k, R_k, B_k, V_k)$, using the same n independent observations on every block. Choose the network budgets as in Theorem A.4, sufficiently large to contain the approximation networks constructed in its proof, and set

$$\widehat{S}^P(x,t) = \widehat{S}_k^P(x,t), \qquad t \in I_k,$$

with an arbitrary consistent convention at block endpoints. Let $\widehat{\mu}_\tau^{\widehat{S}^P}$ be the conditional law, given the training sample, of the reverse process equation A.11 at time $T-\tau$. Define $\mathsf{R}_{\mathcal{M}} : \mathbb{R}^D \to \mathcal{M}$ to equal $\Pi_{\mathcal{M}}$ on the $\rho_0/2$-tube and a fixed point of $\mathcal{M}$ outside this tube.

Theorem A.5 (Distribution recovery with known geometry: Statement 1 of Theorem 3.4 in the main text)**.** *Under the assumptions of Theorem A.4, the estimator defined above satisfies, for each fixed* $\gamma \in (0,1]$,

$$\mathbb{E}\Big[d_\gamma\Big((\mathsf{R}_{\mathcal{M}})_{\#} \widehat{\mu}_\tau^{\widehat{S}^P}, p_* \Big) \Big] \lesssim_{\log} n^{-\frac{\alpha+\gamma}{2\alpha+d}} + n^{-1/2} + e^{-c/\eta^2}. \tag{A.12}$$

The expectation is over the training sample and any randomness used to construct the score estimator.

Since $\eta \asymp n^{-1/(2\alpha+d)}$, the exponential term is negligible compared with every polynomial rate.

B PROOF OF THEOREM A.4

We first show some localization properties of the score.

The determinant of the matrix $\Sigma_t(y) = \sigma_t^2 Q_T(y) + \nu_t^2 Q_N(y)$ is $\sigma_t^{2d}\nu_t^{2(D-d)}$ and is independent of y. Let

$$\Lambda_t(y) = \Sigma_t^{-1}(y) = \sigma_t^{-2} Q_T(y) + \nu_t^{-2} Q_N(y)$$

and

$$\Psi_t(x,y) = \frac{\|Q_T(y)(x - m_t y)\|^2}{\sigma_t^2} + \frac{\|Q_N(y)(x - m_t y)\|^2}{\nu_t^2}.$$

By differentiating, the score can be written as

$$s_t(x) = \frac{\nabla p_t}{p_t} = -\frac{\int_{\mathcal{M}} \Lambda_t(y)(x - m_t y) K_t(x,y) f(y)\, d\,\mathrm{vol}(y)}{\int_{\mathcal{M}} K_t(x,y) f(y) d\,\mathrm{vol}(y)}, \tag{B.1}$$

where $K_t(x,y) = \exp(-\Psi_t(x,y)/2)$ is the kernel. Write

$$\begin{aligned} \mathcal{D}_t(x) &= \int_{\mathcal{M}} K_t(x,y) f(y)\, d\,\mathrm{vol}_{\mathcal{M}}(y), \\ \mathcal{N}_t(x) &= \int_{\mathcal{M}} \Lambda_t(y)(x - m_t y) K_t(x,y) f(y)\, d\,\mathrm{vol}_{\mathcal{M}}(y), \\ s_t &= -\mathcal{N}_t/\mathcal{D}_t. \end{aligned} \tag{B.2}$$

Superscripts on truncated or gated integrals below indicate their weights or approximation orders.

Lemma B.1 (Localization)**.** *For every $A > 0$, there exist constants $C_0, C_{\mathrm{den}}, C_1 > 0$, depending only on A and the structural bounds, such that the following holds for all sufficiently large n. Define*

$$\mathcal{T}_t = \{x : \mathrm{dist}(x, m_t\mathcal{M}) \le C_0 \nu_t \sqrt{\log n}\}, \qquad \mathcal{A}_t(x) = \{y \in \mathcal{M} : \Psi_t(x,y) \le C_1 \log n\}.$$

The following estimates hold uniformly over $t \in [\tau, T]$; in (ii)–(iv), assume $x \in \mathcal{T}_t$.

(i)

$$\mathbb{P}(X_t \notin \mathcal{T}_t) \le n^{-2A}, \qquad \mathbb{E}\big[\|s_t(X_t)\|^2 \mathbf{1}_{X_t \notin \mathcal{T}_t}\big] \le C\sigma_t^{-2} n^{-A}. \tag{B.3}$$

(ii) For $x \in \mathcal{T}_t$, $\mathcal{D}_t(x) \ge c\,\sigma_t^d\, n^{-C_{\mathrm{den}}}$.

(iii) Let $\omega : \mathcal{M} \to [0,1]$ be measurable and satisfy $\omega \equiv 1$ on $\mathcal{A}_t(x)$. Multiplying both integrands in equation B.1 by ω changes the score by at most n^{-A}/σ_t and retains at least $\mathcal{D}_t(x)/2$ in the denominator. In particular, for $x \in \mathcal{T}_t$,

$$\|s_t(x)\| \le C\frac{\sqrt{\log n}}{\sigma_t}. \tag{B.4}$$

(iv) If $t \le h^2$, nearest projection $m_t y_x$ of x onto $m_t\mathcal{M}$ is unique and every $y \in \mathcal{A}_t(x)$ satisfies $\|y - y_x\| \lesssim h\sqrt{\log n}$.

By Tang & Yang (2024, Lemma C.2), for $r \le r_{\mathrm{geo}}$ we may choose an r-cover of $\mathcal{M}$ with at most Cr^{-d} centers and bounded overlap of fixed enlargements. We use the normalized-gate construction of their Lemma C.3 and its proof in Section D.4. The geometric changes here are the moving centers $m_t y_j$ and the tube width $C_0\nu_t\sqrt{\log n}$; the score envelope remains on the scale $\sqrt{\log n}/\sigma_t$.

Specifically, let J_{loc} be the number of local gates $\rho_j \in [0,1]$, with $\max_j \rho_j = 1$ on the set $\mathscr{T}_k = \{(x,t) : t \in [t_k, t_{k+1}],\ x \in \mathcal{T}_t\}$. Let $s = s(x,t)$ be the target. If globally bounded local networks S_j satisfy $\|S_j\|_\infty \le V$ and $\|S_j - s\| \le \epsilon$ on $\mathscr{T}_k \cap \mathrm{supp}\,\rho_j$, then

$$S_{\mathrm{loc}} = \frac{\sum_{j=1}^{J_{\mathrm{loc}}} \rho_j S_j}{1 \vee \sum_{j=1}^{J_{\mathrm{loc}}} \rho_j} \qquad \text{and} \qquad \|S_{\mathrm{loc}} - s\|_{\infty, \mathscr{T}_k} \le \epsilon. \tag{B.5}$$

This holds because on $\mathscr{T}_k$, the denominator is strictly $\sum_j \rho_j \ge 1$, making S_{loc} a convex combination of the local networks. (The $1 \vee \cdot$ floor safely caps the denominator outside the tube). Because the

cover has bounded overlap, clipping the approximated gates to zero outside their support ensures that a uniform network approximation error of δ adds at most $O(V\delta)$ to the final output. Thus, the construction achieves the required target accuracy with only a polylogarithmic parameter overhead per chart, safely retaining the depth and width scaling of the constituent local modules.

We use throughout the piecewise affine cutoff

$$\psi(u) = \begin{cases} 1, & u \le 4, \\ (9-u)/5, & 4 < u < 9, \\ 0, & u \ge 9. \end{cases} \tag{B.6}$$

Throughout, we will use the following approximation results. The corresponding exponential, multiplication, and reciprocal statements also appear in Tang & Yang (2024, Lemmas C.5–C.7).

Lemma B.2 (Lemma F.12 in Oko et al. (2023)). *Take $\epsilon > 0$ arbitrarily. There exists a neural network $\phi_{\exp} \in \Phi(L, W, R, B)$ such that*

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

holds, where $L = O(\log^2 \epsilon^{-1})$, $W = O(\log \epsilon^{-1})$, $R = O(\log^2 \epsilon^{-1})$, and $B = \exp(O(\log^2 \epsilon^{-1}))$. Moreover $|\phi_{\exp}(x)| \le \epsilon$ for $x \ge \log(3\epsilon^{-1})$.

Lemma B.3 (Adapted from Lemma F.6 in Oko et al. (2023)). *Let $m \ge 2$, $C \ge 1$, and $0 \le \epsilon_{\text{error}} \le 1$. For any $\epsilon > 0$, there exists a neural network $\phi_{\text{mult}} \in \Phi(L, W, R, B)$ with $L = O\big(\log m\{1 + \log m + \log \epsilon^{-1} + m \log C\}\big)$, $W = 48m$ $R = O\big(m\{1 + \log m + \log \epsilon^{-1} + m \log C\}\big)$, $B = C^m$, such that*

$$\left| \phi_{\text{mult}}(x'_1, ..., x'_m) - \prod_{i=1}^{m} x_i \right| \le \epsilon + mC^{m-1}\epsilon_{\text{error}}$$

for all $x \in [-C, C]^m$ and $x' \in \mathbb{R}^m$ with $\|x - x'\|_\infty \le \epsilon_{\text{error}}$, and $|\phi_{\text{mult}}(x)| \le C^m$ for all $x \in [-C, C]^m$. Repeated inputs give monomial networks $\phi_{\text{mon}}(x; \omega)$ for x^ω; constants and degree-one monomials are affine maps.

For a polynomial $P(x) = \sum_{\iota \in \mathcal{I}_P} a_\iota x^\iota$ with $J_P = |\mathcal{I}_P|$ terms, define

$$\phi_{\text{poly},P}(x) = \sum_{\iota \in \mathcal{I}_P} a_\iota \phi_{\text{mon}}(x; \iota).$$

This is an affine combination of monomial modules; choosing their errors at most $\delta/(1 + \sum_\iota |a_\iota|)$ gives error at most δ for exact inputs. For vector- or matrix-valued polynomials this convention applies entrywise. We also use the same target-based notation for evaluation by a specified arithmetic circuit with shared intermediates (such as the matrix recurrence below). The stability term for perturbed inputs in Lemma B.3 follows directly from the standard telescoping product inequality on $[-C, C]^m$ after coordinate-wise clipping.

Lemma B.4 (Lemma F.7 in Oko et al. (2023)). *For $0 < \epsilon < 1$ there exists $\phi_{\text{rec}} \in \Phi(L, W, R, B)$ with $L \le O(\log^2 \epsilon^{-1})$, $W = O(\log^3 \epsilon^{-1})$, $R = O(\log^4 \epsilon^{-1})$, and $B = O(\epsilon^{-2})$ such that*

$$\left| \phi_{\text{rec}}(x') - \frac{1}{x} \right| \le \epsilon + \frac{|x' - x|}{\epsilon^2}, \text{ for all } x \in [\epsilon, \epsilon^{-1}] \text{ and } x' \in \mathbb{R}.$$

Lemma B.5 (Lemma F.9 in Oko et al. (2023)). *For $0 < \epsilon < 1$, there exists $\phi_{\text{root}} \in \Phi(L, W, R, B)$ with $L \le O(\log^2 \epsilon^{-1})$, $W = O(\log^3 \epsilon^{-1})$, $R = O(\log^4 \epsilon^{-1})$, and $B = O(\epsilon^{-1})$ such that*

$$|\phi_{\text{root}}(x') - \sqrt{x}| \le \epsilon + \frac{|x' - x|}{\sqrt{\epsilon}}, \text{ for all } x \in [\epsilon, \epsilon^{-1}] \text{ and } x' \in \mathbb{R}.$$

For $0 < a \le b$ and $0 < \delta < 1/2$, write $\phi^\delta_{\text{rec};[a,b]}$ and $\phi^\delta_{\text{root};[a,b]}$ for modules preceded by exact clipping to $[a, b]$ and approximating the indicated function there to error δ. Choose the parameter in Lemmas B.4–B.5 at most $\min\{\delta, a, (1 + b)^{-1}, 1/4\}$. Splitting at the clipped input gives

$$\begin{aligned} &|\phi^\delta_{\text{rec};[a,b]}(x') - x^{-1}| \le \delta + a^{-2}|x' - x|, \\ &|\phi^\delta_{\text{root};[a,b]}(x') - \sqrt{x}| \le \delta + (2\sqrt{a})^{-1}|x' - x|, \qquad x \in [a, b]. \end{aligned} \tag{B.7}$$

Their depth, width, sparsity, and log parameter bound are respectively $O(\log^2 \Lambda)$, $O(\log^3 \Lambda)$, $O(\log^4 \Lambda)$, and $O(\log^2 \Lambda)$, where $\Lambda = 2 + \delta^{-1} + a^{-1} + b$. Below, $\phi_{\mathrm{rec};a}$ abbreviates such a module with floor a, a sufficiently large stated or contextually bounded upper endpoint b, and the numerical tolerance prescribed in the relevant error budget. Thus it approximates $x \mapsto 1/(x \vee a)$ throughout the relevant input range. Other module tolerances are likewise suppressed when their error budget is specified.

Lemma B.6. *Under equation A.4, for $0 < \epsilon < 1/2$ there exist networks $\phi_m, \phi_\sigma \in \Phi(L, W, R, B)$ approximating m_t, σ_t uniformly to error ϵ for all $t \geq 0$, where*

$$L = O(\log^2 \epsilon^{-1}), \quad W = O(\log^3 \epsilon^{-1}), \quad R = O(\log^4 \epsilon^{-1}), \quad \log B = O(\log^2 \epsilon^{-1}).$$

Proof. We use the polynomial–exponential construction of Oko et al. (2023, Lemma B.1) and Tang & Yang (2024, Lemma C.4). Condition A.4 provides the necessary schedule regularity for this construction: on intervals of a sufficiently small fixed length, it bounds the degree-m Taylor remainder by $C2^{-m}$. Up to time $C \log \epsilon^{-1}$, there are $O(\log \epsilon^{-1})$ such intervals, and beyond that time we have $m_t \leq \epsilon/4$. We approximate the primitive $\int_0^t b_s ds$ on overlapping intervals, using its exact value at each center as a constant weight, and then compose the degree-$O(\log \epsilon^{-1})$ polynomial and exponential modules. Piecewise linear gates and input clipping yield the stated bounds. For σ_t, we first approximate $1 - m_t^2$ to error $O(\epsilon^4)$, clip the result to $[(\epsilon/8)^2, 1]$, and apply Lemma B.5 with a sufficiently small accuracy tolerance. The clipping operation alters the square root by at most $\epsilon/8$, safely handling the boundary at $t = 0$. Because all required precisions are fixed powers of ϵ, the stated depth and width orders persist. This explicit dependence on quantitative Taylor remainders also highlights why a purely qualitative C^∞ assumption would be insufficient. □

Corollary B.7. *On $[\tau, T]$, for every $0 < \epsilon < 1/2$ there exist $\phi_{\nu^2}, \phi_\nu, \phi_{\sigma^{-1}}, \phi_{\nu^{-2}}$ approximating $\nu_t^2, \nu_t, \sigma_t^{-1}, \nu_t^{-2}$ to uniform error ϵ. They obey the schedule bounds of Lemma B.6 with ϵ replaced by $\epsilon_* = c\epsilon n^{-c_*}$ for a sufficiently large fixed c_*.*

Proof. Set $\phi_{\nu^2} = 1 - (1 - \eta^2)\phi_{\mathrm{mult}}(\phi_m, \phi_m)$. Apply the root and reciprocal modules to ϕ_{ν^2} on $[\eta^2, 1]$ to obtain ϕ_ν and $\phi_{\nu^{-2}}$, respectively. Apply the reciprocal module to ϕ_σ on $[c\sqrt{\tau}, 1]$ to obtain $\phi_{\sigma^{-1}}$. By equation A.9, the stability constants in equation B.7 are bounded by fixed powers of n, independently of the chosen numerical tolerance. Approximate the input functions and each arithmetic operation to error at most $c\epsilon n^{-c_*}$, with c_* large enough for this fixed number of compositions. The propagated error is at most ϵ. The size bounds follow by composition and parallelization (Oko et al., 2023, Lemmas F.1–F.3). □

Lemma B.8. *Let $C \geq 1$ bound every input coordinate in absolute value, let $0 < \epsilon < 1$, and let the inputs be available to accuracy $\epsilon_{\mathrm{error}}$. Then:*

(i) For $u, v \in \mathbb{R}^D$, the network $\phi_{\mathrm{dot}}(u, v) = \sum_{b=1}^D \phi_{\mathrm{mult}}(u_b, v_b)$ lies in $\Phi(L, W, R, B)$ with $L = O(\log \epsilon^{-1} + \log C) + 1, W = O(D), R = O\big(D(\log \epsilon^{-1} + \log C)\big), B \leq C^2$ and

$$|\phi_{\mathrm{dot}}(u', v') - \langle u, v \rangle| \leq D(\epsilon + 2C\epsilon_{\mathrm{error}}), \qquad |\phi_{\mathrm{dot}}| \leq DC^2.$$

(ii) For $A_{\mathrm{mat}} \in \mathbb{R}^{p \times q}$, $B_{\mathrm{mat}} \in \mathbb{R}^{q \times r}$, the network $[\phi_{\mathrm{mm}}(A_{\mathrm{mat}}, B_{\mathrm{mat}})]_{ij} = \sum_{b=1}^q \phi_{\mathrm{mult}}([A_{\mathrm{mat}}]_{ib}, [B_{\mathrm{mat}}]_{bj})$ lies in $\Phi(L, W, R, B)$ with the same depth and weight bound, $W = O(pqr)$, $R = O\big(pqr(\log \epsilon^{-1} + \log C)\big)$, and entrywise error $q(\epsilon + 2C\epsilon_{\mathrm{error}})$, entrywise bound qC^2.

Proof. Both networks are a single layer of parallel copies of ϕ_{mult} with $m = 2$ followed by one affine summation layer. Lemma B.3 with $m = 2$ gives each copy depth $O(\log \epsilon^{-1} + \log C)$, width $O(1)$, sparsity $O(\log \epsilon^{-1} + \log C)$, parameter bound C^2, and error $\epsilon + 2C\epsilon_{\mathrm{error}}$; parallelizing D copies (resp. pqr copies) and appending the summation layer preserves the depth and parameter bound, adds one to the depth, and multiplies the width and sparsity by the number of copies (Oko et al., 2023, Lemmas F.1–F.3). Summing D (resp. q) such terms multiplies the error and the output bound by that many, giving the stated bounds. □

B.1 HIGHER TIME REGIME: $h^2 \le t_k < t_{k+1} \le T$

In this regime we have $\nu_t \asymp \sigma_t \asymp (t \wedge 1)^{1/2}$. We use the covering, factorization, Taylor expansion, and network assembly of Tang & Yang (2024, Section C.1), with a polylogarithmic refinement of the local cover. This makes the varying exponent small while preserving the complexity.

Put $\ell_n = \log n$ and choose a coarse partition $\{\chi_j\}_{j=1}^{J_{\mathrm{c},k}}$ at radius $r_k = c\{(t_k \wedge 1)\ell_n\}^{1/2} \wedge r_{\mathrm{geo}}$, with $J_{\mathrm{c},k} = O(J_k)$ and $\operatorname{supp}\chi_j \subset B(y_j, Cr_k)$. Using the piecewise affine function ψ in equation B.6, we construct the local gates

$$r_k^{\mathrm{gate}} = C_g \nu_{t_k}\sqrt{\ell_n}, \qquad \rho_j(x,t) = \psi\left(\frac{\|x - m_t y_j\|^2}{(r_k^{\mathrm{gate}})^2}\right).$$

Choose C_g sufficiently large so ρ_j equals one whenever $\operatorname{supp}\chi_j$ meets $\mathcal{A}_t(x)$. By construction, its support satisfies $\|x - m_t y_j\| \le C\nu_t\sqrt{\ell_n}$, with fixed containment margins. For each j, refine the partition on $\operatorname{supp}\chi_j$ using centers y_{ja} and nonnegative functions ψ_{ja}, indexed by $1 \le a \le J_{\mathrm{f},j,k}$, satisfying

$$r_{\mathrm{f},k} = \frac{c_{\mathrm{f}}\sigma_{t_k}}{\ell_n^2}, \qquad \sum_a \psi_{ja} = 1 \text{ on } \operatorname{supp}\chi_j, \qquad \operatorname{supp}(\chi_j\psi_{ja}) \subset B(y_{ja}, Cr_{\mathrm{f},k}). \tag{B.8}$$

The covering bound gives at most $C(1 + r_k/r_{\mathrm{f},k})^d \le C\ell_n^{5d/2}$ fine cells per coarse cell. Thus the total number of pairs (j,a) is $\lesssim_{\log} J_k$. We define the fine-cell weights and local coordinates:

$$\chi_{ja} = \chi_j\psi_{ja}, \quad M_{ja} = \int_{\mathcal{M}} \chi_{ja} f \, d\,\mathrm{vol}, \quad \sum_{j,a} M_{ja} = 1, \quad r_{ja} = x - m_t y_{ja}, \quad q_{ja} = \tfrac{1}{2} r_{ja}^\top \Lambda_t(y_{ja}) r_{ja}.$$

On an active coarse gate and within the fine support, $\|r_{ja}\| \le C\sigma_t\sqrt{\ell_n}$. Because the field is uniformly Lipschitz in local coordinates, we bound the variation for $v = y - y_{ja}$ by $\|\Lambda_t(y) - \Lambda_t(y_{ja})\| \le C\sigma_t^{-2}\|v\|$. We can hence expand the exponent $\Delta_{ja} := \frac{1}{2}\Psi_t(x,y) - q_{ja}$:

$$\begin{aligned}\Delta_{ja} &= -m_t r_{ja}^\top \Lambda_t(y_{ja})v + \tfrac{1}{2}m_t^2 v^\top \Lambda_t(y_{ja})v \\ &\quad + \tfrac{1}{2}(r_{ja} - m_t v)^\top\{\Lambda_t(y) - \Lambda_t(y_{ja})\}(r_{ja} - m_t v),\end{aligned} \tag{B.9}$$

$$0 \le q_{ja} \le C\ell_n, \qquad |\Delta_{ja}| \le C\left\{\frac{r_{\mathrm{f},k}\sqrt{\ell_n}}{\sigma_t} + \frac{r_{\mathrm{f},k}^2}{\sigma_t^2} + r_{\mathrm{f},k}\ell_n\right\} \le \ell_n^{-1}.$$

The last bound holds for a sufficiently small structural c_{f}, using $\sigma_t \asymp \sigma_{t_k} \le 1$ on the block. The last term in equation B.9 accounts for the varying anisotropic projector.

For an accuracy exponent $A_{\mathrm{num}} > 0$, let $m_{\mathrm{exp}} = \lceil (A_{\mathrm{num}} + 2)\ell_n / \log\ell_n \rceil$ and $P_{m_{\mathrm{exp}}}(u) = \sum_{j=0}^{m_{\mathrm{exp}}} (-u)^j/j!$ be the truncated Taylor series. Then

$$K_t(x,y) = e^{-q_{ja}}e^{-\Delta_{ja}}, \qquad \sup_{|u| \le \ell_n^{-1}} |e^{-u} - P_{m_{\mathrm{exp}}}(u)| \le \frac{e^{1/\ell_n}\ell_n^{-(m_{\mathrm{exp}}+1)}}{(m_{\mathrm{exp}}+1)!} \le n^{-A_{\mathrm{num}}}. \tag{B.10}$$

Let

$$\begin{aligned}\mathcal{D}_{ja} &= \int_{\mathcal{M}} \chi_{ja}(y) K_t(x,y) f(y) \, d\,\mathrm{vol}(y) \\ \mathcal{N}_{ja} &= \int_{\mathcal{M}} \chi_{ja}(y)\Lambda_t(y)(x - m_t y) K_t(x,y) f(y) \, d\,\mathrm{vol}(y).\end{aligned}$$

Replacing K_t by the polynomial approximation $K_{ja}^{(m_{\mathrm{exp}})} = e^{-q_{ja}} P_{m_{\mathrm{exp}}}(\Delta_{ja})$ yields the approximate integrals $\mathcal{D}_{ja}^{(m_{\mathrm{exp}})}, \mathcal{N}_{ja}^{(m_{\mathrm{exp}})}$. Since $e^{-q_{ja}} \le 1$ and $\|\Lambda_t(y)(x - m_t y)\| \le \sigma_t^{-1}\sqrt{\Psi_t(x,y)}$, we have

$$|\mathcal{D}_{ja}^{(m_{\mathrm{exp}})} - \mathcal{D}_{ja}| \le n^{-A_{\mathrm{num}}} M_{ja}, \qquad \|\mathcal{N}_{ja}^{(m_{\mathrm{exp}})} - \mathcal{N}_{ja}\| \le C\sigma_t^{-1}\sqrt{\ell_n}\, n^{-A_{\mathrm{num}}} M_{ja}. \tag{B.11}$$

Summing with $0 \le \rho_j \le 1$ preserves these bounds because $\sum_{j,a} M_{ja} = 1$. Write $\mathcal{D}^\rho = \sum_{j,a}\rho_j\mathcal{D}_{ja}$ and $\mathcal{N}^\rho = \sum_{j,a}\rho_j\mathcal{N}_{ja}$ for the exact gated integrals. The superscript $\rho, (m_{\mathrm{exp}})$ denotes the same sums with each cell integral replaced by its Taylor approximation.

Set

$$\Upsilon(y) = \big(y, \{[Q_T(y)]_{bc}\}_{b\le c}\big) \in \mathbb{R}^{m_D}, \qquad m_D = D + \tfrac{D(D+1)}{2}, \qquad \theta = (x, m_t, \sigma_t^{-2}, \nu_t^{-2}).$$

For formal variables (y, Q_{mat}), with Q_{mat} a symmetric $D \times D$ matrix, in place of $\Upsilon(y)$, define the polynomials

$$\begin{aligned} \mathscr{V}(y, Q_{\mathrm{mat}}; \theta) &= \nu_t^{-2}(x - m_t y) + (\sigma_t^{-2} - \nu_t^{-2}) Q_{\mathrm{mat}}(x - m_t y), \\ \mathscr{E}_{ja}(y, Q_{\mathrm{mat}}; \theta) &= \tfrac{1}{2}(x - m_t y)^\top \mathscr{V}(y, Q_{\mathrm{mat}}; \theta) - q_{ja}. \end{aligned} \tag{B.12}$$

They have degrees at most 2 and 3 in (y, Q_{mat}). Expand

$$P_{m_{\exp}}(\mathscr{E}_{ja}) = \sum_{|\iota| \le 3m_{\exp}} c^{\mathcal{D}}_{ja,\iota}(\theta) \Upsilon^\iota, \qquad \mathscr{V} P_{m_{\exp}}(\mathscr{E}_{ja}) = \sum_{|\iota| \le 3m_{\exp}+2} c^{\mathcal{N}}_{ja,\iota}(\theta) \Upsilon^\iota.$$

The coefficients are explicit polynomials in θ, with $y_{ja}, Q_T(y_{ja})$ fixed. With the time-independent moments

$$\mathfrak{m}_{ja}[\iota] = \int_{\mathcal{M}} \chi_{ja}(y) f(y) \Upsilon(y)^\iota \, d\,\mathrm{vol}(y), \qquad \mathcal{D}_{ja}^{(m_{\exp})} = e^{-q_{ja}} \sum_\iota c^{\mathcal{D}}_{ja,\iota}(\theta) \mathfrak{m}_{ja}[\iota], \tag{B.13}$$

the analogous expression with $c^{\mathcal{N}}$ gives $\mathcal{N}_{ja}^{(m_{\exp})}$. These moments are fixed weights of the population approximator; all dependence on (x, t) is displayed in the coefficient polynomials and q_{ja}. In particular, the full time-dependent kernel was expanded. The number of moments per fine cell obeys

$$J_{\mathrm{mom}} \le \binom{m_D + 3m_{\exp} + 2}{m_D} = O_D((\log n)^{m_D}) = (\log n)^{O(D^2)}. \tag{B.14}$$

All sums and multiplications by fixed coefficients below are affine layers. Put

$$\begin{aligned} \phi_{\sigma^{-2}} &= \phi_{\mathrm{mult}}(\phi_{\sigma^{-1}}, \phi_{\sigma^{-1}}), & \phi_\theta &= (x, \phi_m, \phi_{\sigma^{-2}}, \phi_{\nu^{-2}}), \\ \phi_{r,ja} &= x - \phi_m(t) y_{ja}, & \phi_{\Lambda,ja} &= \phi_{\sigma^{-2}}(t) Q_T(y_{ja}) + \phi_{\nu^{-2}}(t) Q_N(y_{ja}), \\ \phi_{q,ja} &= \tfrac{1}{2}\phi_{\mathrm{dot}}\big(\phi_{r,ja}, \phi_{\mathrm{mm}}(\phi_{\Lambda,ja}, \phi_{r,ja})\big), & \phi_{e,ja} &= \phi_{\exp}\big(\mathrm{Clip}_{[0,C\ell_n]}\, \phi_{q,ja}\big). \end{aligned} \tag{B.15}$$

With $\phi_{r,j} = x - \phi_m(t) y_j$, the gate network is explicitly $\phi_{\rho_j} = \psi((r_k^{\mathrm{gate}})^{-2} \phi_{\mathrm{dot}}(\phi_{r,j}, \phi_{r,j}))$. For $c^{\mathcal{D}}_{ja,\iota}(\theta) = \sum_\zeta a^{\mathcal{D}}_{ja,\iota,\zeta} \theta^\zeta$, define

$$\phi_{c^{\mathcal{D}},ja,\iota} = \sum_\zeta a^{\mathcal{D}}_{ja,\iota,\zeta} \phi_{\mathrm{mon}}(\phi_\theta; \zeta), \qquad \phi_{\mathcal{D},ja} = \phi_{\mathrm{mult}}\left(\phi_{e,ja}, \sum_\iota \mathfrak{m}_{ja}[\iota] \phi_{c^{\mathcal{D}},ja,\iota}\right).$$

Define $\phi_{c^{\mathcal{N}},ja,\iota}$ and the vector $\phi_{\mathcal{N},ja}$ by the same formulas with $c^{\mathcal{N}}$ in place of $c^{\mathcal{D}}$. Approximating the coarse distance gates by ϕ_{ρ_j} gives

$$\begin{aligned} \phi_{\mathcal{D}} &= \sum_{j,a} \phi_{\mathrm{mult}}(\phi_{\rho_j}, \phi_{\mathcal{D},ja}), & \phi_{\mathcal{N}} &= \sum_{j,a} \phi_{\mathrm{mult}}(\phi_{\rho_j}, \phi_{\mathcal{N},ja}), \\ \varphi_k &= \mathrm{Clip}_{[-V_k,V_k]^D}\left[-\phi_{\mathrm{mult}}\big(\phi_{\mathcal{N}}, \phi_{\mathrm{rec};d_{0,k}/2}(\phi_{\mathcal{D}})\big)\right]. \end{aligned} \tag{B.16}$$

Here $d_{0,k} = c\sigma_{t_k}^d n^{-C_{\mathrm{den}}}$ is a lower bound for the exact gated denominator; its constant is fixed sufficiently small.

All higher-time network inputs are first clipped exactly to $\mathcal{B}_k = [-C_x\sqrt{\ell_n}, C_x\sqrt{\ell_n}]^D \times I_k$, where C_x is chosen so that this clipping is inactive on $\mathscr{T}_k$. Because the coefficient polynomials have degree $O(m_{\exp})$ with $m_{\exp} = O(\ell_n)$, and each cell contains a polylogarithmic number of monomials, the intermediate arithmetic values on this box are bounded by $M_n = \exp\{C\ell_n^2\}$. Expanding each coefficient into monomials creates at most $O(m_{\exp})$ polynomial-sized product factors along any computational path. Consequently, the propagated numerical error is at most $\exp\{C\ell_n^2\}$ times the primitive operation accuracy.

To control gate leakage, let F_{ja} denote a scalar component of $\mathcal{D}_{ja}^{(m_{\exp})}$ or $\mathcal{N}_{ja}^{(m_{\exp})}$, and let $\widehat{F}_{ja}$ denote its network realization. The preceding range bound also gives $|\widehat{F}_{ja}| \le M_n$ on $\mathcal{B}_k$, after

increasing C. For primitive and schedule accuracy δ_n, put $\epsilon_n = \exp\{C\ell_n^2\}\delta_n$. Uniformly on this box, $|\phi_{\rho_j} - \rho_j| \le \epsilon_n$, while $|\widehat{F}_{ja} - F_{ja}| \le \epsilon_n$ wherever $\rho_j > 0$. Since $0 \le \rho_j \le 1$, these bounds imply

$$|\phi_{\rho_j}\widehat{F}_{ja} - \rho_j F_{ja}| \le \epsilon_n(M_n + 1).$$

On $\{\rho_j = 0\}$ this uses only the gate error and the bound on $\widehat{F}_{ja}$; it does not require a local approximation there. There are $\lesssim_{\log} J_k$ cells, so summing this estimate and the errors of the final multiplication modules, with $\delta_n = \exp\{-C_{\mathrm{num}}\ell_n^2\}$ and C_{num} sufficiently large, ensures that the discrete network outputs satisfy:

$$|\phi_{\mathcal{D}} - \mathcal{D}^{\rho,(m_{\exp})}| \le n^{-A_{\mathrm{num}}}, \qquad \|\phi_{\mathcal{N}} - \mathcal{N}^{\rho,(m_{\exp})}\| \le n^{-A_{\mathrm{num}}}.$$

For $x \in \mathcal{T}_t$, the effective integral weight $y \mapsto \sum_j \rho_j(x,t)\chi_j(y)$ takes values in $[0,1]$ and equals one on $\mathcal{A}_t(x)$. Applying Lemma B.1(ii)–(iii) guarantees a lower bound on the exact denominator $\mathcal{D}^\rho \ge d_{0,k}$, a truncation error of n^{-A}/σ_t, and a bounded ratio $\|\mathcal{N}^\rho/\mathcal{D}^\rho\| \le C\sqrt{\ell_n}/\sigma_t$. To safely compute this ratio in the network, we floor the learned denominator at $d_{0,k}/2$. Given empirical numerator and denominator errors $\epsilon_{\mathcal{N}}$ and $\epsilon_{\mathcal{D}}$, the flooring condition $|\widehat{\mathcal{D}} \vee (d_{0,k}/2) - \mathcal{D}^\rho| \le \epsilon_{\mathcal{D}}$ directly yields the algebraic stability bound:

$$\left\| \frac{\widehat{\mathcal{N}}}{\widehat{\mathcal{D}} \vee (d_{0,k}/2)} - \frac{\mathcal{N}^\rho}{\mathcal{D}^\rho} \right\| \le \frac{2}{d_{0,k}} \left\{ \epsilon_{\mathcal{N}} + \left\| \frac{\mathcal{N}^\rho}{\mathcal{D}^\rho} \right\| \epsilon_{\mathcal{D}} \right\}. \tag{B.17}$$

Combining the continuous Taylor approximation errors from equation B.11 with these discrete numerical bounds produces the total network error:

$$\sigma_t \|\varphi_k - s_t\| \le C\sqrt{\ell_n}\, n^{C_{\mathrm{den}} + d\kappa - A_{\mathrm{num}}} + Cn^{-A}. \tag{B.18}$$

Because the ranges of the final reciprocal and multiplication operations in equation B.16 are bounded by polynomials in n, their resultant errors are fully absorbed by the primitive accuracy δ_n. Setting $A \ge 1$ and $A_{\mathrm{num}} > C_{\mathrm{den}} + d\kappa + 1$ simplifies the right-hand side, yielding the final uniform bound $\|\varphi_k - s_t\| \lesssim_{\log} n^{-1/2}/\sigma_t$ on $\mathcal{T}_t$.

The network evaluates $\lesssim_{\log} J_k$ fine cells, each requiring the computation of polylogarithmically many moments and coefficient monomials. Because the computed products contain at most $O(m_{\exp})$ factors, evaluating them sequentially—with a depth of $O(\ell_n^2)$ per multiplication—yields a total multiplication depth of $O(\ell_n^3)$. The schedule modules require a depth and log-parameter bound of $O(\ell_n^4)$ to achieve the rigorous target accuracy δ_n. Since only a strictly bounded number of these structural stages occur sequentially, parallelizing the evaluations over all cells and moments achieves the desired complexity bounds $W_k + R_k \lesssim_{\log} J_k$. With the pointwise bounds and network architecture established, we bound the integrated score matching loss over the time interval. Because the time integral of σ_t^{-2} is $O(1)$ for blocks below $t = 1$ and at most $O(\log n)$ for terminal blocks, integrating our pointwise bound on the active tubular neighborhood yields:

$$\mathcal{E}_k(\varphi_k) \lesssim_{\log} n^{-1} \int_{t_k}^{t_{k+1}} \sigma_t^{-2}\, dt \lesssim_{\log} n^{-1}.$$

Finally, the trivial network clipping bound combined with the marginal mass bounds from Lemma B.1(i) absorb the error on the complement of the active tubular neighborhood $\mathcal{T}_t$, completing the proof. □

Remark B.9. *The same construction with $r_{\mathrm{f},k} = c_{\mathrm{f}}\sigma_{t_k} n^{-\zeta}/\log n$, for fixed $\zeta > 0$, gives $|\Delta_{ja}| \le n^{-\zeta}$ and requires only $m_{\exp} = \lceil (A_{\mathrm{num}} + 2)/\zeta \rceil$ Taylor terms. Its fine-cell count is $\lesssim_{\log} J_k n^{\zeta d}$, so this direct construction gives $W_k + R_k \lesssim_{\log} J_k n^{\zeta d}$ rather than the higher-time case bound. Using that class on every higher-time block would add $n^{\zeta d}$ to the complexity term in the present oracle inequality. Our logarithmic refinement retains the same small-exponent argument and the stated statistical rate: it corresponds to $\zeta_n = \log\log n/\log n$, so $n^{-\zeta_n} = 1/\log n$. Tang & Yang (2024) use the particular radius $n^{-2\zeta}(\log n)^{-2}$, which gives the small-exponent bound on their large-time range $t \gtrsim n^{-2\zeta}(\log n)^{-3}$; it is not a uniform replacement down to h^2 without relating ζ to that cutoff.*

B.2 SMALL TIME REGIME: $\tau \le t_k < t_{k+1} \le h^2$

Let $L_n = C_L\sqrt{\log n}$ with $C_L \ge \max\{1, C_0, \sqrt{C_1}\}$ from Lemma B.1 and set $r_h = C_h h L_n$, where C_h is chosen so that $C_0\nu_t\sqrt{\log n} \le r_h$ for $t \le h^2$. Let $\{y_j\}$ be an r_h-cover of $\mathcal{M}$ and set

$$\rho_j(x,t) = \psi\Big(\frac{\|x - m_t y_j\|^2}{r_h^2}\Big),$$

with ψ as in equation B.6. If a point y_j is within r_h of the nearest point y_x, and $x \in \mathcal{T}_t$, then $\|x - m_t y_x\| \le r_h$. By the triangle inequality and bound $m_t \le 1$, we have $\|x - m_t y_j\| \le \|x - m_t y_x\| + m_t\|y_x - y_j\| \le r_h + m_t r_h \le 2r_h$. Squaring it gives $\le 4r_h^2$, meaning the evaluation point falls exactly in the $u \le 4$ case of ψ, ensuring $\rho_j(x,t)$ is exactly 1. Furthermore, whenever $\rho_j(x,t) > 0$, Lemma B.1(iv) ensures that all active integration variables $y \in \mathcal{A}_t(x)$ lie within a distance Cr_h of y_j. Consequently, a sufficiently enlarged chart domain centered at y_j safely contains all such active points and the inverse-coordinate domains required below. Let G^* be the exact parametrization of the chart, and let G be its Taylor polynomial of degree $\lceil\beta\rceil - 1$. Also let v be the Taylor polynomial of degree $\lceil\alpha\rceil - 1$ of the exact volume density

$$v^*(z) = f(G^*(z))\sqrt{\det(\nabla G^*(z)^\top \nabla G^*(z))}.$$

By Assumption A.2, there exists an orthonormal frame U^* such that $Q_T(G^*(z)) = U^*(z)U^*(z)^\top$. Let H be its Taylor polynomial of degree $\lceil\beta - 1\rceil - 1$, and $U = H(H^\top H)^{-1/2}$. In the extended chart domain, Taylor approximation and the uniform positive definiteness of the Gram matrix $H^\top H$ gives for sufficiently large n:

$$\begin{aligned} \|G - G^*\|_\infty &\lesssim_{\log} h^\beta, & \|\nabla G - \nabla G^*\|_\infty + \|U - U^*\|_\infty &\lesssim_{\log} h^{\beta-1}, \\ \|\nabla U - \nabla U^*\|_\infty &\lesssim_{\log} h^{\beta-2}, & \|v - v^*\|_\infty &\lesssim_{\log} h^\alpha. \end{aligned} \tag{B.19}$$

All Taylor polynomials are centered at the chart origin. Their degrees are fixed and their coefficients are uniformly bounded. In particular, $\|\nabla v\| + \|\nabla^2 G\| + \|\nabla^2 U\| \le C$ on a fixed smaller neighborhood; when $\alpha \le 1$, v is constant. These are bounds for the polynomial and orthonormalized surrogates, not extra derivatives of f or U^*.

Lemma B.10. *Fix a chart j and a small-time pair (x,t) with $\rho_j(x,t) > 0$. On the corresponding extended chart, define*

$$F^*_{x,t}(z) = U^*(z)^\top(x - m_t G^*(z)), \qquad F_{x,t}(z) = U(z)^\top(x - m_t G(z)).$$

These maps admit inverse branches $\mathsf{Z}^*_{x,t}$ and $\mathsf{Z}_{x,t}$ on $\|q\|_\infty \le 2\sigma_t L_n$, taking values in the enlarged chart. Suppressing (x,t) below, derivatives of F^*, F are taken with respect to z, and derivatives of Z^*, Z with respect to q. Uniformly over all such (x,t,q),

$$\|\nabla\mathsf{Z}(q)\| + \|\nabla\mathsf{Z}^*(q)\| \le C, \tag{B.20}$$

$$\|\mathsf{Z}(q) - \mathsf{Z}^*(q)\| \lesssim_{\log} h^\beta, \qquad \|\nabla\mathsf{Z}(q) - \nabla\mathsf{Z}^*(q)\| \lesssim_{\log} h^{\beta-1}. \tag{B.21}$$

Define

$$r^*_\perp(q) = \big[I_D - U^*(\mathsf{Z}^*(q))U^*(\mathsf{Z}^*(q))^\top\big]\big[x - m_t G^*(\mathsf{Z}^*(q))\big],$$

and define $r_\perp$ analogously using U, G, Z. Then, uniformly on the same domain,

$$\|r_\perp(q)\| + \|r^*_\perp(q)\| \le C\nu_t L_n, \qquad \|r_\perp(q) - r^*_\perp(q)\| \lesssim_{\log} h^\beta, \tag{B.22}$$

and

$$\|U(\mathsf{Z}(q)) - U^*(\mathsf{Z}^*(q))\| \lesssim_{\log} h^{\beta-1}.$$

Remark B.11 (Residuals for a transverse field)**.** *For a general field, $Q_N\nabla G^* = 0$ is false: differentiating the exact complementary residual produces the extra term $-m_t Q_N \nabla G^* \nabla\mathsf{Z}^*$. Lemma B.10 uses only residual sizes and comparisons at matched coordinates, never this cancellation, and therefore applies uniformly to $Q_T = P_T$ and to the transverse model.*

Then we conduct a change variables in the exact and approximate integrals in the expression for the score. We first change to $q = F(z)$ and then to $u = q/\sigma_t$ on $\mathcal{U}_n = [-L_n, L_n]^d$, and set

$$\begin{aligned} a^*(q) &= v^*(\mathsf{Z}^*(q))|\det\nabla\mathsf{Z}^*(q)|, & a(q) &= v(\mathsf{Z}(q))|\det\nabla\mathsf{Z}(q)|, \\ \vartheta^*(q) &= a^*(q)e^{-\|r^*_\perp(q)\|^2/(2\nu_t^2)}, & \vartheta(q) &= a(q)e^{-\|r_\perp(q)\|^2/(2\nu_t^2)}. \end{aligned}$$

Let μ^* and μ have densities on $\mathcal{U}_n$ proportional to $e^{-\|u\|^2/2}\vartheta^*(\sigma_t u)$ and $e^{-\|u\|^2/2}\vartheta(\sigma_t u)$, respectively. The domain of integration is

$$\Omega_t(x) = \{G^*(\mathsf{Z}^*(\sigma_t u)) : u \in \mathcal{U}_n\}.$$

Note that $\mathcal{A}_t(x) \subseteq \Omega_t(x)$ (where $\mathcal{A}_t(x)$ is defined as in Lemma B.1) since $y \in \mathcal{A}_t(x)$ implies $\|Q_T(y)(x - m_t y)\| \le \sigma_t\sqrt{C_1 \log n}$ and $C_L \ge \sqrt{C_1}$. Applying Lemma B.1(iii) with $\omega = \mathbf{1}_{\Omega_t(x)}$ hence controls the replacement of s_t by the truncated ratio on this cube. To bound the residual over the complement $\Omega_t(x) \setminus \mathcal{A}_t(x)$, observe that at any $y = G^*(\mathsf{Z}^*(\sigma_t u))$,

$$\tfrac{1}{2}\Psi_t(x,y) = \tfrac{1}{2}\|u\|^2 + \frac{\|r_\perp^*(\sigma_t u)\|^2}{2\nu_t^2}, \qquad e^{-\|u\|^2/2}\vartheta^*(\sigma_t u) = a^*(\sigma_t u)\, e^{-\Psi_t(x,y)/2}. \tag{B.23}$$

By definition, any point in $\Omega_t(x) \setminus \mathcal{A}_t(x)$ has energy $\Psi_t > C_1 \log n$, bounding the kernel $e^{-\Psi_t/2}$ by $n^{-C_1/2}$. Since the volume of $\mathcal{U}_n$ is $O(L_n^d) = O((\log n)^{d/2})$, the unnormalized residual mass is at most $O\big((\log n)^{d/2} n^{-C_1/2}\big)$. Normalizing by the rescaled denominator (bounded below by $cn^{-C_{\mathrm{den}}}$) yields an excess mass contribution of at most $O\big((\log n)^{d/2} n^{C_{\mathrm{den}} - C_1/2}\big)$. An identical argument holds for the numerator, as the additional spatial factor $\Lambda_t(y)(x - m_t y)$ introduces only a negligible polynomial multiplier $O(\sqrt{\log n}/\sigma_t)$, giving the truncation bound.

Lemma B.12. *Uniformly on the active tubular neighborhood,*

$$\sup_q \Big|\frac{\vartheta(q)}{\vartheta^*(q)} - 1\Big| \lesssim_{\log} h^\alpha, \qquad \|\mu - \mu^*\|_{\mathrm{TV}} \lesssim_{\log} h^\alpha. \tag{B.24}$$

Moreover the chart-approximated score

$$\widetilde{s} = -\sigma_t^{-1}\mathbb{E}_\mu\big[U(\mathsf{Z}(\sigma_t u))\, u\big] - \nu_t^{-2}\mathbb{E}_\mu\big[r_\perp(\sigma_t u)\big] \tag{B.25}$$

satisfies, for every fixed A and the corresponding localization constants,

$$\|\widetilde{s} - s_t\| \lesssim_{\log} \frac{h^\alpha}{\sigma_t} + \frac{n^{-A}}{\sigma_t}. \tag{B.26}$$

We now realize the chart-approximated score by a ReLU network.

Lemma B.13. *Fix $A > 0$ and an interval $[t_k, t_{k+1}]$ with $t_{k+1} \le h^2$, and let $\widetilde{s}$ be the chart-approximated score equation B.25 on a valid chart gate. There exists a neural network $\phi_{\mathrm{loc}}(x,t) \in \Phi(L, W, R, B, V_k)$ such that*

$$\sup_{t\in[t_k,t_{k+1}]}\ \sup_{x\,:\,\rho_j(x,t)>0,\ x\in\mathcal{T}_t} \big\|\phi_{\mathrm{loc}}(x,t) - \mathrm{Clip}_{[-V_k,V_k]^D}\,\widetilde{s}(x,t)\big\| \le \frac{n^{-A}}{\sigma_{t_k}},$$

where $L = O\big((\log n)^4\big), W = O\big((\log n)^{d+3}\big), R = O\big((\log n)^{d+4}\big), B = \exp\big(O\big((\log n)^4\big)\big), V_k = C_V\frac{\sqrt{\log n}}{\sigma_{t_k}}$. The implicit constants depend only on A and the structural bounds, and in particular are uniform over charts and over blocks with $t_{k+1} \le h^2$.

Fix $A \ge 3$. For each chart, let $\phi_{\mathrm{loc},j}$ be the local network of Lemma B.13, clipped at V_k. Define $\phi_{r,j} = x - \phi_m(t)y_j$ and $\phi_{\rho_j} = \psi(r_h^{-2}\phi_{\mathrm{dot}}(\phi_{r,j}, \phi_{r,j}))$. The block network is explicitly

$$\varphi_k = \mathrm{Clip}_{[-V_k,V_k]^D}\left[\phi_{\mathrm{mult}}\left(\sum_j \phi_{\mathrm{mult}}(\phi_{\rho_j}, \phi_{\mathrm{loc},j}),\ \phi_{\mathrm{rec};1}\Big(\sum_j \phi_{\rho_j}\Big)\right)\right]. \tag{B.27}$$

Choose gate and arithmetic errors small enough to absorb the number of charts. The normalizing integrals in equation B.25 lie between n^{-C} and C, since $a, a^* \asymp 1$ and $\|r_\perp\|, \|r_\perp^*\| \le C\nu_t L_n$, so Lemma B.13, equation B.26, and the gluing bound before combine to give

$$\sup_{x\in\mathcal{T}_t}\|\varphi_k(x,t) - s_t(x)\| \lesssim_{\log} \frac{h^\alpha + n^{-A}}{\sigma_t}, \qquad t \in I_k. \tag{B.28}$$

Because adjacent grid points have ratio at most two and $\sigma_t^2 \asymp t$, for $t_{k+1} \le h^2$ we have

$$\int_{t_k}^{t_{k+1}} \sigma_t^{-2}\, dt \le C\log(t_{k+1}/t_k) \le C\log 2.$$

Squaring equation B.28 and integrating therefore contributes $\lesssim_{\log} h^{2\alpha} + n^{-2A} \lesssim_{\log} h^{2\alpha}$ to the expected loss, where the n^{-2A} term is strictly dominated for $A \geq 3$. On $\mathcal{T}_t$, the clipping operator at V_k is nonexpansive toward s_t by the score envelope bound equation B.4, leaving the local approximation intact. Off the tube, the network output is deterministically bounded by $\|\varphi_k\| \leq \sqrt{D}V_k$. Combined with the marginal tail mass bounds from Lemma B.1(i), the complement space contributes at most an order of $V_k^2 n^{-2A} + \sigma_{t_k}^{-2} n^{-A}$ to the expected loss, which is trivially absorbed into the n^{-A} remainder by choosing A sufficiently large.

Finally, we bound the architectural complexity. By the covering bound at radius r_h, the global network evaluates $\lesssim_{\log} h^{-d}$ charts. By Lemma B.13, each local network possesses strictly polylogarithmic width and sparsity. Because the gating and final normalization layers add only a further polylogarithmic factor, the total complexity of the block network scales as $W_k + R_k \lesssim_{\log} h^{-d}$, completing the proof.

C Proof of Theorem A.5

C.1 Score to Distribution Error

Throughout this section we condition on the training sample, so that the blockwise empirical estimator $\widehat{S}$ of Section C.2, trained with an admissible field Q_T, is a fixed function. Each $\widehat{S}_k$ lies in $\Phi(L_k, W_k, R_k, B_k, V_k)$. Hence $\widehat{S}$ is Borel measurable and, for each t, piecewise affine in x, with a Lipschitz constant in x that is uniform over $t \in [\tau, T]$. Since $\sigma_t \leq C\sigma_{t_k}$ for $t \in I_k$,

$$\|\widehat{S}(x,t)\| \leq \sqrt{D}\, V_k \leq C_S \frac{\sqrt{\log n}}{\sigma_t}, \qquad x \in \mathbb{R}^D,\ t \in I_k, \tag{C.1}$$

with C_S depending only on C_V and structural constants. In particular, equation A.11 has a unique nonexplosive strong solution. Set $w_k = 1 \wedge (t_k \log n)^{\gamma/2}$. Constants below do not depend on the training sample.

We begin by showing that an upper bound on the bounded IPM can be extended to an upper bound on the full IPM through truncation. Recall that Π_{out} is projection onto the ball $\overline{B}(0, R_{\text{out}})$.

Lemma C.1. *For every law λ on $\mathbb{R}^D$,*

$$d_\gamma((\Pi_{\text{out}})_\# \lambda, \mu_\eta^Q) \lesssim \bar{d}_\gamma(\lambda, \mu_\eta^Q) + e^{-c/\eta^2}. \tag{C.2}$$

Proposition C.2 (Transfer inequality)**.** *Conditionally on the training sample, for every fixed $A > 0$,*

$$\bar{d}_\gamma(\widehat{\mu}_\tau^{\widehat{S}}, \mu_\eta^Q) \leq Ce^{-cT} + C\tau^{\gamma/2} + C\sum_{k=0}^{K-1} w_k \mathcal{E}_k(\widehat{S})^{1/2} + Cn^{-A}. \tag{C.3}$$

Consequently, with expectations over the training sample,

$$\begin{aligned}\mathbb{E}\,\bar{d}_\gamma(\widehat{\mu}_\tau^{\widehat{S}}, \mu_\eta^Q) + \mathbb{E}\, d_\gamma(\widehat{\mu}_{\tau,\text{out}}^{\widehat{S}}, \mu_\eta^Q) \leq Ce^{-cT} + C\tau^{\gamma/2} + C\sum_{k=0}^{K-1} w_k \big\{\mathbb{E}\mathcal{E}_k(\widehat{S})\big\}^{1/2} \\ + Cn^{-A} + Ce^{-c/\eta^2}.\end{aligned} \tag{C.4}$$

It remains to transfer the raw endpoint bound to the target distribution in the exact-field case.

Lemma C.3. *Let $\mathsf{R}_{\mathcal{M}}$ be as in Theorem A.5. For every law λ on $\mathbb{R}^D$,*

$$d_\gamma\big((\mathsf{R}_{\mathcal{M}})_\# \lambda, p_*\big) \leq C\bar{d}_\gamma(\lambda, \mu_\eta^P) + Ce^{-c/\eta^2}.$$

C.2 Empirical score estimation and completion of the recovery theorem

Throughout this section, the field Q_T of Assumption A.2 is fixed, and we suppress the superscript Q. Let

$$X_t(z) = m_t\{z + \eta Q_N(z)\xi\} + \sigma_t \xi_{\text{iso}},$$

where $\xi, \xi_{\text{iso}} \sim \mathrm{N}(0, I_D)$ are independent. On each block $I_k = [t_k, t_{k+1}]$, define the augmentation-averaged loss

$$\ell_{k,S}(z) := \frac{1}{t_k} \int_{t_k}^{t_{k+1}} t\, \mathbb{E}_{\xi,\xi_{\text{iso}}} \|S(X_t(z), t) + \xi_{\text{iso}}/\sigma_t\|^2 \, dt, \qquad \widehat{S}_k \in \arg\min_{S \in \Phi_k} \frac{1}{n} \sum_{i=1}^{n} \ell_{k,S}(Z_i). \tag{C.5}$$

Here Φ_k is the network class of Theorem A.4, and the blockwise estimator is $\widehat{S}(x,t) = \widehat{S}_k(x,t)$ for $t \in I_k$. Following standard theoretical practice for score matching, we analyze the empirical risk with respect to the finite data sample while retaining the exact expectation over the Gaussian augmentation variables. This isolates the geometric approximation and data-dependent generalization errors from the routine uniform concentration of Monte Carlo noise sampling. The factor t/t_k is bounded above and below on I_k, so it affects the risks below only by universal constants.

Conditionally on $Z = z$, the pair $(\xi_{\text{iso}}, X_t(z))$ is jointly Gaussian and

$$\mathrm{Cov}(\xi_{\text{iso}}, X_t \mid Z = z) = \sigma_t I_D, \qquad \mathrm{Var}(X_t \mid Z = z) = \Sigma_t(z).$$

Consequently, the Gaussian conditioning formula gives

$$\mathbb{E}\left[-\frac{\xi_{\text{iso}}}{\sigma_t} \,\middle|\, X_t = x, Z = z \right] = -\Sigma_t(z)^{-1}(x - m_t z) = -\Lambda_t(z)(x - m_t z). \tag{C.6}$$

The right-hand side is the score of the conditional density of X_t given $Z = z$. Fisher's identity and the tower property therefore imply

$$\mathbb{E}\left[-\frac{\xi_{\text{iso}}}{\sigma_t} \,\middle|\, X_t = x \right] = s_t(x). \tag{C.7}$$

Thus, as in the isotropic case (Vincent, 2011), the usual denoising label remains unbiased for the marginal score, even though it is not equal pathwise to the conditional score in equation C.6. Expanding the square in equation C.5 and using equation C.7 yields

$$\mathbb{E}\ell_{k,S}(Z) - \mathbb{E}\ell_{k,s}(Z) = \mathcal{R}_k(S) := \frac{1}{t_k} \int_{t_k}^{t_{k+1}} t\, \mathbb{E}\|S(X_t, t) - s_t(X_t)\|^2 \, dt. \tag{C.8}$$

Since $1 \le t/t_k \le 2$ on I_k,

$$\mathcal{E}_k(S) \le \mathcal{R}_k(S) \le 2\mathcal{E}_k(S). \tag{C.9}$$

Proposition C.4. *There is a constant depending only on the structural bounds in Assumptions A.1–A.2 such that, uniformly in k, n, and the admissible field Q_T,*

$$\mathbb{E}\mathcal{E}_k(\widehat{S}_k) \lesssim_{\log} \inf_{S \in \Phi_k} \mathcal{E}_k(S) + \frac{R_k + 1}{n}. \tag{C.10}$$

If the empirical criterion is minimized only up to a measurable excess $\zeta_k \ge 0$, the right-hand side acquires the additional term $C\mathbb{E}\zeta_k$.

Completion of Theorem A.5. The scales in equation A.7 satisfy

$$\tau^{\gamma/2} = h^{\alpha+\gamma}, \qquad \frac{h^{-d}}{n} = h^{2\alpha}, \qquad n^{-1/2} = h^{\alpha + d/2}. \tag{C.11}$$

Combining Theorem A.4 with Proposition C.4 and the second identity above gives

$$\mathbb{E}\mathcal{E}_k(\widehat{S}_k) \lesssim_{\log} \begin{cases} h^{2\alpha}, & t_{k+1} \le h^2, \\ J_k/n, & t_k \ge h^2. \end{cases} \tag{C.12}$$

For the small-time blocks, the geometric time grid and $w_k \lesssim_{\log} t_k^{\gamma/2}$ give

$$\sum_{t_{k+1} \le h^2} w_k \{\mathbb{E}\mathcal{E}_k(\widehat{S}_k)\}^{1/2} \lesssim_{\log} h^{\alpha} \sum_{t_{k+1} \le h^2} t_k^{\gamma/2} \lesssim_{\log} h^{\alpha+\gamma}. \tag{C.13}$$

The final inequality follows because the sum is geometric and is controlled by its largest term, $(h^2)^{\gamma/2}$.

For $h^2 \le t_k \le 1$, the definitions of w_k and J_k yield

$$w_k(J_k/n)^{1/2} \le Cn^{-1/2}(\log n)^{\gamma/2}\left(t_k^{\gamma/2} + t_k^{\gamma/2-d/4}\right).$$

Summing over the geometric grid gives

$$\sum_{h^2\le t_k\le 1} w_k(J_k/n)^{1/2} \le Cn^{-1/2}(\log n)^{\gamma/2}\begin{cases} h^{\gamma-d/2}, & d>2\gamma,\\ 1+\log(1/h), & d=2\gamma,\\ 1, & d<2\gamma. \end{cases} \tag{C.14}$$

In the first case, the lower-endpoint contribution satisfies

$$n^{-1/2}h^{\gamma-d/2} = h^{\alpha+d/2}h^{\gamma-d/2} = h^{\alpha+\gamma}. \tag{C.15}$$

The blocks with $t_k \ge 1$ contribute $\lesssim_{\log} n^{-1/2}$.

Substituting equation C.11–equation C.15 into equation C.4, and choosing C_T so that $e^{-cT} \le n^{-1}$, gives, for every admissible field,

$$\mathbb{E}\,\bar{d}_\gamma(\widehat{\mu}_\tau^{\widehat{S}}, \mu_\eta^Q) + \mathbb{E}\,d_\gamma(\widehat{\mu}_{\tau,\mathrm{out}}^{\widehat{S}}, \mu_\eta^Q) \lesssim_{\log} n^{-\frac{\alpha+\gamma}{2\alpha+d}} + n^{-1/2} + e^{-c/\eta^2}. \tag{C.16}$$

Now take $Q_T = P_T$. Applying Lemma C.3 conditionally on the training sample with $\lambda = \widehat{\mu}_\tau^{\widehat{S}^P}$, taking expectations, and using equation C.16 proves equation A.12. If the empirical minimizers have excesses ζ_k, the same argument adds $C\sum_k w_k(\mathbb{E}\zeta_k)^{1/2}$ to the distributional bounds. □

D ESTIMATED GEOMETRY

Split the n observations into a geometry sample of size n_{geo} and an independent score sample of size n_{sc}, with $n_{\mathrm{geo}} \asymp n_{\mathrm{sc}} \asymp n$. Conditionally on the geometry sample, the constructions of the preceding sections are applied to the score sample, with h, η, τ, T in equation A.7 defined using n_{sc} in place of n.

Assumption D.1 (Regular reconstructed surface)**.** *On a geometry sample event $\mathcal{G}_{\mathrm{geo}}$, the reconstruction represents a compact embedded boundaryless d-manifold $\widehat{\mathcal{M}} \subset \overline{B}(0, R_0+1)$, with reach at least $\rho_* > 0$ and a uniformly regular C^β atlas in the sense of Assumption A.1. Its chart radii, chart count, derivative and conditioning bounds, and ρ_* are independent of sample size and geometry sample outcome. Moreover,*

$$\epsilon_M := \sup_{z\in\mathcal{M}} \mathrm{dist}(z, \widehat{\mathcal{M}}) \le \epsilon_0 \le \rho_*/8, \tag{D.1}$$

where ϵ_0 is a sufficiently small structural constant. The surface, its normal projector, and its tubular nearest projection are determined measurably by the geometry sample.

A Hausdorff error bound implies equation D.1. The support estimator of Aamari & Levrard (2019) attains the rate $(\log n/n)^{\beta/d}$ in Hausdorff distance for integer β, but is an unstructured union of polynomial patches and need not itself satisfy Assumption D.1. Theorem D.3 is therefore stated conditionally on a regular reconstruction.

For $z \in \mathcal{M}$, define

$$\pi(z) = \Pi_{\widehat{\mathcal{M}}}(z), \qquad Q_N(z) = P_{N,\widehat{\mathcal{M}}}(\pi(z)), \qquad Q_T(z) = I_D - Q_N(z). \tag{D.2}$$

The next proposition verifies the field condition used in the score proof.

Proposition D.2. *Under Assumption D.1, the field in equation D.2 satisfies Assumption A.2 uniformly on $\mathcal{G}_{\mathrm{geo}}$. In particular,*

$$\max_j \|Q_T \circ G_j^*\|_{C^{\beta-1}} \le C, \qquad \sup_{z\in\mathcal{M}} \|Q_T(z) - P_T(z)\|_{\mathrm{op}} \le C\sqrt{\epsilon_M} \le \tfrac{1}{2}. \tag{D.3}$$

For the independent score observations, define

$$X_{0,i}^Q = Z_i + \eta P_{N,\widehat{\mathcal{M}}}(\pi(Z_i))\xi_i, \tag{D.4}$$

where $\xi_i \sim \mathrm{N}(0, I_D)$ are independent augmentation noises. This display specifies the augmentation law; the theoretical empirical objective averages over fresh augmentation noises as in equation C.5, rather than fixing one perturbation per observation. Train the blockwise score estimators by equation C.5 using the field

$$Q_N(z) = P_{N,\widehat{\mathcal{M}}}(\pi(z)), \qquad Q_T(z) = I_D - Q_N(z).$$

Conditionally on $Z_i = z$ and the geometry sample, the forward mean and covariance are

$$\mathbb{E}[X^Q_{t,i} \mid Z_i = z] = m_t z, \qquad \mathrm{Cov}(X^Q_{t,i} \mid Z_i = z) = \sigma_t^2 Q_T(z) + \nu_t^2 Q_N(z).$$

Consequently, the score approximation and estimation results apply conditionally on the geometry sample.

Let $\widehat{z}_0 \in \widehat{\mathcal{M}}$ be selected measurably from the reconstructed manifold and define the bounded output map

$$\mathsf{R}_{\widehat{\mathcal{M}}}(x) = \begin{cases} \Pi_{\widehat{\mathcal{M}}}(x), & \mathrm{dist}(x, \widehat{\mathcal{M}}) < \rho_*/2, \\ \widehat{z}_0, & \mathrm{dist}(x, \widehat{\mathcal{M}}) \geq \rho_*/2. \end{cases} \tag{D.5}$$

Use equation D.5 on every outcome where the reconstruction is a manifold satisfying the stated regularity and containment bounds. On such outcomes, define the training field at every query point by

$$Q_N(z) = P_{N,\widehat{\mathcal{M}}}\big(\mathsf{R}_{\widehat{\mathcal{M}}}(z)\big), \qquad Q_T(z) = I_D - Q_N(z).$$

On $\mathcal{G}_{\mathrm{geo}}$ this agrees with equation D.2 for every $z \in \mathcal{M}$. On other outcomes, use a fixed rank-$(D-d)$ orthogonal projector for Q_N and a fixed measurable output map into $\overline{B}(0, R_{\mathrm{out}})$.

Theorem D.3 (Recovery with a reconstructed manifold)**.** *Assume the standing model and drift conditions and Assumption D.1. Let $\mathscr{F}_{\mathrm{geo}}$ be the sigma-field generated by the geometry sample, and train the blockwise empirical score estimators using equation D.4. Then, almost surely on $\mathcal{G}_{\mathrm{geo}}$,*

$$\mathbb{E}\left[d_\gamma\left((\mathsf{R}_{\widehat{\mathcal{M}}})_{\#}\widehat{\mu}_\tau^{\widehat{S}^Q}, p_* \right) \,\middle|\, \mathscr{F}_{\mathrm{geo}} \right] \lesssim_{\log} n_{\mathrm{sc}}^{-\frac{\alpha+\gamma}{2\alpha+d}} + n_{\mathrm{sc}}^{-1/2} + \epsilon_M^\gamma + e^{-c/\eta^2}. \tag{D.6}$$

Corollary D.4 (The second statement of Theorem 3.4 in the main text)**.** *Let $\beta \geq 2$ be an integer, and assume in addition that $\mathcal{M}$ is connected. Let $\mathfrak{M}_\beta$ be the class of compact, embedded, boundaryless C^β d-manifolds contained in $\overline{B}(0, R_0 + 1)$, with reach at least $\rho_* > 0$ and with the C^β constants of their inverse-projection parametrizations bounded by L_* in the sense of Aamari & Levrard (2019). Assume that $\rho_* \leq \rho_0$ and that L_* is sufficiently large that $\mathcal{M} \in \mathfrak{M}_\beta$.*

Let $\widetilde{\mathcal{M}}$ be the support estimator of Aamari & Levrard (2019, Theorem 6), constructed from the independent clean geometry sample using the bandwidth and threshold prescribed in that theorem for the fixed model parameters. Since the nonempty compact subsets of $\overline{B}(0, R_0+1)$ form a compact metric space under d_H, the subspace $\mathfrak{M}_\beta$ is separable. Fix a countable d_H-dense subset

$$\{M_i'\}_{i \geq 1} \subset \mathfrak{M}_\beta$$

and, for every $i \geq 1$, fix a point $z_i' \in M_i'$.

Set

$$\epsilon_{n_{\mathrm{geo}}} := \left(\frac{\log n_{\mathrm{geo}}}{n_{\mathrm{geo}}} \right)^{\beta/d}$$

and define

$$\widehat{\imath} := \min\left\{ i \geq 1 : d_H(M_i', \widetilde{\mathcal{M}}) \leq \inf_{j \geq 1} d_H(M_j', \widetilde{\mathcal{M}}) + \epsilon_{n_{\mathrm{geo}}} \right\}, \qquad \widehat{\mathcal{M}} := M_{\widehat{\imath}}'. \tag{D.7}$$

For this reconstruction, define $\mathsf{R}_{\widehat{\mathcal{M}}}$ by equation D.5 on every geometry-sample outcome, with $\widehat{z}_0 = z_{\widehat{\imath}}'$.

Then, for a sufficiently small fixed structural constant $\epsilon_0 > 0$, Assumption D.1 holds on

$$\mathcal{G}_{\mathrm{geo}} := \left\{ d_H(\widetilde{\mathcal{M}}, \mathcal{M}) \leq \epsilon_0/3 \right\}$$

for all sufficiently large n_{geo}. Moreover,

$$\mathbb{E} d_\gamma\left((\mathsf{R}_{\widehat{\mathcal{M}}})_{\#}\widehat{\mu}_\tau^{\widehat{S}^Q}, p_* \right) \lesssim_{\log} n_{\mathrm{sc}}^{-\frac{\alpha+\gamma}{2\alpha+d}} + n_{\mathrm{sc}}^{-1/2} + n_{\mathrm{geo}}^{-\beta\gamma/d} + e^{-c/\eta^2}. \tag{D.8}$$

Proof. Since $M' \mapsto d_H(M', \widetilde{\mathcal{M}})$ is continuous and $\{M'_i\}_{i\geq 1}$ is dense in $\mathfrak{M}_\beta$,

$$\inf_{j\geq 1} d_H(M'_j, \widetilde{\mathcal{M}}) = \inf_{M'\in\mathfrak{M}_\beta} d_H(M', \widetilde{\mathcal{M}}).$$

Because $\mathcal{M} \in \mathfrak{M}_\beta$, the definition of $\widehat{\mathcal{M}}$ and the triangle inequality give

$$d_H(\widehat{\mathcal{M}}, \mathcal{M}) \leq 2d_H(\widetilde{\mathcal{M}}, \mathcal{M}) + \epsilon_{n_{\mathrm{geo}}}. \tag{D.9}$$

The index $\widehat{\imath}$ is measurable because it is defined using countably many measurable Hausdorff distances. Since $\widehat{\mathcal{M}}$ takes values in the countable family $\{M'_i\}_{i\geq 1}$, its tubular projection and normal projector can be chosen jointly measurable in the geometry sample and query point. More precisely, for each i, extend $\Pi_{M'_i}$ outside the $\rho_*/2$ tube by the fixed point z'_i. The resulting maps

$$(\mathcal{X}_{\mathrm{geo}}, x) \longmapsto \mathsf{R}_{\widehat{\mathcal{M}}}(x), \qquad (\mathcal{X}_{\mathrm{geo}}, x) \longmapsto P_{N,\widehat{\mathcal{M}}}\big(\mathsf{R}_{\widehat{\mathcal{M}}}(x)\big)$$

are jointly measurable.

By construction, $\widehat{\mathcal{M}} \in \mathfrak{M}_\beta$ on every geometry-sample outcome. In particular, it is a compact embedded boundaryless C^β d-manifold contained in $\overline{B}(0, R_0 + 1)$, with reach at least ρ_* and with uniformly bounded C^β parametrization constants. These properties provide the chart radii, chart count, derivative bounds, and conditioning bounds required by Assumption D.1, with constants depending only on

$$(\rho_*, L_*, R_0, d, D).$$

On $\mathcal{G}_{\mathrm{geo}}$, and for $\epsilon_{n_{\mathrm{geo}}} \leq \epsilon_0/3$, equation D.9 gives

$$\epsilon_M = \sup_{z\in\mathcal{M}} \mathrm{dist}(z, \widehat{\mathcal{M}}) \leq d_H(\widehat{\mathcal{M}}, \mathcal{M}) \leq \epsilon_0.$$

Together with the membership $\widehat{\mathcal{M}} \in \mathfrak{M}_\beta$ and the preceding measurability properties, this verifies Assumption D.1 on $\mathcal{G}_{\mathrm{geo}}$, provided that ϵ_0 is chosen as a sufficiently small multiple of ρ_*. Proposition D.2 then shows that the pulled-back field satisfies Assumption A.2.

Assumption A.1, including the upper and lower bounds on the density with respect to the volume measure of $\mathcal{M}$, the connectedness assumption of this corollary, and the independent clean geometry sample verify the hypotheses of Aamari & Levrard (2019, Theorem 6) with $k = \beta$. Therefore,

$$\mathbb{E}d_H(\widetilde{\mathcal{M}}, \mathcal{M}) \lesssim \epsilon_{n_{\mathrm{geo}}}. \tag{D.10}$$

Markov's inequality gives

$$\mathbb{P}(\mathcal{G}^c_{\mathrm{geo}}) \leq \frac{3}{\epsilon_0}\mathbb{E}d_H(\widetilde{\mathcal{M}}, \mathcal{M}) \lesssim \epsilon_{n_{\mathrm{geo}}} \leq \epsilon^\gamma_{n_{\mathrm{geo}}}, \tag{D.11}$$

where the last inequality uses $0 < \gamma \leq 1$ and $\epsilon_{n_{\mathrm{geo}}} \leq 1$.

Because $\widehat{\mathcal{M}} \in \mathfrak{M}_\beta$ on every outcome, $\mathsf{R}_{\widehat{\mathcal{M}}}$ always maps into $\overline{B}(0, R_0 + 1)$. Since p_* is supported in $\overline{B}(0, R_0)$,

$$d_\gamma\left((\mathsf{R}_{\widehat{\mathcal{M}}})_\# \widehat{\mu}_\tau^{\widehat{S}^Q}, p_*\right) \leq (2R_0 + 1)^\gamma$$

on $\mathcal{G}^c_{\mathrm{geo}}$. Consequently, the bad-event contribution is $O(\epsilon^\gamma_{n_{\mathrm{geo}}})$ by equation D.11.

On $\mathcal{G}_{\mathrm{geo}}$, Theorem D.3 applies. Moreover, equation D.9, the monotonicity and concavity of $x \mapsto x^\gamma$ and Jensen's inequality give

$$\begin{aligned}\mathbb{E}\left[\epsilon^\gamma_M \mathbf{1}_{\mathcal{G}_{\mathrm{geo}}}\right] &\leq \mathbb{E}\left[\left(2d_H(\widetilde{\mathcal{M}}, \mathcal{M}) + \epsilon_{n_{\mathrm{geo}}}\right)^\gamma\right] \\ &\leq \left(2\mathbb{E}d_H(\widetilde{\mathcal{M}}, \mathcal{M}) + \epsilon_{n_{\mathrm{geo}}}\right)^\gamma \lesssim \epsilon^\gamma_{n_{\mathrm{geo}}}.\end{aligned} \tag{D.12}$$

Taking expectations in equation D.6, using equation D.11 and equation D.12, and recalling that

$$\epsilon^\gamma_{n_{\mathrm{geo}}} = \left(\frac{\log n_{\mathrm{geo}}}{n_{\mathrm{geo}}}\right)^{\beta\gamma/d},$$

proves equation D.8. □

Remark D.5 (Noninteger smoothness and interpretation)**.** *The preceding corollary assumes integer β because Aamari & Levrard (2019, Theorem 6) is stated for integer C^k regularity. For noninteger β, the same argument with $k = \lfloor\beta\rfloor$ gives the geometric contribution $\lesssim_{\log} n_{\mathrm{geo}}^{-\lfloor\beta\rfloor\gamma/d}$. Obtaining the exponent $\beta\gamma/d$ for noninteger β would require a corresponding Hölder-smooth extension of the support estimation theorem.*

The minimum-distance projection onto $\mathfrak{M}_\beta$ is an existence construction for the statistical upper bound and is not claimed to be computationally efficient. It produces a regular reconstruction attaining the minimax Hausdorff rate up to logarithmic factors, but differs from the practical local geometric estimators used in the experiments.

E Proof of Technical Lemmas

E.1 Proof of Lemma B.1

Proof. Conditionally on Z, $X_t - m_t Z$ has covariance $\Sigma_t(Z) \preceq \nu_t^2 I_D$, which gives the probability bound in (i) after choosing $C_0 = C_0(A)$.

Put $\mathsf{H}_t = \Lambda_t(Z)(X_t - m_t Z)$. Conditionally on Z we have $\mathsf{H}_t \sim \mathrm{N}(0, \Lambda_t(Z))$ because $\Sigma_t = \Lambda_t^{-1}$. Hence, by Fisher's identity and the Gaussian moment formulas,

$$\begin{aligned} s_t(X_t) &= -\mathbb{E}[\mathsf{H}_t \mid X_t], \\ \mathbb{E}\|s_t(X_t)\|^2 &\overset{(1)}{\leq} \mathbb{E}\|\mathsf{H}_t\|^2 = \mathbb{E}\,\mathrm{tr}\,\Lambda_t(Z) = \frac{d}{\sigma_t^2} + \frac{D-d}{\nu_t^2} \overset{(2)}{\leq} \frac{D}{\sigma_t^2}, \\ \mathbb{E}\|\mathsf{H}_t\|^4 &= \mathbb{E}\big\{(\mathrm{tr}\,\Lambda_t(Z))^2 + 2\,\mathrm{tr}(\Lambda_t(Z)^2)\big\} \leq \frac{D(D+2)}{\sigma_t^4}, \end{aligned} \tag{E.1}$$

where (1) is conditional Jensen and (2) uses $\nu_t \geq \sigma_t$. Cauchy–Schwarz with the fourth moment and the probability bound gives the second half of equation B.3.

For (ii), we first consider the case $t \geq h^2$. Then $\nu_t \asymp \sigma_t$. Because $x \in \mathcal{T}_t$, the distance to its nearest point y_x is at most $C_0\nu_t\sqrt{\log n}$. For any y in a chart ball of radius $c\sigma_t$ about y_x, the triangle inequality ensures $\|x - m_t y\| \leq C(C_0)\nu_t\sqrt{\log n}$. This bounds the energy by $\Psi_t(x,y) \leq C(C_0)\log n$ on this entire ball. Since the manifold geometry guarantees this ball has volume at least $c\sigma_t^d$, integrating the exponential lower bound $e^{-\Psi_t}$ over this volume directly gives (ii). For the case $t \leq h^2$, we have that $m_t \asymp 1$, $\nu_t \asymp h$, and $h\sqrt{\log n} = o(\rho_0)$, so nearest projection onto $m_t\mathcal{M}$ is unique. Center chart coordinates at y_x and set

$$F^*(z) = U^*(z)^\top(x - m_t G^*(z)), \qquad J_0 = U^*(0)^\top \nabla G^*(0),$$

so $\|J_0^{-1}\| \leq C$ by Lemma A.3. We seek a zero of F^*, representing a point where the residual is strictly normal to the approximate bundle. On a ball of radius $C(C_0)\nu_t\sqrt{\log n}$,

$$\big\|I_d + (m_t J_0)^{-1}\nabla F^*(z)\big\| \leq C\big(\|\nabla U^*\|\,\|x - m_t G^*(z)\| + \|U^{*\top}\nabla G^*(z) - J_0\|\big) = o(1),$$

The first term is bounded by the total residual $O(\nu_t\sqrt{\log n})$, and the second by the Lipschitz variation $O(\|z\|) = O(\nu_t\sqrt{\log n})$. At the center, stationarity of the true nearest point ensures $x - m_t y_x \perp P_T(y_x)$, which gives

$$\|F^*(0)\| = \|Q_T(y_x)(x - m_t y_x)\| = \|(Q_T(y_x) - P_T(y_x))(x - m_t y_x)\| \leq C\epsilon_P C_0 \nu_t\sqrt{\log n}.$$

The contraction mapping theorem applied to $z \mapsto z + (m_t J_0)^{-1}F^*(z)$ therefore supplies a zero z_c in that ball. On a $c\sigma_t$ ball about z_c we have $\|F^*\| \leq C\sigma_t$ and complementary residual at most $C(C_0)\nu_t\sqrt{\log n}$, so $\Psi_t \leq C\log n$ there and integration gives (ii) with C_{den} depending only on C_0 and structural constants. For $Q_T = P_T$ the zero is the nearest point itself.

For (iii), having fixed C_0 and C_{den} in the preceding parts, choose

$$C_1/2 > A + C_{\mathrm{den}} + (d+4)\kappa + 4, \tag{E.2}$$

where κ is the polynomial lower-bound exponent in equation A.9. On $\mathcal{T}_t$ we have $\|x - m_t y\| \leq C\sqrt{\log n}$ for all $y \in \mathcal{M}$, and off $\mathcal{A}_t(x)$ the kernel is at most $n^{-C_1/2}$. Writing $\mathcal{D}^\omega, \mathcal{N}^\omega$ for the retained denominator and numerator,

$$0 \leq \mathcal{D}_t - \mathcal{D}^\omega \leq n^{-C_1/2}, \qquad \|\mathcal{N}_t - \mathcal{N}^\omega\| \leq C\sigma_t^{-2}\sqrt{\log n}\, n^{-C_1/2},$$

and $\|\mathcal{N}_t/\mathcal{D}_t\| \le C\sigma_t^{-2}\sqrt{\log n}$. The denominator floor and equation E.2 first give $\mathcal{D}^\omega \ge \mathcal{D}_t/2$. Subtracting the ratios,

$$\begin{aligned}\Big\|\frac{\mathcal{N}_t}{\mathcal{D}_t} - \frac{\mathcal{N}^\omega}{\mathcal{D}^\omega}\Big\| &\le \frac{\|\mathcal{N}_t - \mathcal{N}^\omega\| + \|\mathcal{N}_t/\mathcal{D}_t\|(\mathcal{D}_t - \mathcal{D}^\omega)}{\mathcal{D}^\omega} \\ &\le C\sigma_t^{-(d+2)}\sqrt{\log n}\, n^{C_{\mathrm{den}} - C_1/2} \le n^{-A}/\sigma_t.\end{aligned} \tag{E.3}$$

The last inequality uses $\sigma_t \ge cn^{-\kappa}$ and equation E.2. Taking $\omega = \mathbf{1}_{\mathcal{A}_t(x)}$ and using $\|\Lambda_t(y)(x - m_t y)\| \le C\sqrt{\log n}/\sigma_t$ on the active set proves equation B.4. An active y satisfies $\|x - m_t y\| \le \nu_t\sqrt{C_1 \log n}$; with the same bound at y_x and $\nu_t \asymp h$, the triangle inequality gives (iv). □

E.2 Proof of Lemma B.10

Proof. By differentiating, we get that

$$\nabla F^* = (\nabla U^*)^\top (x - m_t G^*) - m_t U^{*\top} \nabla G^*.$$

Define the central Jacobian as $J_0 = U^*(0)^\top \nabla G^*(0)$, which is uniformly invertible by Lemma A.3. On the extended chart domain, the first term of ∇F^* is $O(r_h)$. Since $U^{*\top}\nabla G^*$ is $C^{\beta-1}$ with $\beta \ge 2$, it is Lipschitz continuous, and hence can be expanded as $-m_t J_0 + O(r_h)$. Preconditioning on J_0 hence implies

$$\theta_0 := \sup_z \|I_d + (m_t J_0)^{-1}\nabla F^*(z)\| + \sup_z \|I_d + (m_t J_0)^{-1}\nabla F(z)\| \le Cr_h = o(1). \tag{E.4}$$

Integrating equation E.4 along the line segment between any z_1, z_2 in the enlarged chart yields

$$\|(m_t J_0)^{-1}(F^*(z_1) - F^*(z_2)) + (z_1 - z_2)\| \le \theta_0 \|z_1 - z_2\|.$$

This immediately implies the inverse Lipschitz bound

$$\|F^*(z_1) - F^*(z_2)\| \;\ge\; \frac{1 - \theta_0}{\|(m_t J_0)^{-1}\|}\,\|z_1 - z_2\| \;\ge\; c\, m_t \|z_1 - z_2\|. \tag{E.5}$$

An identical argument holds for F, which proves equation B.20, provided they exist.

To this end, for $\|q\|_\infty \le 2\sigma_t L_n$, consider

$$\Theta_q(z) = z + (m_t J_0)^{-1}(F^*(z) - q),$$

whose fixed points satisfy $F^*(z) = q$. The bound equation E.4 implies that Θ_q is a θ_0-contraction. Because the chart is centered at $G^*(0) = y_j$, the gating condition $\rho_j > 0$ guarantees

$$\|F^*(0)\| \le \|x - m_t y_j\| \le 3r_h, \quad \|\Theta_q(0)\| \lesssim r_h + \sigma_t L_n \lesssim r_h.$$

By selecting a ball of radius $C' r_h$ with C' large enough so that $\|\Theta_q(0)\| \le (1 - \theta_0)C' r_h$, Θ_q preserves this ball. The contraction mapping theorem thus gives us a fixed point. Applying the same argument to F follows since $\|F(0) - F^*(0)\| \lesssim_{\log} h^\beta = o(r_h)$.

Now if we evaluate the maps at a common point, equation B.19 and $\nu_t \asymp h$ imply that $\|F - F^*\| \lesssim_{\log} h^\beta$ and $\|\nabla F - \nabla F^*\| \lesssim_{\log} h^{\beta-1}$. The four terms of $\nabla(F - F^*)$ are bounded by $h^{\beta-2}O(hL_n)$, $O(h^\beta)$, $O(h^{\beta-1})$ and $O(h^{\beta-1})$. Combining equation E.5 and $\|F - F^*\| \lesssim_{\log} h^\beta$ implies $\|\mathsf{Z} - \mathsf{Z}^*\| \lesssim_{\log} h^\beta$. As for the Jacobians,

$$\begin{aligned}\|\nabla F(\mathsf{Z}) - \nabla F^*(\mathsf{Z}^*)\| &\le \|\nabla F(\mathsf{Z}) - \nabla F(\mathsf{Z}^*)\| + \|\nabla F(\mathsf{Z}^*) - \nabla F^*(\mathsf{Z}^*)\| \\ &\le \|\nabla^2 F\|_\infty \|\mathsf{Z} - \mathsf{Z}^*\| + \|\nabla F - \nabla F^*\|_\infty \\ &\lesssim_{\log} h^\beta + h^{\beta-1} \lesssim_{\log} h^{\beta-1}.\end{aligned} \tag{E.6}$$

Since $\nabla \mathsf{Z} = (\nabla F \circ \mathsf{Z})^{-1}$ and matrix inversion is Lipschitz on this uniformly conditioned class, equation B.21 follows. An analogous argument applied to the bounded derivative of U provides the same frame bound. Lastly, equation B.22 follows by adding and subtracting at $\mathsf{Z}^*(q)$ as both unprojected residuals are $O(hL_n)$ before normal projection is applied. □

E.3 Proof of Lemma B.12

Proof. Note that a, a^* are bounded from above and below by structural constants. The triangle inequality implies

$$\begin{aligned}\|a - a^*\|_\infty &\le C\big(\|v(\mathsf{Z}) - v(\mathsf{Z}^*)\|_\infty + \|v(\mathsf{Z}^*) - v^*(\mathsf{Z}^*)\|_\infty + \|\nabla\mathsf{Z} - \nabla\mathsf{Z}^*\|_\infty\big) \\ &\lesssim_{\log} h^\beta + h^\alpha + h^{\beta-1} \lesssim_{\log} h^\alpha,\end{aligned} \tag{E.7}$$

where C contains the (bounded) Jacobian determinants and density factors. For the exponents, we have by Lemma B.10 that

$$\frac{\big|\|r_\perp\|^2 - \|r_\perp^*\|^2\big|}{2\nu_t^2} \le \frac{(\|r_\perp\| + \|r_\perp^*\|)\,\|r_\perp - r_\perp^*\|}{2\nu_t^2} \lesssim_{\log} \frac{\nu_t L_n \cdot h^\beta}{\nu_t^2} \lesssim_{\log} h^{\beta-1}. \tag{E.8}$$

since $\nu_t \asymp h$. As $\alpha \le \beta - 1$, combining equation E.7 and equation E.8 gives the first claim. Now suppose $|\vartheta/\vartheta^* - 1| \le \epsilon < 1/2$. Then their integrals obey the same relative bound and their normalized densities have L^1-distance at most $2\epsilon/(1-\epsilon)$. Taking $\epsilon = C(\log n)^c h^\alpha$ for the bound just obtained gives us the TV bound claim for sufficiently large n.

For the tangential summand of the score, we write T_1, T_1^* for the first summand of equation B.25 and its starred version. Then by the bound $\|u\| \le \sqrt{d}L_n$ and the frame bound in Lemma B.10,

$$\|T_1 - T_1^*\| \le \frac{\sqrt{d}L_n}{\sigma_t}\big(\sup_q \|U(\mathsf{Z}(q)) - U^*(\mathsf{Z}^*(q))\| + 2\|\mu - \mu^*\|_{\mathrm{TV}}\big) \lesssim_{\log} \frac{h^{\beta-1} + h^\alpha}{\sigma_t}. \tag{E.9}$$

For the normal summand, we use T_2, T_2^* for the second summand of equation B.25. We then have

$$\|T_2 - T_2^*\| \le \frac{1}{\nu_t^2}\Big\{\sup_q \|r_\perp(q) - r_\perp^*(q)\| + 2\sup_q \|r_\perp^*(q)\|\,\|\mu - \mu^*\|_{\mathrm{TV}}\Big\} \lesssim_{\log} \frac{h^\beta}{\nu_t^2} + \frac{L_n\, h^\alpha}{\nu_t}. \tag{E.10}$$

Now notice that since $\sigma_t \le \nu_t \asymp h$ and $\beta - 1 \ge \alpha$, multiplying equation E.10 by σ_t gives us

$$\sigma_t \cdot \frac{h^\beta}{\nu_t^2} \lesssim h^{\beta-2}\sigma_t \lesssim h^{\beta-1} \le h^\alpha, \qquad \sigma_t \cdot \frac{h^\alpha}{\nu_t} \lesssim h^{\alpha-1}\sigma_t \lesssim h^\alpha. \tag{E.11}$$

Together with equation E.9 and the cube-domain truncation error, this proves equation B.26. □

E.4 Proof of Lemma B.13

Proof. Fix a chart and write $r(z) = x - m_t G(z)$ and $J_F(z) = \nabla_z F(z)$. We first give the finite networks approximating $U, \nabla U, F, J_F, \mathsf{Z}$, then display the integrands and the score network.

Evaluate $G, H, v, \nabla G, \nabla H$ by monomial networks and affine layers. To construct $U = H(H^\top H)^{-1/2}$, set $\mathsf{G}_H = H^\top H$ and choose $B_* > \sup\|\mathsf{G}_H\|$. With $c_j = 4^{-j}\binom{2j}{j}$ and $\mathsf{E}_H = I_d - \mathsf{G}_H/B_*$,

$$\mathsf{G}_H^{-1/2} = B_*^{-1/2}\sum_{j=0}^{\infty} c_j \mathsf{E}_H^j, \qquad \|\mathsf{E}_H\| \le \theta < 1. \tag{E.12}$$

For $m_{\mathrm{ser}} = \lceil C_A \log n \rceil$, let $P_{m_{\mathrm{ser}}}(\mathsf{G}_H)$ be the truncated sum. Its directional derivative is the polynomial

$$\nabla P_{m_{\mathrm{ser}}}(\mathsf{G}_H)[E] = -B_*^{-3/2}\sum_{j=1}^{m_{\mathrm{ser}}} c_j \sum_{a=0}^{j-1} \mathsf{E}_H^a E \mathsf{E}_H^{j-1-a}.$$

The value and derivative tails are bounded by $C\theta^{m_{\mathrm{ser}}}$ and $C\sum_{j>m_{\mathrm{ser}}} j\theta^{j-1}$, respectively. Let

$$\begin{aligned}\phi_{\mathsf{G}_H} &= \phi_{\mathrm{mm}}(\phi_H^\top, \phi_H), \\ \phi_U &= \phi_{\mathrm{mm}}\big(\phi_H, \phi_{\mathrm{poly},P_{m_{\mathrm{ser}}}}(\phi_{\mathsf{G}_H})\big), \\ \phi_{\nabla_b U} &= \phi_{\mathrm{mm}}\big(\phi_{\nabla_b H}, \phi_{\mathrm{poly},P_{m_{\mathrm{ser}}}}(\phi_{\mathsf{G}_H})\big) + \phi_{\mathrm{mm}}\big(\phi_H, \phi_{\mathrm{poly},\nabla P_{m_{\mathrm{ser}}}}(\phi_{\mathsf{G}_H}, \phi_{\nabla_b \mathsf{G}_H})\big),\end{aligned} \tag{E.13}$$

where $\phi_{\nabla_b \mathsf{G}_H}$ uses $(\nabla_b H)^\top H + H^\top \nabla_b H$. To control the network size, we evaluate the truncated matrix polynomial and its directional derivative via a shared Horner recurrence. For an input direction E, initialize $\mathsf{B}_{m_{\mathrm{ser}}} = c_{m_{\mathrm{ser}}} I_d$ and $\mathsf{C}_{m_{\mathrm{ser}}} = 0$, and for $j = m_{\mathrm{ser}} - 1, \dots, 0$ use

$$\mathsf{B}_j = c_j I_d + \mathsf{E}_H \mathsf{B}_{j+1}, \qquad \mathsf{C}_j = -(E/B_*)\mathsf{B}_{j+1} + \mathsf{E}_H \mathsf{C}_{j+1}.$$

Then $P_{m_{\mathrm{ser}}}(\mathsf{G}_H) = B_*^{-1/2}\mathsf{B}_0$ and $\nabla P_{m_{\mathrm{ser}}}(\mathsf{G}_H)[E] = B_*^{-1/2}\mathsf{C}_0$. The two polynomial modules in equation E.13 mean the ReLU realizations of these recurrences, which require $O(\log n)$ matrix products for each of the fixed number of coordinate directions. Their intermediate values remain bounded on the stated real spectral domain.

For a uniformly invertible $d \times d$ matrix J, the formula $J^{-1} = \mathrm{adj}(J)/\det J$ yields the network

$$\phi_{\mathrm{inv}}(J) = \phi_{\mathrm{mult}}\left(\phi_{\mathrm{adj}}(J),\, \phi_{\mathrm{mult}}\big(\phi_{\det}(J), \phi_{\mathrm{rec};c_J^2/2}(\phi_{\mathrm{mult}}(\phi_{\det}(J), \phi_{\det}(J)))\big)\right). \tag{E.14}$$

Here $|\det J| \geq c_J$ on the valid domain, and $\phi_{\det}, \phi_{\mathrm{adj}}$ evaluate their fixed polynomials using multiplication modules.

Define

$$\begin{aligned}
\phi_r(x,t,z) &= x - \phi_{\mathrm{mult}}(\phi_m(t), \phi_G(z)),\\
\phi_F(x,t,z) &= \phi_{\mathrm{mm}}(\phi_U(z)^\top, \phi_r(x,t,z)),\\
[\phi_{J_F}(x,t,z)]_{ab} &= \sum_{c=1}^{D} \phi_{\mathrm{mult}}([\phi_{\nabla_b U}(z)]_{ca}, [\phi_r]_c) - \phi_{\mathrm{mult}}\left(\phi_m(t), \sum_{c=1}^{D} \phi_{\mathrm{mult}}([\phi_U]_{ca}, [\phi_{\nabla G}]_{cb})\right).
\end{aligned} \tag{E.15}$$

The exact Newton sequence from $z^{(0)} = 0$ is

$$z^{(\ell+1)} = z^{(\ell)} - J_F(z^{(\ell)})^{-1}\{F(z^{(\ell)}) - q\}, \qquad \|z^{(\ell)} - \mathsf{Z}(q)\| \leq C(Cr_h)^{2^\ell}. \tag{E.16}$$

Uniform bounds on $J_F^{-1}, \nabla^2 F$ and $\mathsf{Z}(q) = O(r_h)$ justify this sequence on an enlarged chart. Since $r_h \lesssim_{\log} n^{-1/(2\alpha+d)}$, a fixed number $\ell_* = \ell_*(A)$ gives any prescribed inverse-polynomial accuracy. Its ReLU realization is the recursive composition

$$\begin{aligned}
\phi_z^{(0)} &= 0,\\
\phi_z^{(\ell+1)} &= \mathrm{Clip}_{\mathcal{C}_j}\left[\phi_z^{(\ell)} - \phi_{\mathrm{mm}}\left(\phi_{\mathrm{inv}}(\phi_{J_F}(x,t,\phi_z^{(\ell)})),\, \phi_F(x,t,\phi_z^{(\ell)}) - q\right)\right], \qquad \phi_{\mathsf{Z}} = \phi_z^{(\ell_*)}.
\end{aligned} \tag{E.17}$$

We choose the coordinate box $\mathcal{C}_j$ strictly inside the enlarged chart with a fixed relative margin around every valid iterate. Clipping ensures exact evaluation points stay within the valid domain, while uniform conditioning and finite induction control numerical perturbations. Taking the polynomial and primitive accuracies sufficiently small preserves the strict determinant bounds required for equation E.14. (This adapts the Newton construction from Tang & Yang (2024, Lemma C.8); the conditioning and convergence for this matched coordinate map were verified in Lemma B.10).

For $u \in \mathcal{U}_n$, put $z_u = \mathsf{Z}(\sigma_t u)$, $r_u = x - m_t G(z_u)$, and define

$$U_u = U(z_u), \quad r_{\perp,u} = r_u - U_u U_u^\top r_u, \quad \vartheta_u = \frac{v(z_u)}{|\det J_F(z_u)|} \exp\left(-\frac{\|r_{\perp,u}\|^2}{2\nu_t^2}\right). \tag{E.18}$$

The three integrals to approximate are

$$\mathcal{D} = \int_{\mathcal{U}_n} e^{-\|u\|^2/2}\vartheta_u\, du, \qquad \mathcal{N}_\parallel = \int_{\mathcal{U}_n} e^{-\|u\|^2/2}\vartheta_u U_u u\, du, \qquad \mathcal{N}_\perp = \int_{\mathcal{U}_n} e^{-\|u\|^2/2}\vartheta_u r_{\perp,u}\, du. \tag{E.19}$$

Thus the local numerator is $\mathcal{N} = \sigma_t^{-1}\mathcal{N}_\parallel + \nu_t^{-2}\mathcal{N}_\perp$, and $\widetilde{s} = -\mathcal{N}/\mathcal{D}$. Here $\mathcal{N}_\parallel$ and $\mathcal{N}_\perp$ are the unscaled tangent-frame and complementary-residual moments, respectively. Lemma B.12 and the preceding denominator bounds give, with $d_{1,k} = cn^{-C_{\mathrm{den}}}$ for a sufficiently small fixed $c > 0$,

$$d_{1,k} \leq \mathcal{D} \leq C, \qquad \|\mathcal{N}_\parallel\| \leq \sqrt{d} L_n \mathcal{D}, \qquad \|\mathcal{N}_\perp\| \leq C\nu_t L_n \mathcal{D}. \tag{E.20}$$

We prove directly that each full integrand in equation E.19 has an approximation by a polynomial in u of degree $O(\log n)$, to any prescribed inverse-polynomial accuracy. All polynomial degrees below refer to u or z with (x,t) fixed; their coefficients may depend on (x,t).

Fix a target exponent $A_{\mathrm{int}} > 0$ and an integer $p \geq 2$, to be chosen depending on A_{int} and the structural constants but not on n. Work on an enlarged ball $\|z\| \leq C_{\mathrm{ball}} r_h$ containing the inverse coordinates, with a fixed relative margin. Since $H(0) = U^*(0)$ has orthonormal columns, the polynomial matrix $\mathsf{E}(z) = H(z)^\top H(z) - I_d$ satisfies $\mathsf{E}(0) = 0$, $\|\mathsf{E}(z)\| \leq C r_h$, and $\|\nabla \mathsf{E}(z)\| \leq C$. Define

$$U^{[p]}(z) = H(z) \sum_{j=0}^{p} (-1)^j c_j \mathsf{E}(z)^j, \qquad c_j = 4^{-j} \binom{2j}{j}.$$

The real binomial series for $(I_d + \mathsf{E})^{-1/2}$, with its termwise first derivative, gives

$$\|U^{[p]} - U\|_\infty \leq C_p r_h^{p+1}, \qquad \|\nabla U^{[p]} - \nabla U\|_\infty \leq C_p r_h^p. \tag{E.21}$$

Indeed, the two tails are bounded by $C \sum_{j>p} (C r_h)^j$ and $C \sum_{j>p} j (C r_h)^{j-1}$, respectively. The degrees and the first two derivative bounds of $U^{[p]}$ depend only on p and the structural constants.

Put $F^{[p]}(z) = U^{[p]}(z)^\top (x - m_t G(z))$ and $J_c = \nabla F^{[p]}(0)$. Uniform transversality and $\|x - m_t G(0)\| \leq 3 r_h$ imply

$$\|J_c^{-1}\| \leq C_p, \qquad \sup_{\|z\| \leq C_{\mathrm{ball}} r_h} \|I_d - J_c^{-1} \nabla F^{[p]}(z)\| \leq C_p r_h.$$

For $u \in \mathcal{U}_n$, define the polynomial iteration

$$w^{(0)}(u) = 0, \qquad w^{(\ell+1)}(u) = w^{(\ell)}(u) - J_c^{-1}\{F^{[p]}(w^{(\ell)}(u)) - \sigma_t u\}, \qquad w_p = w^{(p)}. \tag{E.22}$$

The matrix J_c^{-1} is independent of u. The bound $\|J_c^{-1}(F^{[p]}(0) - \sigma_t u)\| \leq C r_h$ and the preceding contraction estimate show, after choosing C_{ball} large enough, that every iterate stays in the enlarged ball. For $z_u = \mathsf{Z}(\sigma_t u)$, write $e_\ell = \sup_{u \in \mathcal{U}_n} \|w^{(\ell)}(u) - z_u\|$. Since $F(z_u) = \sigma_t u$, $\|x - m_t G(z_u)\| \leq C r_h$, and equation E.21 holds,

$$e_0 \leq C r_h, \qquad e_{\ell+1} \leq C_p r_h e_\ell + C_p r_h^{p+2}, \qquad e_p \leq C_p r_h^{p+1}.$$

Thus w_p is a polynomial in u of degree bounded independently of n. More explicitly, if $b_p = \max\{1, \deg F^{[p]}\}$, its degree is at most b_p^p. Large constants depending on p do not introduce a power of n.

Next set

$$\mathfrak{d}_0 = \det J_c, \qquad \mathfrak{d}_p(u) = \det \nabla F^{[p]}(w_p(u)), \qquad a_p(u) = \frac{1}{|\mathfrak{d}_0|} \sum_{j=0}^{p} \left(-\frac{\mathfrak{d}_p(u) - \mathfrak{d}_0}{\mathfrak{d}_0} \right)^j.$$

Here $|\mathfrak{d}_0| \geq c_p > 0$ and $|\mathfrak{d}_p/\mathfrak{d}_0 - 1| \leq C_p r_h$. For large n, $\mathfrak{d}_p$ and $\mathfrak{d}_0$ have the same sign. The finite geometric-series identity, the coordinate error, and equation E.21 consequently give

$$\sup_{u \in \mathcal{U}_n} \left| a_p(u) - \frac{1}{|\det J_F(z_u)|} \right| \leq C_p r_h^p. \tag{E.23}$$

Define further polynomials

$$\begin{aligned}
\overline{U}_p(u) &= U^{[p]}(w_p(u)), & \overline{r}_p(u) &= x - m_t G(w_p(u)), \\
\overline{r}_{\perp,p}(u) &= (I_D - \overline{U}_p(u) \overline{U}_p(u)^\top) \overline{r}_p(u), & A_{0,p}(u) &= v(w_p(u)) a_p(u), \\
A_{\parallel,p}(u) &= A_{0,p}(u) \overline{U}_p(u) u, & A_{\perp,p}(u) &= A_{0,p}(u) \overline{r}_{\perp,p}(u), \\
E_p(u) &= \tfrac{1}{2}\|u\|^2 + \frac{\|\overline{r}_{\perp,p}(u)\|^2}{2\nu_t^2}.
\end{aligned}$$

Their degrees are bounded by a constant depending only on p and the structural constants. On the real cube, $E_p \geq 0$. Since $\nu_t \asymp h$ on a small-time block and $r_h \asymp h L_n$, the preceding estimates give

$$\|\overline{U}_p - U_u\|_\infty + \|\overline{r}_{\perp,p} - r_{\perp,u}\|_\infty \leq C_p r_h^{p+1}, \qquad 0 \leq E_p(u) \leq C_p \log n.$$

Let $g_0, g_\parallel, g_\perp$ denote the scalar or vector full integrands in equation E.19, including $e^{-\|u\|^2/2}$. The reciprocal error equation E.23, bounded polynomial derivatives, and $|e^{-a} - e^{-b}| \leq |a - b|$ for $a, b \geq 0$ show that

$$\max_{\diamond \in \{0,\parallel,\perp\}} \|g_\diamond - A_{\diamond,p} e^{-E_p}\|_\infty \leq C_p (1 + L_n)^2 (1 + \nu_t^{-2}) r_h^p \leq C_p n^{C_{\mathrm{alg}}} r_h^p. \tag{E.24}$$

The exponent C_{alg} can be chosen independently of p: only the displayed schedule factor and powers of L_n enter it. As $r_h \le n^{-1/(2(2\alpha+d))}$ for all sufficiently large n, a fixed large $p = p(A_{\mathrm{int}})$ makes equation E.24 at most $\frac{1}{2}n^{-A_{\mathrm{int}}}$.

Finally take the finite Taylor polynomial

$$T_M(a) = \sum_{j=0}^{M} \frac{(-a)^j}{j!}, \qquad M = \lceil C_{A_{\mathrm{int}}} \log n \rceil.$$

For $0 \le a \le C_p \log n$, the real Taylor remainder satisfies

$$|e^{-a} - T_M(a)| \le \frac{a^{M+1}}{(M+1)!} \le \left(\frac{eC_p \log n}{M+1}\right)^{M+1}.$$

The amplitudes $A_{\diamond,p}$ are bounded by a fixed polynomial in n. Increasing $C_{A_{\mathrm{int}}}$ therefore gives the polynomial approximants

$$P_\diamond(u) = A_{\diamond,p}(u) T_M(E_p(u)), \qquad \deg P_\diamond \le C'_{A_{\mathrm{int}}} \log n, \qquad \max_{\diamond} \|g_\diamond - P_\diamond\|_\infty \le n^{-A_{\mathrm{int}}}. \tag{E.25}$$

This construction is uniform over all valid (x,t) and charts.

If $P(u) = \sum_\iota b_\iota u^\iota$, its integral is the explicit moment sum

$$\int_{[-L_n,L_n]^d} P(u)\,du = \sum_\iota b_\iota \prod_{j=1}^{d} \begin{cases} 2L_n^{\iota_j+1}/(\iota_j+1), & \iota_j \text{ even}, \\ 0, & \iota_j \text{ odd}. \end{cases}$$

For the network implementation, use the equivalent positive-weight evaluation formula: tensor Gauss–Legendre quadrature with $m_{\mathrm{quad}} = \lceil C''_{A_{\mathrm{int}}} \log n \rceil$ nodes per coordinate, chosen so $2m_{\mathrm{quad}} - 1$ is at least the degree in equation E.25. This formula integrates each $P_\diamond$ exactly, has positive weights w_i, and satisfies $\sum_i w_i = (2L_n)^d$; these are the elementary polynomial-exactness properties of Gauss quadrature (DLMF, Section 3.5(v)). Consequently, componentwise,

$$\left| \int_{\mathcal{U}_n} g_\diamond(u)\,du - \sum_i w_i g_\diamond(u_i) \right| \le 2(2L_n)^d \|g_\diamond - P_\diamond\|_\infty \le 2(2L_n)^d n^{-A_{\mathrm{int}}}. \tag{E.26}$$

For vector integrands the same bound holds in Euclidean norm up to a fixed dimensional constant. Increasing A_{int} gives any required inverse-polynomial integration accuracy with $J_{\mathrm{quad}} = m_{\mathrm{quad}}^d = O((\log n)^d)$ nodes.

Let $(u_i, w_i)_{i=1}^{J_{\mathrm{quad}}}$ be the tensor nodes and positive weights, with $J_{\mathrm{quad}} = m_{\mathrm{quad}}^d$, and set $\lambda_i = w_i e^{-\|u_i\|^2/2}$. All node sums below run from 1 to J_{quad}. These are fixed network weights. At each node put

$$\begin{aligned}
\phi_{z,i} &= \phi_{\mathsf{Z}}(x,t,\phi_\sigma(t)u_i), \quad \phi_{U,i} = \phi_U(\phi_{z,i}), \quad \phi_{r,i} = \phi_r(x,t,\phi_{z,i}), \\
\phi_{r_\perp,i} &= \phi_{r,i} - \phi_{\mathrm{mm}}\big(\phi_{U,i}, \phi_{\mathrm{mm}}(\phi_{U,i}^\top, \phi_{r,i})\big), \\
\phi_{a,i} &= \phi_{\mathrm{mult}}\left(\phi_v(\phi_{z,i}), \phi_{\mathrm{rec};c_J/2}(|\phi_{\det}(\phi_{J_F}(x,t,\phi_{z,i}))|)\right), \\
\phi_{e,i} &= \phi_{\exp}\left(\mathrm{Clip}_{[0,C\log n]}\left[\tfrac{1}{2}\phi_{\mathrm{mult}}\big(\phi_{\nu^{-2}}(t), \phi_{\mathrm{dot}}(\phi_{r_\perp,i}, \phi_{r_\perp,i})\big)\right]\right), \\
\phi_{\vartheta,i} &= \phi_{\mathrm{mult}}(\phi_{a,i}, \phi_{e,i}).
\end{aligned} \tag{E.27}$$

Then form the finite sums and the local score network

$$\begin{aligned}
\phi_{\mathcal{D}} &= \sum_i \lambda_i \phi_{\vartheta,i}, \qquad \phi_{\mathcal{N}_\parallel} = \sum_i \lambda_i \phi_{\mathrm{mult}}\big(\phi_{\vartheta,i}, \phi_{\mathrm{mm}}(\phi_{U,i}, u_i)\big), \\
\phi_{\mathcal{N}_\perp} &= \sum_i \lambda_i \phi_{\mathrm{mult}}(\phi_{\vartheta,i}, \phi_{r_\perp,i}), \\
\phi_{\mathrm{loc}} &= \mathrm{Clip}_{[-V_k,V_k]^D}\left[-\phi_{\mathrm{mult}}\left(\phi_{\mathrm{mult}}(\phi_{\sigma^{-1}}, \phi_{\mathcal{N}_\parallel}) + \phi_{\mathrm{mult}}(\phi_{\nu^{-2}}, \phi_{\mathcal{N}_\perp}),\ \phi_{\mathrm{rec};d_{1,k}/2}(\phi_{\mathcal{D}})\right)\right].
\end{aligned} \tag{E.28}$$

Choose quadrature, coordinate, matrix-series, and primitive accuracies so that the sum of the three errors for $\phi_{\mathcal{D}}, \phi_{\mathcal{N}_\parallel}, \phi_{\mathcal{N}_\perp}$ is at most

$$\delta_{\mathrm{int},k} = \frac{c\, d_{1,k} n^{-A}}{1 + L_n + \sigma_{t_k} \nu_{t_k}^{-2}}. \tag{E.29}$$

This is an inverse-polynomial precision. The quadrature weights have polylogarithmic total mass, and all real intermediate ranges and relevant Lipschitz constants are polynomial in n. Thus a sufficiently small inverse-polynomial primitive accuracy and a sufficiently large fixed ℓ_*, $m_{\mathrm{ser}}/\log n$, and $m_{\mathrm{quad}}/\log n$ achieve this budget. Products with the schedule modules and the final reciprocal are evaluated at accuracies that make their additional error at most half the final tolerance. Using equation E.20 and the floored-ratio inequality gives

$$\sigma_{t_k} \|\phi_{\mathrm{loc}} - \mathrm{Clip}_{[-V_k, V_k]^D}\, \widetilde{s}\| \le C d_{1,k}^{-1} \delta_{\mathrm{int},k} (1 + L_n + \sigma_{t_k} \nu_{t_k}^{-2}) \le n^{-A}, \tag{E.30}$$

after reducing the constant in equation E.29 to allow for those final arithmetic errors.

The base chart polynomials have fixed degree, and the Newton iteration requires only a fixed number of compositions. The shared matrix recurrence computes $O(\log n)$ sequential products of fixed-size matrices; because binary multiplication at inverse-polynomial precision requires depth and sparsity $O(\log n)$, this recurrence demands depth and sparsity $O((\log n)^2)$. At the level of individual tensor nodes, the reciprocal modules dictate the local bounds, requiring width $O((\log n)^3)$ and sparsity $O((\log n)^4)$, with the schedule, exponential, and arithmetic modules comfortably fitting within these envelopes. Consequently, each individual quadrature node evaluates with depth $O((\log n)^2)$, width $O((\log n)^3)$, and sparsity $O((\log n)^4)$. Parallelizing these computations across the $J_{\mathrm{quad}} = O((\log n)^d)$ tensor nodes, and applying the final affine summation, reciprocal, and clipping operations, yields the stated global width $W = O((\log n)^{d+3})$ and sparsity $R = O((\log n)^{d+4})$. Including the identity connections necessary to carry the fixed-dimensional inputs and outputs through the layers, the overall network depth and log-parameter bounds strictly satisfy the claimed $O((\log n)^4)$ envelopes. □

E.5 Proof of Lemma C.1

Proof. The truncation map Π_{out} is nonexpansive, and its range has a diameter of at most $2R_{\mathrm{out}}$. Thus, we can bound the distance between mapped points by the minimum of their original distance and the diameter. Pointwise, for any coupling of x and y, this yields:

$$\|\Pi_{\mathrm{out}}(x) - \Pi_{\mathrm{out}}(y)\|^\gamma \le \max\{1, (2R_{\mathrm{out}})^\gamma\}(1 \wedge \|x - y\|^\gamma).$$

Taking the infimum over all valid couplings between λ and μ_η^Q directly bounds the distance between the pushed-forward measures:

$$d_\gamma\big((\Pi_{\mathrm{out}})_\# \lambda, (\Pi_{\mathrm{out}})_\# \mu_\eta^Q\big) \lesssim \bar{d}_\gamma(\lambda, \mu_\eta^Q). \tag{E.31}$$

Next, we bound the truncation error on the target distribution. Because $X_0 = Z + \eta Q_N(Z)\xi$ under μ_η^Q, and the manifold support satisfies $\|Z\| \le R_0$, the truncation is only active when the normal noise exceeds the strictly positive margin $R_{\mathrm{out}} - R_0$. Thus,

$$\mathbb{E}\|X_0 - \Pi_{\mathrm{out}}(X_0)\|^\gamma \le \mathbb{E}[(\eta\|\xi\|)^\gamma \mathbf{1}_{\eta\|\xi\| > R_{\mathrm{out}} - R_0}] \le C e^{-c/\eta^2}.$$

where the final bound follows from standard Gaussian tail concentration. This controls the distance $d_\gamma\big((\Pi_{\mathrm{out}})_\# \mu_\eta^Q, \mu_\eta^Q\big)$. Applying the triangle inequality through the intermediate measure $(\Pi_{\mathrm{out}})_\# \mu_\eta^Q$ yields the claim. □

E.6 Proof of Proposition C.2

We first prove the following lemma.

Lemma E.1. *Start both processes at p_T and let both use the exact reverse score down to forward time t_{k+1}. On $[t_k, t_{k+1}]$, the first uses s_t and the second uses $\widehat{S}$; below t_k both use $\widehat{S}$. Let λ_k^0, λ_k^1 be their endpoint laws at forward time τ. Then, for every fixed $A > 0$,*

$$\bar{d}_\gamma(\lambda_k^0, \lambda_k^1) \le C w_k\, \mathcal{E}_k(\widehat{S})^{1/2} + C n^{-A}. \tag{E.32}$$

Proof. Keep the fixed-model convention of Section A. The exact reverse process is the time reversal of the isotropic OU process started at μ_η^Q. We follow the entropy and coupling strategy of Tang & Yang (2024, Appendix D.1), establishing the displacement estimate directly.

Both processes have the same state $V_{\mathrm{com}} \sim p_{t_{k+1}}$ at forward time t_{k+1}. On I_k the reference process uses s_t and has marginals p_t, whereas the other process uses $\widehat{S}$. Their drift difference is $2b_t(\widehat{S} - s_t)$ and their diffusion coefficient is $\sqrt{2b_t}$, so their block path laws satisfy

$$\mathrm{KL}(\mathsf{P}^0_{\mathrm{blk}} \| \mathsf{P}^1_{\mathrm{blk}}) \le \frac{1}{2}\mathbb{E}\int_{t_k}^{t_{k+1}} \left\| \frac{2b_t(\widehat{S}(X_t,t) - s_t(X_t))}{\sqrt{2b_t}} \right\|^2 dt \le b_+ \, \mathcal{E}_k(\widehat{S}). \tag{E.33}$$

To justify the change of measure, stop the canonical paths upon exit from a ball. On the stopped interval the drift difference is bounded, so Girsanov's theorem applies. Nonexplosion and lower semicontinuity of relative entropy allow passage to increasing balls. For $t \ge \tau$, the mixture formula gives $\|s_t(x)\| \le \sigma_\tau^{-2}(\|x\| + R_0)$, and $\widehat{S}$ is bounded by equation C.1, so both drifts have at most linear growth and the processes are nonexplosive. The reference energy is finite by the Fisher bound equation E.1, the envelope equation C.1, and $\tau > 0$.

Append the common $\widehat{S}$-driven continuation from t_k to τ, and let $\mathsf{P}^0, \mathsf{P}^1$ denote the resulting joint laws of V_{com} and the continuation paths. Appending a common transition kernel does not increase relative entropy, so Pinsker's inequality gives $d_{\mathrm{TV}}(\mathsf{P}^0, \mathsf{P}^1) \le C\mathcal{E}_k(\widehat{S})^{1/2}$. Since the two laws have the same V_{com}-marginal, couple their continuation paths maximally conditional on V_{com}. If $\mathcal{B}$ denotes the event that the coupled paths differ, then

$$\mathbb{P}(\mathcal{B}) = d_{\mathrm{TV}}(\mathsf{P}^0, \mathsf{P}^1) \le C\mathcal{E}_k(\widehat{S})^{1/2},$$

and their endpoints Y^0, Y^1 agree on $\mathcal{B}^c$.

Suppose first that $t_{k+1} \le 1$. We show that, under either marginal law,

$$\mathbb{P}\Big(\|Y^i - V_{\mathrm{com}}\| > C_A\sqrt{t_{k+1}\log n}\Big) \le Cn^{-A}, \qquad i = 0, 1. \tag{E.34}$$

The forward marginals have uniformly sub-Gaussian tails: conditional on $Z = z$ their means are $m_t z$ with $\|z\| \le R_0$, and their covariances satisfy $\Sigma_t(z) \preceq I_D$ by equation A.9. Hence

$$\mathbb{P}_{X\sim p_t}\Big(\|X\| > C_A\sqrt{\log n}\Big) \le Cn^{-A} \tag{E.35}$$

uniformly in $t \in [\tau, T]$.

Each path in equation E.34 consists of at most two consecutive segments: an exact segment on I_k, present only for λ_k^0, followed by an $\widehat{S}$-driven segment down to τ. We bound the displacement across each segment by $C_A\sqrt{t_{k+1}\log n}$ outside an event of probability Cn^{-A}; the two bounds then add, and adjusting C_A gives equation E.34.

For the exact segment, the states at forward times t_k, t_{k+1} have the forward OU joint law

$$X_{t_{k+1}} = \frac{m_{t_{k+1}}}{m_{t_k}} X_{t_k} + \zeta_k^{\mathrm{OU}}, \qquad \zeta_k^{\mathrm{OU}} \sim \mathrm{N}\left(0, \left[1 - \frac{m_{t_{k+1}}^2}{m_{t_k}^2}\right] I_D\right),$$

with ζ_k^{OU} independent of X_{t_k}. Since $|1 - m_{t_{k+1}}/m_{t_k}| \le C(t_{k+1} - t_k) \le Ct_{k+1}$ and $1 - m_{t_{k+1}}^2/m_{t_k}^2 \le Ct_{k+1}$, applying equation E.35 to X_{t_k} and Gaussian concentration to ζ_k^{OU} gives

$$\|X_{t_{k+1}} - X_{t_k}\| \le Ct_{k+1}\sqrt{\log n} + C_A\sqrt{t_{k+1}\log n} \le C_A\sqrt{t_{k+1}\log n}$$

outside an event of probability Cn^{-A}, using $t_{k+1} \le 1$.

For the $\widehat{S}$-driven segment, run from forward time v down to r with $\tau \le r \le v \le t_{k+1}$ and set $\ell = v - r$. In reverse elapsed time its dynamics are

$$dY_u = b_{v-u}\{Y_u + 2\widehat{S}(Y_u, v-u)\}\, du + \sqrt{2b_{v-u}}\, dB_u, \qquad 0 \le u \le \ell.$$

Variation of constants and the envelope equation C.1 yield

$$\|Y_\ell - Y_0\| \le C\ell\|Y_0\| + C\sqrt{\log n}\int_r^v \frac{dt}{\sigma_t} + \|\mathcal{G}_\ell\|,$$

where the centered Gaussian stochastic convolution is

$$\mathcal{G}_\ell := \int_0^\ell \exp\left(\int_a^\ell b_{v-s}\,ds\right)\sqrt{2b_{v-a}}\,dB_a.$$

Since $\ell \le 1$, its covariance is bounded by $2b_+e^{2b_+}\ell I_D$. The initial state has law $p_{t_{k+1}}$ for λ_k^1 and p_{t_k} for λ_k^0, so equation E.35 gives $\|Y_0\| \le C_A\sqrt{\log n}$ outside an event of probability Cn^{-A}; since $\ell \le t_{k+1} \le 1$ we have $\ell \le \sqrt{t_{k+1}}$, whence $C\ell\|Y_0\| \le C_A\sqrt{t_{k+1}\log n}$. For the drift term, $\sigma_t \ge c\sqrt{t}$ on $(0,1]$ gives

$$C\sqrt{\log n}\int_r^v \frac{dt}{\sigma_t} \le C\sqrt{\log n}\int_0^{t_{k+1}} t^{-1/2}\,dt \le C\sqrt{t_{k+1}\log n}, \tag{E.36}$$

and Gaussian concentration bounds $\|\mathcal{G}_\ell\|$ by $C_A\sqrt{\ell\log n} \le C_A\sqrt{t_{k+1}\log n}$ outside an event of probability Cn^{-A}. This proves equation E.34.

Let $\mathcal{H}$ be the event that both displacement bounds hold, so that $\mathbb{P}(\mathcal{H}^c) \le Cn^{-A}$. On $\mathcal{B}\cap\mathcal{H}$,

$$1 \wedge \|Y^0 - Y^1\|^\gamma \le 1 \wedge \left(2C_A\sqrt{t_{k+1}\log n}\right)^\gamma \le Cw_k,$$

since $t_{k+1} \le 2t_k$. The cost vanishes on $\mathcal{B}^c$ and is at most one on $\mathcal{H}^c$, so

$$\bar{d}_\gamma(\lambda_k^0,\lambda_k^1) \le \mathbb{E}\big[(1\wedge\|Y^0-Y^1\|^\gamma)\mathbf{1}_{\mathcal{B}\cap\mathcal{H}}\big] + \mathbb{P}(\mathcal{H}^c) \le Cw_k\,\mathcal{E}_k(\widehat{S})^{1/2} + Cn^{-A}.$$

If $t_{k+1} > 1$, then $t_k > 1/2$ and $w_k = 1$, and the same conclusion follows from the bounded cost and the bound on $\mathbb{P}(\mathcal{B})$. □

The exact reverse process started at p_T ends at p_τ by time reversal, and the coupling $X_\tau = m_\tau X_0 + \sigma_\tau\xi_{\text{iso}}$ gives

$$\bar{d}_\gamma(p_\tau,\mu_\eta^Q) \le \mathbb{E}\|(m_\tau-1)X_0 + \sigma_\tau\xi_{\text{iso}}\|^\gamma \le \left(\mathbb{E}\|(m_\tau-1)X_0+\sigma_\tau\xi_{\text{iso}}\|^2\right)^{\gamma/2} \le C\tau^{\gamma/2},$$

where we used $1-m_\tau \asymp \tau$, $\sigma_\tau^2 \asymp \tau$, and $\mathbb{E}\|X_0\|^2 \le C$.

For sufficiently large T, $\Sigma_T(y) = I_D + O(m_T^2)$ uniformly in y. The Gaussian relative-entropy formula gives

$$\mathrm{KL}\big(\mathrm{N}(m_Ty,\Sigma_T(y))\,\|\,\mathrm{N}(0,I_D)\big) = \tfrac{1}{2}\{\operatorname{tr}\Sigma_T - D - \log\det\Sigma_T + m_T^2\|y\|^2\} \le Cm_T^2. \tag{E.37}$$

Convexity and Pinsker therefore yield $d_{\mathrm{TV}}(p_T,\mathrm{N}(0,I_D)) \le Ce^{-cT}$.

For $0 \le j \le K$ let λ_j be the endpoint law at forward time τ of the process started at p_T that uses s_t for $t > t_j$ and $\widehat{S}$ for $t \le t_j$, so that $\lambda_0 = p_\tau$ and λ_K is the law generated by $\widehat{S}$ throughout, started at p_T. These interpolating processes are used only in the analysis; the sampler never evaluates s_t. Adjacent hybrids differ only on I_j and therefore have exactly the form of Lemma E.1. Applying it on each block with tail accuracy $n^{-(A+1)}$ and using the triangle inequality for $\bar{d}_\gamma$,

$$\bar{d}_\gamma(\lambda_K,p_\tau) \le \sum_{j=0}^{K-1}\left(Cw_j\mathcal{E}_j(\widehat{S})^{1/2} + Cn^{-(A+1)}\right) \le C\sum_j w_j\mathcal{E}_j(\widehat{S})^{1/2} + Cn^{-A},$$

the last step because $K = O(\log n)$. The law $\widehat{\mu}_\tau^{\widehat{S}}$ differs from λ_K only in its initialization, both being obtained by applying the same $\widehat{S}$-driven reverse kernel; data processing and $\bar{d}_\gamma \le d_{\mathrm{TV}}$ therefore give $\bar{d}_\gamma(\widehat{\mu}_\tau^{\widehat{S}},\lambda_K) \le d_{\mathrm{TV}}(\mathrm{N}(0,I_D),p_T) \le Ce^{-cT}$. Combining these two bounds with the early-stopping estimate by the triangle inequality proves equation C.3.

The constants in equation C.3 do not depend on the training sample. Taking expectations and using $\mathbb{E}\,\mathcal{E}_k(\widehat{S})^{1/2} \le \{\mathbb{E}\mathcal{E}_k(\widehat{S})\}^{1/2}$ by Jensen's inequality bounds $\mathbb{E}\,\bar{d}_\gamma(\widehat{\mu}_\tau^{\widehat{S}},\mu_\eta^Q)$. Lemma C.1, applied conditionally with $\lambda = \widehat{\mu}_\tau^{\widehat{S}}$ and then averaged, bounds the second term and proves equation C.4.

E.7 Proof of Lemma C.3

Proof. Take a near-optimal coupling (X, X_0) of $X \sim \lambda$ and $X_0 \sim \mu_\eta^P$ for the metric $\bar{d}_\gamma$. By the gluing lemma, we can expand this to a joint distribution (X, X_0, Z, ξ) such that the conditional law of (Z, ξ) given X_0 matches the exact generative process $X_0 = Z + \eta P_N(Z)\xi$, where $Z \sim p_*$. Under this joint distribution, the pair $(\mathsf{R}_\mathcal{M}(X), Z)$ forms a valid coupling between the pushed-forward measure $(\mathsf{R}_\mathcal{M})_\# \lambda$ and the true data distribution p_*.

Define the high-probability event $E = \{\eta\|\xi\| \le \rho_0/4\}$, which guarantees the additive noise does not push X_0 outside the manifold's valid tubular neighborhood. On E, the vector $X_0 - Z$ is strictly normal to $\mathcal{M}$ at Z and has a length strictly less than the reach ρ_0. By the fundamental properties of the tubular neighborhood, the nearest-point projection is unique, meaning $\Pi_\mathcal{M}(X_0) = Z$ exactly.

To bound the coupling cost on E, we analyze two spatial regimes for the displacement $\|X - X_0\|$:

1. **($\|X - X_0\| \le \rho_0/4$):** By the triangle inequality, the distance from X to the manifold is at most $\|X - X_0\| + \|X_0 - Z\| \le \rho_0/2$. On this $\rho_0/2$-tube, the metric projection $\Pi_\mathcal{M}$ is well-defined and 2-Lipschitz. Because both X and X_0 lie in this tube, we have
$$\|\mathsf{R}_\mathcal{M}(X) - Z\| = \|\Pi_\mathcal{M}(X) - \Pi_\mathcal{M}(X_0)\| \le 2\|X - X_0\|.$$

2. **($\|X - X_0\| > \rho_0/4$):** Both points $\mathsf{R}_\mathcal{M}(X)$ and Z lie on the compact manifold $\mathcal{M}$, so their maximum distance is trivially bounded by the diameter $2R_0$. Because the displacement between X and X_0 is bounded away from zero, the truncated cost $1 \wedge \|X - X_0\|^\gamma$ is strictly bounded below by $c(\rho_0) > 0$. Consequently, there exists a constant C' such that
$$\|\mathsf{R}_\mathcal{M}(X) - Z\|^\gamma \le (2R_0)^\gamma \le C'\big(1 \wedge \|X - X_0\|^\gamma\big).$$

Combining both regimes, we find that on the event E, the cost is uniformly bounded by the coupled displacement: $\|\mathsf{R}_\mathcal{M}(X) - Z\|^\gamma \le C\big(1 \wedge \|X - X_0\|^\gamma\big)$.

Finally, on the complement event E^c, the distance between any two manifold points remains bounded by $2R_0$, yielding a maximum cost of $(2R_0)^\gamma$. Standard Gaussian tail concentration for the scaled noise $\eta\xi$ gives $\mathbb{P}(E^c) \le Ce^{-c/\eta^2}$.

Taking expectations over the joint distribution, the total transport cost is bounded by the sum of the expected cost on E and the tail mass from E^c. Taking the infimum over all such couplings completes the proof. □

E.8 Proof of Proposition C.4

Proof. Fix $S_0 \in \Phi_k$, and write $\mathbb{E}_n g = n^{-1}\sum_{i=1}^n g(Z_i)$ and $\mathbb{E}_* g = \int g\, dp_*$. We verify the envelope, Bernstein, and entropy conditions for the standard localized empirical-risk argument.

Let V_k be the output bound. Jensen's inequality and the Gaussian second moment give, uniformly in z and $S \in \Phi_k$,
$$0 \le \ell_{k,S}(z) \le CM_k, \qquad M_k := \int_{t_k}^{t_{k+1}} \frac{t}{t_k}\left(V_k^2 + \frac{D}{\sigma_t^2}\right) dt. \tag{E.38}$$

Using $t_{k+1} \le 2t_k$, $\sigma_t^2 \asymp t$ for $t \le 1$, $\sigma_t \asymp 1$ for $t \ge 1$, and the bounds on V_k, we obtain
$$M_k \le \begin{cases} C\log n, & t_k \le 1, \\ Ct_k \log n \le C(\log n)^2, & t_k > 1. \end{cases} \tag{E.39}$$

Write the difference of two squared losses as the inner product of $S - S_0$ with $S + S_0 + 2\xi_{\text{iso}}/\sigma_t$. Applying Cauchy–Schwarz first in $(t, \xi, \xi_{\text{iso}})$ conditionally on $Z = z$, and then averaging over z, gives
$$\begin{aligned}\mathbb{E}\{\ell_{k,S}(Z) - \ell_{k,S_0}(Z)\}^2 &\le CM_k \int_{I_k} \frac{t}{t_k}\mathbb{E}\|S(X_t, t) - S_0(X_t, t)\|^2 dt \\ &\le CM_k\{\mathcal{R}_k(S) + \mathcal{R}_k(S_0)\}. \end{aligned}\tag{E.40}$$

The last inequality follows by inserting s_t and using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$.

Fix $A \ge 3$. On $[-C_A\sqrt{\log n}, C_A\sqrt{\log n}]^D \times I_k$, take a uniform cover of the clipped networks with accuracy

$$\delta_k = \frac{n^{-A}}{C(1+T+M_k)}. \tag{E.41}$$

Conditionally on $Z = z$, the mean of $X_t(z)$ is uniformly bounded and its covariance satisfies $\Sigma_t(z) \preceq I_D$. Taking C_A sufficiently large makes the contribution from the complement of this box at most Cn^{-A}, uniformly in z and t. Inside the box, Cauchy–Schwarz and equation E.38 show that the network cover induces a uniform cover of the losses with accuracy Cn^{-A}. If $\mathfrak{N}_k$ denotes its cardinality, the standard sparse-ReLU covering bound gives

$$\log \mathfrak{N}_k \le C R_k L_k \log\{(n/\delta_k)(1+W_k)(1+B_k)(1+R_k)\} \lesssim_{\log} R_k + 1. \tag{E.42}$$

Bernstein's inequality, a union bound over the loss cover, and equation E.40 imply that, with probability at least $1 - e^{-u}$, simultaneously over $S \in \Phi_k$,

$$\begin{aligned} |(\mathbb{E}_n - \mathbb{E}_*)(\ell_{k,S} - \ell_{k,S_0})| \le \frac{1}{4}\{\mathcal{R}_k(S) + \mathcal{R}_k(S_0)\} \\ + \frac{C M_k(\log \mathfrak{N}_k + u + 1)}{n} + Cn^{-A}. \end{aligned} \tag{E.43}$$

By empirical optimality, $\mathbb{E}_n(\ell_{k,\widehat{S}_k} - \ell_{k,S_0}) \le \zeta_k$, while equation C.8 gives

$$\mathbb{E}_*(\ell_{k,S} - \ell_{k,S_0}) = \mathcal{R}_k(S) - \mathcal{R}_k(S_0).$$

Combining these displays and absorbing $\frac{1}{4}\mathcal{R}_k(\widehat{S}_k)$ into the left-hand side yields

$$\mathcal{R}_k(\widehat{S}_k) \le \frac{5}{3}\mathcal{R}_k(S_0) + C\left\{\zeta_k + \frac{M_k(\log \mathfrak{N}_k + u + 1)}{n} + n^{-A}\right\}.$$

Integrating in u, using equation E.39, equation E.42, and equation C.9, and then taking the infimum over S_0 proves equation C.10. This is the localized bounded-loss argument used in Oko et al. (2023, Theorem C.4); the calculations above verify its hypotheses for the anisotropic augmentation. □

E.9 Proof of Proposition D.2

Proof. On a fixed smaller reconstructed chart $\widehat{G}$, choose a local orthonormal normal frame $\widehat{V}_N(u) \in \mathbb{R}^{D\times(D-d)}$ by Gram orthonormalization. The atlas bounds give uniform $C^{\beta-1}$ bounds on $\widehat{V}_N$ and on $P_{T,\widehat{\mathcal{M}}}(\widehat{G}(u)) = \nabla\widehat{G}(u)(\nabla\widehat{G}(u)^\top\nabla\widehat{G}(u))^{-1}\nabla\widehat{G}(u)^\top$. The normal-coordinate map

$$\widehat{F}_{\mathrm{tube}}(u,a) = \widehat{G}(u) + \widehat{V}_N(u)a, \qquad a \in \mathbb{R}^{D-d}, \quad \|a\| \le \rho_*/2,$$

is uniformly $C^{\beta-1}$. Its differential has uniformly bounded inverse: the tangential block is $(I - \mathcal{S}_{\widehat{V}_N(u)a})\nabla\widehat{G}(u)$, where $\mathcal{S}_\nu$ is the shape operator of $\widehat{\mathcal{M}}$ in normal direction ν: $\mathcal{S}_\nu v = -P_{T,\widehat{\mathcal{M}}}\nabla_v\widetilde{\nu}$ for a local normal extension $\widetilde{\nu}$ of ν. Here I is the tangent-space identity, and the shape operator satisfies $\|\mathcal{S}_{\widehat{V}_N(u)a}\|_{\mathrm{op}} \le \|a\|/\rho_* \le 1/2$ by the reach bound. The normal block is an isometry and the remaining block is bounded. The quantitative inverse function theorem on smaller patches therefore gives uniform $C^{\beta-1}$ bounds for the tubular projection and for the tangent projector composed with it. This argument also covers noninteger β with the stated Hölder convention; for the integer projection-regularity statement see Leobacher & Steinicke (2021, Theorem 2). Composition with the charts of $\mathcal{M}$ proves the first bound in equation D.3.

For the second, fix a unit $v \in T_z\mathcal{M}$. Uniform C^2 charts give a curve $c(r) \in \mathcal{M}$ with $c(0) = z$ and $c(r) = z + rv + O(r^2)$. Put $p = \pi(z)$ and $p_r = \pi(c(r))$. Projection is 2-Lipschitz on the half-reach tube, hence $\|p_r - p\| \le Cr$. The reach bound also implies

$$\|P_{N,\widehat{\mathcal{M}}}(p)(p_r - p)\| \le \frac{\|p_r - p\|^2}{2\rho_*}.$$

Indeed, compare squared distances from $p + a\mathfrak{n}$ to p and p_r for each unit normal $\mathfrak{n}$ and $0 < a < \rho_*$, then let $a \uparrow \rho_*$. Both $\|c(r) - p_r\|$ and $\|z - p\|$ are at most ϵ_M, so

$$\|Q_N(z)v\| \le C\big(r + \epsilon_M/r\big).$$

Take $r = \sqrt{\epsilon_M}$, or let $r \downarrow 0$ if $\epsilon_M = 0$. Equal-rank orthogonal projectors satisfy $\|Q_T - P_T\|_{\mathrm{op}} = \|Q_N P_T\|_{\mathrm{op}}$, proving the bound. Finally, let $V_{T,j}$ be a uniformly regular orthonormal tangent frame of $\mathcal{M}$ on each smaller chart. The explicit formula

$$U_j^* = (Q_T \circ G_j^*) V_{T,j} \big(V_{T,j}^\top (Q_T \circ G_j^*) V_{T,j}\big)^{-1/2}$$

gives a frame for Q_T: its Gram matrix is uniformly positive definite by the projector discrepancy bound. Smoothness of inverse square root on this uniformly conditioned class gives the required $C^{\beta-1}$ frame bounds. Shrinking ϵ_0 ensures transversality. □

E.10 Proof of Theorem D.3

Proof. Condition on $\mathscr{F}_{\mathrm{geo}}$ and on the event $\mathcal{G}_{\mathrm{geo}}$. Then $\widehat{\mathcal{M}}$, π, Q_T, and Q_N are fixed. Moreover, Proposition D.2 shows that Q_T satisfies the regularity and transversality conditions required by Theorem A.4. Since the score sample is independent of the geometry sample, the observations used for score estimation remain i.i.d. from p_*. Hence Theorem A.4 and Proposition C.4 give

$$\mathbb{E}\left[\mathcal{E}_k(\widehat{S}_k^Q) \,\middle|\, \mathscr{F}_{\mathrm{geo}}\right] \lesssim_{\log} \begin{cases} h^{2\alpha}, & t_{k+1} \le h^2, \\ J_k/n_{\mathrm{sc}}, & t_k \ge h^2. \end{cases} \tag{E.44}$$

Apply Proposition C.2 conditionally on the training sample and on $\mathscr{F}_{\mathrm{geo}}$. Taking conditional expectations given $\mathscr{F}_{\mathrm{geo}}$ and using Jensen's inequality yields

$$\begin{aligned} &\mathbb{E}\left[\bar{d}_\gamma\left(\widehat{\mu}_\tau^{\widehat{S}^Q}, \mu_\eta^Q\right) \,\middle|\, \mathscr{F}_{\mathrm{geo}}\right] \\ &\quad \le Ce^{-cT} + C\tau^{\gamma/2} + C\sum_{k=0}^{K-1} w_k \left\{\mathbb{E}\left[\mathcal{E}_k(\widehat{S}_k^Q) \,\middle|\, \mathscr{F}_{\mathrm{geo}}\right]\right\}^{1/2} + Cn_{\mathrm{sc}}^{-A}. \end{aligned} \tag{E.45}$$

Substituting equation E.44 and applying the same dyadic summations as in equation C.13–equation C.15 gives

$$\mathbb{E}\left[\bar{d}_\gamma\left(\widehat{\mu}_\tau^{\widehat{S}^Q}, \mu_\eta^Q\right) \,\middle|\, \mathscr{F}_{\mathrm{geo}}\right] \lesssim_{\log} n_{\mathrm{sc}}^{-\frac{\alpha+\gamma}{2\alpha+d}} + n_{\mathrm{sc}}^{-1/2}. \tag{E.46}$$

It remains to transfer the raw distribution bound through the projection onto the reconstructed manifold. Throughout this step, we condition on the geometry sample and work on $\mathcal{G}_{\mathrm{geo}}$. Recall that

$$\pi := \Pi_{\widehat{\mathcal{M}}}\big|_{\mathcal{M}}, \qquad \varepsilon_M := \sup_{z \in \mathcal{M}} \|z - \pi(z)\|, \qquad Q_N(z) = P_{N,\widehat{\mathcal{M}}}(\pi(z)),$$

$\mathrm{reach}(\widehat{\mathcal{M}}) \ge \rho_*$, and that $\mathsf{R}_{\widehat{\mathcal{M}}}$ agrees with $\Pi_{\widehat{\mathcal{M}}}$ on the $\rho_*/2$-tube and maps into $\widehat{\mathcal{M}}$ everywhere. By the reconstruction assumptions, $\widehat{\mathcal{M}}$ lies in a fixed bounded set and $\varepsilon_M \le \rho_*/8$. Below, $\pi_\# p_*$ is the law of the projected observation $\pi(Z)$ for $Z \sim p_*$.

First, we identify the projected target of the ideal perturbed law. Let

$$X_0^Q = Z + \eta Q_N(Z)\xi, \qquad Z \sim p_*, \qquad \xi \sim \mathcal{N}(0, I_D),$$

with Z and ξ independent. Since $\pi(Z)$ is the nearest point to Z on $\widehat{\mathcal{M}}$,

$$Z - \pi(Z) \in N_{\pi(Z)}\widehat{\mathcal{M}}.$$

Consequently,

$$X_0^Q - \pi(Z) = Z - \pi(Z) + \eta Q_N(Z)\xi \in N_{\pi(Z)}\widehat{\mathcal{M}}. \tag{E.47}$$

On the event

$$A := \{\eta\|\xi\| \le \rho_*/8\},$$

we have $\|X_0^Q - \pi(Z)\| \le \rho_*/4$. The reach condition therefore implies

$$\mathsf{R}_{\widehat{\mathcal{M}}}(X_0^Q) = \Pi_{\widehat{\mathcal{M}}}(X_0^Q) = \pi(Z).$$

Thus projection cancels the normal perturbation on A, but it returns $\pi(Z)$ rather than Z. Moreover,

$$\mathbb{P}(A^c) \le Ce^{-c/\eta^2}.$$

We next control the effect of the learned diffusion error. For any probability law λ on $\mathbb{R}^D$, couple $X \sim \lambda$ and $X_0^Q \sim \mu_\eta^Q$ nearly optimally for the bounded Hölder transport cost $1 \wedge \|X - X_0^Q\|^\gamma$. Extend this coupling by the conditional law of (Z, ξ) given X_0^Q.

On

$$A \cap \{\|X - X_0^Q\| < \rho_*/4\},$$

both X and X_0^Q lie in the $\rho_*/2$-tube. Nearest-point projection is uniformly Lipschitz on this tube, and hence

$$\|\mathsf{R}_{\widehat{\mathcal{M}}}(X) - \pi(Z)\|^\gamma \le C\|X - X_0^Q\|^\gamma.$$

On $A \cap \{\|X - X_0^Q\| \ge \rho_*/4\}$, boundedness of $\widehat{\mathcal{M}}$ gives the corresponding bound by $C(1 \wedge \|X - X_0^Q\|^\gamma)$. The same boundedness controls the contribution of A^c. Taking the infimum over couplings yields

$$d_\gamma\big((\mathsf{R}_{\widehat{\mathcal{M}}})_\# \lambda, \pi_\# p_*\big) \le C\bar{d}_\gamma(\lambda, \mu_\eta^Q) + Ce^{-c/\eta^2}. \tag{E.48}$$

Finally, we account for the discrepancy between the projected target and the original distribution. The coupling $(\pi(Z), Z)$ gives

$$\begin{aligned} d_\gamma(\pi_\# p_*, p_*) &\le \int_{\mathcal{M}} \|\pi(z) - z\|^\gamma \, p_*(dz) \\ &\le \varepsilon_M^\gamma. \end{aligned} \tag{E.49}$$

Combining the preceding two bounds by the triangle inequality, we obtain, for every λ,

$$d_\gamma\big((\mathsf{R}_{\widehat{\mathcal{M}}})_\# \lambda, p_*\big) \le C\bar{d}_\gamma(\lambda, \mu_\eta^Q) + \varepsilon_M^\gamma + Ce^{-c/\eta^2}.$$

Apply this inequality with $\lambda = \widehat{\mu}_\tau^{\widehat{S}^Q}$ and take conditional expectations given $\mathscr{F}_{\mathrm{geo}}$. The raw distribution bound established above then gives

$$\mathbb{E}\Big[d_\gamma\Big((\mathsf{R}_{\widehat{\mathcal{M}}})_\# \widehat{\mu}_\tau^{\widehat{S}^Q}, p_*\Big) \,\Big|\, \mathscr{F}_{\mathrm{geo}}\Big] \lesssim_{\log} n_{\mathrm{sc}}^{-\frac{\alpha+\gamma}{2\alpha+d}} + n_{\mathrm{sc}}^{-1/2} + \varepsilon_M^\gamma + e^{-c/\eta^2},$$

as claimed. □