HIERARCHICAL FUNCTIONAL DIFFUSION MODELS WITH GLOBAL AND LOCAL STATISTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Function generation is a promising and challenging problem in generative modeling. Many scientific and spatial data are more naturally viewed as random functions or fields, where the goal is to generate an entire function that can be evaluated at any locations and resolutions. Functional diffusion models provide a natural framework for this setting, but a central difficulty is that the Bayes optimal denoiser is conditioned on an infinite dimensional noisy function. Existing approaches typically rely either on finite dimensional latent representations, which may discard important predictive information, or on a large number of sampled function values, which can be computationally expensive. We propose a hierarchical functional diffusion model that addresses this tradeoff by separating global and local predictive structure. A moderate dimensional diffusion model first generates a global state that captures long-range dependence, after which a conditional functional diffusion model generates the residual using compressed local information around each query. We develop an information based error decomposition that separates discretization, localization, compression, and learning errors, and derive Wasserstein bounds for the resulting functional distribution. Our algorithm is particularly well suited to hierarchical Gaussian process mixture models, which allow globally non-Gaussian and multimodal distributions while requiring Gaussian structure only conditionally at the residual scale. The proposed framework provides a scalable, coarse-to-fine, and resolution free approach to generative modeling on function spaces.

1 INTRODUCTION

Diffusion models have been extensively studied for finite-dimensional data, with applications to image and video generation, molecular design, inverse problems, and related fields (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho et al., 2022b; Hoogeboom et al., 2022; Chung et al., 2022). In many scientific applications, however, the object of interest is more naturally viewed as a spatially indexed random function rather than a vector with fixed discretization. Functional diffusion models are designed to generate such objects in a manner that is consistent across query locations and discretization resolutions (Kerrigan et al., 2023; Pidstrigach et al., 2024; Lim et al., 2025). The diffusion framework extends naturally to this setting: A clean random function F_0 is corrupted to a noisy state F_t, and a time-dependent denoiser is learned to recover the clean data. At a query point q, the Bayes optimal denoiser is $D^*(t, F_t)(q) = \mathbb{E}[F_0(q)|F_t]$. The difficulty is that the conditioning variable F_t is itself a function. Although function-to-function architectures such as neural operators provide ways to parameterize such maps, any practical denoiser must ultimately access the current function through a finite representation (Li et al., 2021; Kovachki et al., 2023). This raises a basic statistical question: how much information about F_t is actually needed to predict the clean value $F_0(q)$?

Existing methods access F_t through operator architectures, latent representations, or function values on a set of context points. These approaches face a tradeoff: strong compression may discard information needed to predict $F_0(q)$, whereas conditioning on a large context is computationally expensive and processes observations that contribute little to a given query, and moreover it ties the architecture to a discretization. Using only the observations near q is attractive, and locality is known to reduce the effective dimension of high dimensional diffusion models (Gottwald et al., 2025), but pure localization fails for functional data whose distant regions vary coherently through a shared global geometry. The relevant question is therefore not whether the functional distribution is local,

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.

but whether its nonlocal predictive dependence can first be represented by a moderate dimensional global state, leaving a conditional residual that is substantially more local, and whether the resulting representation has controllable predictive error and scales across resolutions without expanding the architecture.

Motivated by this view, we propose a coarse-to-fine hierarchical diffusion model. Each function is separated into a moderate dimensional global component and a residual component. A finite dimensional diffusion model first generates the global state, after which a conditional functional diffusion model generates the residual using the compressed local information around each query. Importantly, this construction is not a low dimensional manifold assumption. The residual may retain substantial variance and remain high dimensional. In fact, the global representation need not explain most of the total variance. Instead, it is intended to explain the appropriate nonlocal dependence.

We formalize this principle through an information based error decomposition that separates discretization, localization, compression and learning errors, and study conditions under which these errors are small. We then derive the Wasserstein bounds for the generated functional distribution. The theoretical framework is well suited to hierarchical Gaussian process mixture models, which allow the data distribution to be highly flexible and multimodal. We use this class as a controlled benchmark in our experiments, with conditional residual covariances based on the nonstationary Matérn construction of Paciorek & Schervish (2006).

Our contributions. We develop a scalable and resolution independent hierarchical functional diffusion framework, establish quantitative approximation and generation bounds, and introduce a nonstationary Gaussian process mixture benchmark with closed-form Bayes denoisers for controlled evaluation.

2 Background

Setup and Notation. Let $\chi \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain and let $\mathcal{H} = H^s(\chi), s > d/2$, be the Sobolev space on which the data and noise measures are supported. By the Sobolev embedding $\mathcal{H} \hookrightarrow C(\chi)$, point evaluation $f \mapsto f(q)$ is a bounded linear functional for every $q \in \chi$, so $\mathcal{H}$ is an RKHS where pointwise quantities used in the denoising process are well-defined. On $L^2(\chi)$, elements are instead identified modulo equality almost everywhere. We distinguish support and metric spaces: while all probability measures are supported on $\mathcal{H}$, let $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}})$ be a Hilbert space with continuous embedding $\mathcal{H} \hookrightarrow \mathcal{K}$, and measure distributional errors using $\|\cdot\|_{\mathcal{K}}$. A canonical choice is $\mathcal{K} = L^2(\chi)$. We write $\mathcal{L}(X)$ for the law of X, $\mathcal{P}_2(E)$ for Borel probability measures on a Hilbert space E with finite second moment. Let $\mu_0 \in \mathcal{P}_2(\mathcal{H})$ denote the data distribution with $F_0 \sim \mu_0$ and i.i.d. observations $F_0^i \sim \mu_0$.

Gaussian Measure. The forward noise is a centered Gaussian measure $\mu_1 = N(0, C)$ on $\mathcal{H}$ with a positive, self-adjoint, trace-class covariance operator C. Suppose $F_1 \sim \mathcal{N}(0, C)$, then for every $h \in \mathcal{H}$, $\langle F_1, h\rangle_{\mathcal{H}} \sim \mathcal{N}(0, \langle Ch, h\rangle_{\mathcal{H}})$. A Gaussian process with sample paths in $\mathcal{H}$ induces such a Gaussian measure, and we use the two descriptions interchangeably.

Local Regression. Given context points $X_m = \{x_i\}_{i=1}^m \subset \chi$ and a query $q \in \chi$, define the punctured neighborhood

$$N(q, h) = \{x_i \in X_m : 0 < \|x_i - q\| \le h\}.$$

For a function f, let $l(f; q) \in R^{D_p}$ be the coefficients of the degree-p locally weighted polynomial fitted on $N(q, h)$,

$$\hat{\beta}(q) = \underset{\{\beta_\alpha\}}{\arg\min} \sum_{x_i \in N(q,h)} W_h(\|x_i - q\|) \left[f(x_i) - \sum_{|\alpha| \le p} \beta_\alpha (\frac{x_i - q}{h})^\alpha \right]^2, \quad D_p = \binom{p+d}{d},$$

where $W_h(u) = h^{-d} W(u/h)$ and W is compactly supported on the unit ball. The statistic $l(f; q) = \hat{\beta}(q)$ carries the local geometry and variation information around q, while it is fixed dimensional for fixed (p, d), regardless of the number of context points. We denote the polynomial space on R^d with degree no larger than p as $\mathcal{P}_p$.

Related Work. Relevant directions include functional and infinite dimensional diffusion (Kerrigan et al., 2023; Pidstrigach et al., 2024; Lim et al., 2025; Franzese et al., 2023; Hagemann et al., 2025), neural operator and latent representations (Li et al., 2021; Kovachki et al., 2023; Rombach et al., 2022), context based diffusion (Zhang & Wonka, 2024), coarse-to-fine generation (Ho et al., 2022a; Saharia et al., 2022) and localized diffusion Gottwald et al. (2025). Our closest distinction is that the residual diffusion remains query-based after conditioning on a generated global state, and we explicitly quantify localization and compression errors. We also draw on local polynomial regression Fan (2018) and convergence analyses for diffusion samplers (Chen et al., 2022; Lee et al., 2023; De Bortoli, 2022).

3 Hierarchical Functional Diffusion

3.1 From Simultaneous to Hierarchical Denoising

Zhang & Wonka (2024) represent the noisy function through its values on a context set X_m. The corresponding full context Bayes denoiser is $d^*_{\text{full}}(t,q) = \mathbb{E}[F_0(q) \mid F_t(X_m)]$. This requires the denoiser to process all context observations, so its computational cost can grow with the number of context points. A natural alternative is to represent the noisy function through a global summary $A_t = S_L(F_t)$, which captures its overall characteristics, together with a query dependent local feature $l(F_t; q)$. This gives the simultaneous global-local denoiser

$$d^*_{\text{sim}}(t,q) = \mathbb{E}[F_0(q) \mid S_L(F_t),\, l(F_t;q)]\,. \tag{1}$$

The difficulty is that distant observations may improve prediction at q for two conceptually different reasons. First, $A_t = S_L(F_t)$ is corrupted by the forward process and only partially reveals the clean global state $A_0 = S_L(F_0)$, so distant observations may help resolve uncertainty about A_0. Second, even conditional on A_0, the residual field may retain intrinsic long-range predictive dependence. The simultaneous formulation does not distinguish these two sources of nonlocality, motivating the hierarchical construction below.

3.2 Hierarchical Diffusion Framework

Suppose $F_0 \sim \mu_0, F_1 \sim \mu_1 = \mathcal{N}(0, C)$, and $F_t = \alpha_t F_0 + \sigma_t F_1$. Let $S_L : \mathcal{H} \to \mathbb{R}^L$ be a bounded linear encoder with $A_t = S_L(F_t)$, and let $\mathcal{R} : \mathbb{R}^L \to \mathcal{H}$ be a linear reconstruction decoder satisfying $S_L\mathcal{R} = I_L$, where I_L is the identity map on $\mathbb{R}^L$. The subspace $V_A = \operatorname{ran}\mathcal{R}$ is parameterized by A_0, and we write $V_U = \ker S_L$ for its complement. Let $G_0 = \mathcal{R}A_0$ be the reconstructed function and $U_0 = F_0 - G_0 = (I - \mathcal{R}S_L)F_0$ be the residual. Then $S_L U_0 = 0$, i.e. $U_0 \in V_U$. The linearity yields

$$A_t = \alpha_t A_0 + \sigma_t A_1, \qquad U_t = \alpha_t U_0 + \sigma_t U_1, \tag{2}$$

where the two terminal laws are the pushforwards

$$A_1 = S_L F_1 \sim \nu_A := S_{L\#}\mu_1, \qquad U_1 = (I - \mathcal{R}S_L)F_1 \sim \nu_U := (I - \mathcal{R}S_L)_{\#}\mu_1. \tag{3}$$

Since $\mu_1 = \mathcal{N}(0, C)$, both terminal laws are Gaussian, and $\nu_U = \mathcal{N}(0, C_U)$ with $C_U = (I - \mathcal{R}S_L)C(I - \mathcal{R}S_L)^*$. Thus the same forward family induces a finite dimensional diffusion for the global state and a function space diffusion for the residual.

The joint data law can be represented as: $A_0 \sim \mu_A, G_0 \sim \mathcal{R}_{\#}\mu_A, U_0 \mid A_0 = a \sim \mu_{U_0|A_0=a}$. Writing $H_L(a)(\cdot) = \mathbb{E}[U_0 \mid A_0 = a]$ and $U_0 = H_L(A_0)(x) + \epsilon_0(x)$, we obtain

$$F_0 = \underbrace{G_0 + H_L(A_0)(x)}_{\text{global term}} + \underbrace{\epsilon_0(x)}_{\text{local term}}\,.$$

The residual may still be high dimensional and retain high variance, and conditioning only removes the predictive dependence encoded by A_0.

Denoising. We first learn a finite dimensional denoiser $D_\phi(A_t, t) \approx A_0$ by minimizing

$$\mathbb{E}_{A_0\sim\mu_A, A_1\sim\nu_A, t\in\text{Unif}[0,1]}\|D_\phi(A_t,t) - A_0\|^2.$$

Conditioning on the clean global state A_0, we train the residual denoiser

$$D_\theta(t, A_0, U_t(q), l(U_t;q), q) \approx U_0(q) \tag{4}$$

by minimizing

$$\mathbb{E}_{A_0\sim\mu_A,\, U_0\sim\mu_{U_0|A_0},\, U_1\sim\nu_U,\, t\sim\text{Unif}[0,1],q\sim\bar{\pi}_m}|D_\theta(t, A_0, U_t(q), l(U_t; q), q) - U_0(q)|^2.$$

The query distribution is the mixed law $\bar{\pi}_m = \lambda\pi + (1-\lambda)\pi_m$, $\lambda \in [0,1]$, where $\pi_m = \frac{1}{m}\sum_{i=1}^{m}\delta_{x_i}$ is the empirical uniform measure on X_m and π is a continuous query distribution on χ.

Generation. We first draw $A_1 \sim \nu_A$ and reverse the global diffusion to obtain $\hat{A}_0$. We then draw $U_1 \sim \nu_U$ and reverse the conditional residual diffusion according to (4), using $\hat{A}_0$ as the global conditioning state, to obtain $\hat{U}_0$. Finally,

$$\hat{U}_0 \leftarrow (I - \mathcal{R}S_L)\hat{U}_0, \quad F_{\text{gen}} = \mathcal{R}\hat{A}_0 + \hat{U}_0. \tag{5}$$

The projection guarantees $S_L F_{\text{gen}} = \hat{A}_0$. Both reverse chains are early stopped at $t_0 = \epsilon > 0$, and $\hat{A}_0, \hat{U}_0$ denote the stopped states A_{t_0}, U_{t_0}. Complete training and sampling procedures are given in Algorithms 1 and 2.

Global encoders. Two constructions give a fixed dimensional global state while allowing the residual denoiser to query arbitrary locations in functional diffusion (Appendix A.3): a *basis* encoder, $A_0 = (\langle F_0, \psi_i\rangle_{\mathcal{K}})_{i\le L}$ for $\{\psi_i\}_{i=1}^{L} \subset \mathcal{H}$ orthogonal in $\mathcal{K}$ with $\mathcal{R}a = \sum_{i\le L} a_i\psi_i$, recovered from $F_0(X_m)$ by least squares as a numerical encoder $\hat{S}_L$ satisfying $\hat{S}_L\mathcal{R} = I_L$; and an *anchor* encoder, $A_0 = F_0(X_k)$ on anchors $X_k \subset X_m$, $k \ll m$, with $\mathcal{R}$ the kernel interpolant through the anchors, so that $S_L\mathcal{R} = I_k$.

4 Theoretical Analysis

4.1 Information-Based Error Decomposition

Fix $t \in (0,1]$, $q \in \chi$, and radius $r > 0$. Recall that $N(q,r) = \{x_i \in X_m : 0 < \|x_i - q\| \le r\}$. Let

$$d^*_{\text{Bayes}} = \mathbb{E}[U_0(q) \mid A_0, U_t], \qquad d^*_{\text{full}} = \mathbb{E}[U_0(q) \mid A_0, U_t(q), U_t(X_m)], \tag{6}$$

$$d^*_r = \mathbb{E}[U_0(q) \mid A_0, U_t(q), U_t(N(q,r))], \qquad d^*_{\text{local}} = \mathbb{E}[U_0(q) \mid A_0, U_t(q), \ell(U_t; q)]. \tag{7}$$

Define the discretization, localization, compression and learning errors respectively as

$$\Delta_{\text{disc}} = \mathbb{E}[d^*_{\text{full}} - d^*_{\text{Bayes}}]^2, \Delta_{\text{loc}} = \mathbb{E}[d^*_r - d^*_{\text{full}}]^2$$

$$\Delta_{\text{comp}} = \mathbb{E}[d^*_{\text{local}} - d^*_r]^2, \Delta_{\text{learn}} = \mathbb{E}[D_\theta - d^*_{\text{local}}]^2.$$

Theorem 1. *If the local statistic is measurable with respect to $U_t(N(q,r))$, then*

$$\mathbb{E}[D_\theta - d^*_{\text{Bayes}}]^2 = \Delta_{disc} + \Delta_{loc} + \Delta_{comp} + \Delta_{learn}. \tag{8}$$

Moreover,

$$\underbrace{\mathbb{E}[D_\theta - U_0(q)]^2}_{\textit{training loss}} = \underbrace{\mathbb{E}[D_\theta - d^*_{\text{Bayes}}]^2}_{\textit{transport error}} + \underbrace{\mathbb{E}[d^*_{\text{Bayes}} - U_0(q)]^2}_{\textit{irreducible Bayes risk}}. \tag{9}$$

Hence population minimizers of the usual denoising loss also minimize the error relative to Bayes optimal denoiser that defines the exact reverse flow.

The four terms separate information lost by finite discretization, discarding distant context, compressing the retained neighborhood, and approximating the remaining conditional expectation with a learned model. The proof is an orthogonal projection telescoping argument.

4.2 Generalization Error Bound

Conditioning on $A_0 = a$, the exact residual probability reverse flow can be written

$$\partial_t U_t(q) = A(t)U_t(q) + B(t)d^*_{\text{Bayes}}(t,q), \quad A(t) = \frac{\dot{\sigma}_t}{\sigma_t}, \quad B(t) = \dot{\alpha}_t - \frac{\alpha_t\dot{\sigma}_t}{\sigma_t}. \tag{10}$$

In comparison, the learned flow is

$$\partial_t \hat{U}_t(q) = A(t)\hat{U}_t + B(t)D_\theta(t,q), \quad D_\theta(t,q) = D_\theta(t,a,U_t(q),l(U_t;q),q). \tag{11}$$

We assume that the local statistic $l(y(X_m);q)$, regarded as a function on χ in L^2 and on X_m in $\|\cdot\|_m$, is $C_{\mathcal{L}}$-Lipschitz with respect to the context values in $\|\cdot\|_m$, uniformly over the admissible designs. For the learned denoiser $D_\theta(t,a,u,l,q)$, we assume it is Lipschitz in its global state a, noisy value u and local statistic l with constants $L_A(t)$, $L_u(t)$, $L_l(t)$. See Assumptions 4 and 5 in Appendix C.

Theorem 2 (Conditional Generation Bound)**.** *Let $\mu_{\hat{U}|a}$ denote the law generated by the learned conditional residual flow stopped at $\epsilon > 0$, and let*

$$\eta^2_{mix}(t) = \mathbb{E}_{q\sim\bar{\pi}_m}|D_\theta(t,q) - d^*_{Bayes}(t,q)|^2.$$

Then

$$[\mathbb{E}_{A_0\sim\mu_A}\mathcal{W}_2^2(\mu_{U_0|A_0},\mu_{\hat{U}|A_0})]^{\frac{1}{2}} \le C_\lambda \int_\epsilon^1 g_\epsilon(t)\eta_{mix}(t)dt + \epsilon_{stop}, \tag{12}$$

where $g_\epsilon(t) = |B(t)|\exp(\int_\epsilon^t L_{ode}(s)ds)$, $L_{ode}(t) = |A(t)| + |B(t)|(L_u(t) + L_l(t)C_{\mathcal{L}})$ and $C_\lambda = [\max\{\frac{1}{\lambda}, \frac{|\chi|}{1-\lambda}\}]^{1/2}$.

Thus (12) controls the residual transport error by the denoising error under the mixed query law, which covers both the context grid, whose values evolve autonomously, and off-grid queries, which evolve conditionally on the context through l. Global generation error propagates through the residual stage; under our denoiser this dependence is controlled by the following stability bound.

Proposition 3. *For a learned conditional denoiser D_θ, there exists a constant $L_{gen(\theta)}$ such that, for all a, a', $\mathcal{W}_2(\mu_{\hat{U}|a},\mu_{\hat{U}|a'}) \le L_{gen(\theta)}\|a - a'\|$.*

We can then bound the total Wasserstein error by sum of global and conditional generation error.

Theorem 4. *Write $\mu_{\hat{A}} = \mathcal{L}(\hat{A}_0)$, $\mu_F = \mathcal{L}(F_0)$, $\mu_{\hat{F}} = \mathcal{L}(\hat{F}_0)$, then*

$$\mathcal{W}_2(\mu_F,\mu_{\hat{F}}) \le (\|\mathcal{R}\| + L_{gen(\theta)})\mathcal{W}_2(\mu_A,\mu_{\hat{A}}) + [\mathbb{E}_{A_0\sim\mu_A}\mathcal{W}_2^2(\mu_{U_0|A_0},\mu_{\hat{U}|A_0})]^{\frac{1}{2}}.$$

4.3 Locality And Effective Radius

The localization error Δ_{loc} measures the loss caused by discarding distant context. Let $I = \{q\} \cup N(q,r)$, $J = \{x_i \in X_m : |x_i - q| > r\}$ and write $C = (A_0, U_t(I))$. The conditional correlation ratio

$$\bar{\eta}^2(r) =: \frac{\mathbb{E}[\mathrm{Var}(\mathbb{E}[U_0(q) \mid U_t(J), C] \mid C)]}{\mathbb{E}[\mathrm{Var}(U_0(q) \mid C)]} \in [0,1]$$

measures the fraction of remaining uncertainty explained by the distant context after conditioning on the near field. Proposition 11 (Appendix C.1) shows $\mathbb{E}[d^*_{\text{full}} - d^*_r]^2 = \frac{\bar{\eta}^2(r)}{1-\bar{\eta}^2(r)}V_{\text{full}}$ with $V_{\text{full}} = \mathbb{E}[\mathrm{Var}(U_0(q) \mid A_0, U_t(q), U_t(X_m))]$, so the localization error vanishes exactly when $\bar{\eta}(r) \to 0$.

To obtain explicit rates, we assume that the residual has covariance decay and Gaussianity.

Assumption 1 (Covariance decay of residual)**.** *For $\mu_A - a.e.$ a and all $q, x \in \chi$*

$$|\mathrm{Cov}(U_0(q), U_0(x) \mid A_0)| \le \rho_0(\|q - x\|),$$

where $\rho_0(r)$ is non-increasing and $\int_0^\infty \rho_0(u)^2 u^{d-1}du < \infty$. The residual noise decays no slower, with $|C_U(q,x)| \lesssim \rho_0(\|q-x\|)$.

Assumption 2. *$U_0, U_t(q), U_t(X_m) \mid A_0 = a$ is jointly Gaussian for every relevant $a \in R^L$, $q \in \chi$ and $t > 0$.*

Under Gaussianity, the localization error Δ_{loc} has a closed form.

Theorem 5. *Under Assumption 2, write $Z_I = U_t(I), Z_J = U_t(J)$, $\Sigma = \mathrm{Cov}((Z_I, Z_J) \mid A_0)$, $\sigma_q = \mathrm{Cov}((Z_I, Z_J), U_0(q) \mid A_0)$, and partition them with respect to I, J, i.e., $\Sigma = \begin{bmatrix}\Sigma_{II} & \Sigma_{IJ}\\ \Sigma_{JI} & \Sigma_{JJ}\end{bmatrix}$, $\sigma_q = \begin{bmatrix}\sigma_I\\ \sigma_J\end{bmatrix}$, then*

$$\mathbb{E}[d^*_{full} - d^*_r \mid A_0]^2 = w^T S^{-1} w, \quad w = \sigma_J - \Sigma_{JI}\Sigma_{II}^{-1}\sigma_I, \quad S = \Sigma_{JJ} - \Sigma_{JI}\Sigma_{II}^{-1}\Sigma_{IJ}. \tag{13}$$

Here $\bar{u} = \Sigma_{II}^{-1}\sigma_I$ is the kriging vector, the coefficients of the optimal linear predictor of $U_0(q)$ from the near field. Covariance decay alone does not make w small: a far point $x \in J$ just outside $B(q,r)$ and a near point $y \in I$ just inside can be arbitrarily close, so Σ_{xy} may be large, and the second term of $w_x = \sigma_x - \Sigma_{xI}\bar{u}$ is controlled only if the kriging weights near the boundary are small. We therefore impose the screening condition (Stein, 2002).

Assumption 3 (Localized kriging weights)**.** *The total mass of the near-field kriging weights at distance at least s from q decays at least as fast as the covariance envelope,*

$$T(s) = \sum_{y\in I:\|y-q\|>s} |\bar{u}_y| \le \Lambda|\alpha_t|\rho_0(s)/c_0, \quad 0 < s \le r, \tag{14}$$

where $c_0 = \rho_0(0)$.

Theorem 6. *Suppose X_m is quasi-uniform in χ with m points. If Assumption 1 holds with exponential decay $\rho_0(u) = c_0 e^{-u/l_c}$, then*

$$\mathbb{E}[d^*_{full} - d^*_r]^2 \lesssim \frac{\alpha_t^2(1+\Lambda)^2}{\sigma_t^2} \cdot \frac{c_0^2 m l_c r^{d-1}}{\lambda_{\min}(C_U)|\chi|} e^{-r/l_c}.$$

A polynomial envelope $\rho_0(u) = c_0(1+u/l_c)^{-a}$ with $a > d/2$ gives, by the same argument, $\mathbb{E}[d_{\text{full}} - d^*_r]^2 \lesssim \frac{\alpha_t^2}{\sigma_t^2}\frac{m}{|\chi|} r^{d-2a}$ (Proposition 12, Appendix C).

4.4 A MODEL CLASS IN WHICH THE HYPOTHESES HOLD

Theorem 6 assumes conditional joint Gaussianity, which is severe as a statement about μ_0 while mild as one about $\mu_{U_0|A_0}$. In fact, our proposed hypotheses hold naturally under the Hierarchical Gaussian Process mixture model setting.

Proposition 7 (Hierarchical Gaussian Process Mixtures)**.** *Let μ_A be an arbitrary Borel probability measure on R^L, and suppose $U_0 \mid A_0 = a \sim \mathcal{GP}(H_L(a), C_a(x,x'))$, where C_a is the conditional data covariance kernel on V_U, so that $\mu_0 = \int \mathcal{N}(\mathcal{R}a + H_L(a), C_a)\mu_A(da)$ is a continuous mixture of Gaussian measures. Then: (i) for every a and every $t > 0$, $U_0(q), U_t(X_m) \mid A_0 = a$ are jointly Gaussian, so Assumption 2 hold exactly; (ii) μ_0 is in general neither Gaussian nor unimodal, and can be flexible enough as μ_A varies.*

The law μ_A on $\mathbb{R}^L$ carries the multimodality of the data. Gaussianity is required only at the fine scale after the global structure has been conditioned away, and $C_a(x,x')$ depending on a lets the local structure differ across modes: the theory needs conditional locality, not global Gaussianity.

4.5 COMPRESSION ERROR OF LOCAL STATISTICS

Take $h = r$ in implementation so that the local statistic uses all context points in $N = N(q,r)$, and let $n_h = |N|$. Conditioning on A_0 and under Assumption 2, the radius-r predictor is linear, $d^*_r = w_q U_t(q) + w_N^T U_t(N)$, with kriging weights (w_q, w_N) given in closed form in Appendix C.3 (Proposition 13). Let $\Sigma_{N|q} = \mathrm{Cov}(U_t(N) \mid U_t(q), A_0)$, and for $\varphi \in \mathcal{P}_p$, let $W_h\varphi = (W_h(x_i - q)\varphi(x_i - q))_{i\in N}$. The statistic $l(U_t; q)$ is a fixed linear map of $U_t(N)$ whose transpose has range $\{W_h\varphi : \varphi \in \mathcal{P}_p\}$ (Appendix C.3), so the compression error is the distance from the kriging weights to that range in the energy norm of $\Sigma_{N|q}$.

Proposition 8 (Compression error)**.** *Under joint Gaussianity, the compression error*

$$\Delta_{comp} = \min_{\varphi\in\mathcal{P}_p} \|w_N - W_h\varphi\|^2_{\Sigma_{N|q}}, \tag{15}$$

where $W_h\varphi = (W_h(x_i - h)\varphi(x_i - q))_{i\in N}$ and $\|v\|^2_{\Sigma_{N|q}} = v^T\Sigma_{N|q}v$ for any $v \in R^{n_h}$.

Proposition 9 (Compression rate)**.** *Suppose that, conditionally on A_0, the design residual of the local fit is quasi-optimal, $\|U_t(N) - Vl(U_t; q)\|_2^2 \le C_{\text{qo}} \inf_{\varphi\in\mathcal{P}_p} \sum_{x_i\in N} |U_t(x_i) - \varphi(x_i)|^2$, that N satisfies a sampling inequality on $B(q,h)$ with constant C_{samp}, and that U_t restricted to $B(q,h)$ is a stationary field with spectral density $\widehat{C}_t$ and finite spectral moment $S_k(t) = \int_{\mathbb{R}^d} |\xi|^{2k}\widehat{C}_t(\xi)\, d\xi$ for some $k \le \min\{p+1, s\}$. Then*

$$\Delta_{\text{comp}} \le C_{d,k}\, \|w_N\|_2^2\, n_h\, h^{2k}\, S_k(t),$$

with $C_{d,k}$ depending on d, k, C_{qo} and C_{samp} only; $\|w_N\|_2$ and $S_k(t)$ depend on A_0 and stand for their suprema over a.

End-to-end Bound. Combining the results above gives the following end-to-end bound.

Theorem 10. *Under Assumptions 1–8 with exponential decay and $r \geq l_c$, for any $\delta > 0$,*

$$\mathbb{E}\,\mathcal{W}_2(\mu_F, \mu_{\hat{F}}) \lesssim N^{-\frac{\theta+1-\delta}{2(2\theta+L)}} + N_*^{-\frac{\theta'}{2\theta'+L+D_p+d+1}} + N^{-\frac{\theta'}{2\theta'+L}} + m^{-\frac{s}{d}} + r^{\frac{d-1}{2}} e^{-\frac{r}{2l_c}} + r^{k+\frac{d}{2}} + \epsilon_{\text{stop}},$$

where N is the number of i.i.d. function samples and $N_* = N \min\{|\chi|/(r+l_c)^d, |Q|\}$. The precise form is in Appendix D.10.

5 Experiment

5.1 Controlled Gaussian Process Mixture Diffusion

We build a synthetic Gaussian process mixture distribution in which every information-based denoiser has a closed form. On $\chi = [0,1]^2$, we use a real tensor Fourier basis with $M = 29^2 = 841$ modes, the $L = 25$ lowest-frequency modes for global component and the remaining $K = 816$ for residual. The global state $A_0 \in R^{25}$ follows a four component Student-t mixture. Conditional on a, the residual coefficients c_0 are Gaussian with a fixed nonlinear mean $H_L(a)$ and a covariance $\Sigma_0(a)$, which is obtained by projecting one of the four nonstationary anisotropic Matérn kernels of Paciorek–Schervish type selected by a onto the residual Fourier subspace, each with distinct smoothness and spatial correlation. By construction, $U_0(q) = \psi_U(q)^T c_0$. Forward residual noise is a fixed stationary isotropic Matérn-$3/2$ field projected onto the same subspace, deliberately mismatched with the data covariance. For fixed (a, t, q), we compute the conditional variances of $d^*_{\text{Bayes}}, d^*_{\text{full}}, d^*_r, d^*_{\text{local}}$, denoted as $V_{\text{Bayes}}, V_{\text{m}}, V_{\text{r}}, V_{\text{loc}}$ respectively, and then $\Delta_{\text{disc}} = V_m - V_{\text{Bayes}}$, $\Delta_{\text{loc}} = V_r - V_m$, $\Delta_{\text{comp}} = V_{\text{loc}} - V_r$. We also include the query-only denoiser $d^*_{\text{query}} = \mathbb{E}[U_0(q) \mid A_0, U_t(q)]$, and denote V_{query} its variance.

Approximation Behavior. Figure 1 shows the qualitative behavior of three approximation errors against the quantities that control them. Discretization error falls under refinement and vanishes once the design resolves the residual band ($m = K$); localization error decays in r in both the finite Fourier space and the spatial domain construction; the compression error stays below the local gain $V_{\text{query}} - V_r$ at every scale, and the degree-p statistic retains a substantial fraction of the neighborhood's predictive information rather than losing it as the scale shrinks. The same ordering holds in every covariance mode, including the rough and strongly anisotropic ones.

Generation behavior. Figure 2 runs the reverse chain under each information set from a shared terminal noise. The full context and radius-r chains are nearly indistinguishable from the exact Bayes chain, and the compressed chain (ours) stays within a small factor of them ($\mathcal{D} = 0.004$ against 0.001 in mode 0, 0.236 against 0.171 in mode 3), whereas dropping the neighborhood $U_t(N)$ costs an order of magnitude ($\mathcal{D} = 0.030$ and 0.574). The trained neural network denoiser lands between the compressed and the query-only chains in every mode, so the error budget of Theorem 2 is dominated by Δ_{learn} rather than by localization or compression.

Appendix B.1 gives the full benchmark and additional experimental results, including validation of the screening and weight-growth assumptions (Figures 6 and 10), the information ladder and its time dependence (Figures 7 and 11), robustness to the covariance parameters (Figure 3), the role of the polynomial degree (Figure 4), and the denoiser hierarchy at fixed t (Figures 5, 8 and 9).

5.2 Functional Diffusion on Natural Image Generation

The image experiment is primarily diagnostic: pixel values on a fixed grid do not meaningfully test resolution independence. Instead, we use it to construct empirical counterparts of the main terms in Theorem 4, identify which stage contributes most to the error, and test whether the corresponding theory motivated remedies are effective (details and samples in Appendix B.2).

Model and diagnostic. On CIFAR-10 (Krizhevsky et al., 2009), we use a two-level `db4` wavelet transform hierarchy: the global state A_0 is the $8 \times 8 \times 3$ low-pass block ($L = 192$), $G_0 = \mathcal{R}A_0$ its reconstruction on the pixel grid, and $U_0 = F_0 - G_0$. The global stage is a diffusion model on the normalized A_0, a U-Net on the 8×8 block with 250 DDIM steps (Song et al., 2020)). The residual stage uses the denoiser of Algorithm 2, with either a *full context* trunk (a U-Net with attention over

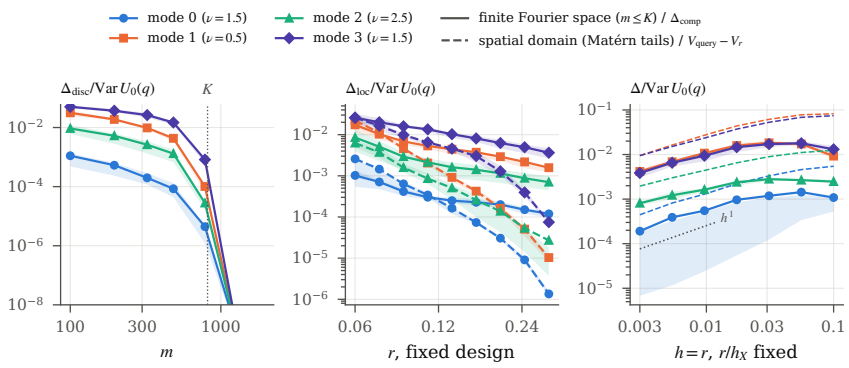


Figure 1: **Approximation errors across four covariance modes.** Discretization error decreases under design refinement and vanishes at the finite-rank saturation point $m = K$; localization error decreases as the radius r grows, both in the finite Fourier space (solid) and spatial domain (dashed); compression error remains below the available local gain $V_{\text{query}} - V_r$ as the local scale varies.

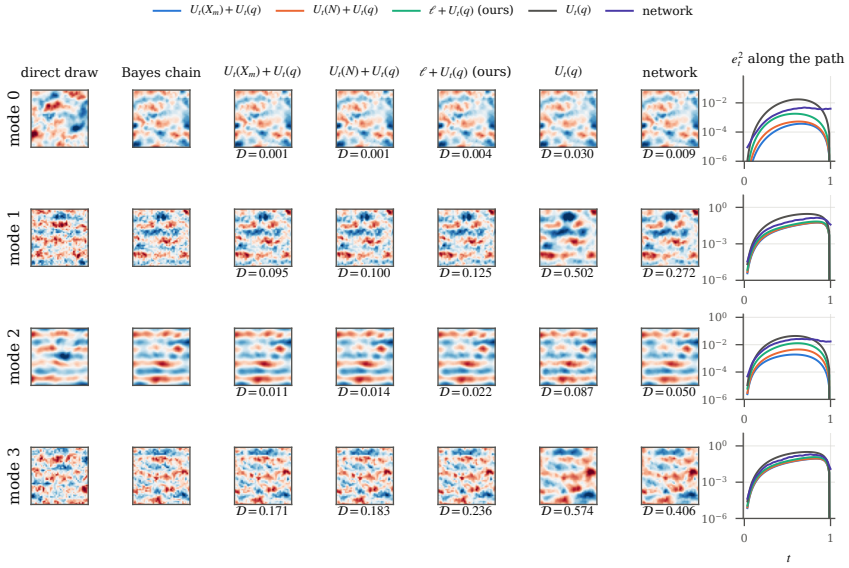


Figure 2: **Residual generation across four covariance modes.** We compare exact Bayes flow with flow using d^*_{full}, d^*_r, d^*_{local}, single query $U_t(q)$, and learned denoiser by reversing from same noise U_1; $\mathcal{D}$ measures the final discrepancy from Bayes flow and the right panels show instantaneous denoising error e_t^2 along Bayes path.

the whole field, the role of d^*_{full}) or a *local* trunk of r stride-one 3×3 convolutions with per-pixel normalization, whose receptive field is $N(q, r)$ exactly. In both cases the head receives $U_t(q)$, the fixed degree-2 local polynomial statistic (leave-one-out, three window scales, no parameters) and the learned trunk features at q, so that a radius-r trunk has the information set of d^*_r and $r = 0$ is that of Algorithm 1 exactly. The residual stage is trained with *conditioning augmentation*: it receives $\mathcal{R}(A_0 + \sigma\,\mathrm{sd}(A_0)\xi)$ with $\sigma \sim \mathrm{Unif}[0, \frac{1}{2}]$ while the target remains U_0, which is meant to enhance the robustness of the residual denoiser to global error and reduce $L_{\text{gen}(\theta)}$. For each sample index we draw one residual terminal noise and use it in two runs differing only in the global state: *gen* uses generated $\hat{A}_0$ while *oracle* uses the true A_0 of a held-out test image. We use $\mathrm{FID}(\text{gen}) - \mathrm{FID}(\text{oracle})$ as an empirical diagnostic of the global error, corresponding to $(\|\mathcal{R}\| + L_{\text{gen}(\theta)})\mathcal{W}_2(\mu_A, \mu_{\hat{A}})$, and $\mathrm{FID}(\text{oracle}) - \mathrm{FID}(\text{real})$ as the diagnostic of the residual error given the true global state. We also report FD_A, the Fréchet distance between generated and real global states (Gelbrich, 1990), as

Table 1: **CIFAR-10, FID-10k, gen / oracle.** The oracle arm uses the true global state of a held-out image; the simultaneous oracle follows its noisy trajectory. Top: model comparison. Bottom: gen and oracle FID of residual denoisers that see the whole field (full) or a window of radius r, with the base global stage.

model	gen	oracle
hierarchical, base	40.7	14.3
hierarchical, reg. global + aug	27.8	**12.3**
hierarchical, class-cond. + aug	**23.0**	**12.3**
simultaneous (one chain)	61.5	64.4
monolithic (no hierarchy)	21.5	–
monolithic, local $r = 4$ trunk	126.8	–

residual trunk	full	$r = 8$	$r = 4$	$r = 0$
oracle (true A_0)	**17.9**	19.6	24.3	49.0
gen (end-to-end)	35.2	**32.6**	38.4	73.5

$\sqrt{\mathrm{FD}_A}$ in percent of the RMS norm of A_0. FIDs use $10{,}000$ samples (Heusel et al., 2017); differences below one point are noise and real test images score 3.15.

Results. Table 1 gives four findings. (i) L_{gen} *is real and reducible.* With the global stage fixed, conditioning augmentation reduces the gap from 27.4 to 18.5 (Table 2), a smaller $L_{\mathrm{gen}(\theta)}$ since $\mathcal{W}_2(\mu_A, \mu_{\hat{A}})$ is unchanged. (ii) *The global stage is limited by memorization.* Its held-out ε-MSE is 1.7 times the training value. Regularization lowers the thumbnail FID from 6.7 to 3.3 and the end-to-end FID from 30.8 to 27.8, and class conditioning to 23.0. (iii) *Locality costs little and reduces global-state sensitivity.* Given the true A_0, $r = 8$ and $r = 4$ are 1.7 and 6.4 FID behind the full-context model, yet end to end $r = 8$ is better (32.6 vs 35.2): a local window cannot propagate a global inconsistency of $\hat{A}_0$. At $r = 0$ (49.0) the degree-2 statistic alone is insufficient on natural images, unlike on the distance fields of Section 5.3. (iv) *The hierarchy makes the residual problem easier; the global stage limits end-to-end quality.* The simultaneous model (61.5 vs 35.2) and the local $r = 4$ trunk without a global state (126.8 vs 38.4) are far worse, while the monolithic full-context model (21.5) is better end to end but worse than the residual stage given the true A_0 (12.3): the hierarchy loses only through its global stage.

5.3 Signed Distance Function Generation

To test the algorithm with a fixed local polynomial statistic, we generate the signed distance functions of MNIST and Fashion-MNIST shapes (Appendix B.3). The global state is the first 64 DCT modes, and the residual denoiser sees the noisy field only through the six coefficients of the local degree-2 polynomial regression; a full-context baseline with the same global state reads the whole residual field. With $20\times$ fewer parameters, the local model has a lower held-out denoising error at nearly every noise level (Table 5), and its generated samples are valid signed distance functions whose eikonal error and topology are close to the data's (Figure 13). On a 64^2 grid the local model keeps its size and has the lower eikonal error, while the baseline triples in size.

6 Conclusion

We proposed a hierarchical functional diffusion model in which a moderate dimensional global state is generated first and a conditional functional diffusion generates the residual from the query value and a fixed dimensional local statistic. The information-based decomposition separates the residual error into discretization, localization, compression and learning errors, and these terms propagate through the reverse flow into the Wasserstein bound on the generated distribution. The Gaussian process mixture benchmark measures every term in closed form. In real applications, the fixed local statistic fits smooth fields such as distance functions well, while natural images need a learned window and a better global stage; for images, higher-resolution and implicit neural representation studies are the natural next step.

AI DISCLOSURE

We used large language models (LLMs) to improve the clarity and presentation of the manuscript, assist with experimental code development and debugging, and help analyze experimental results. All AI-assisted work was reviewed and verified by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 4 and Appendix C, and all proofs are given in Appendix D. Algorithms 1 and 2 give the training and sampling procedures, and Appendices B.1, B.2 and B.3 describe the data, models, training and evaluation of each experiment. Code to reproduce all results will be released upon publication.

REFERENCES

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

Giulio Franzese, Giulio Corallo, Simone Rossi, Markus Heinonen, Maurizio Filippone, and Pietro Michiardi. Continuous-time functional diffusion processes. *Advances in Neural Information Processing Systems*, 36:37370–37400, 2023.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Georg A. Gottwald, Shuigen Liu, Youssef Marzouk, Sebastian Reich, and Xin T. Tong. Localized diffusion models, 2025. URL https://arxiv.org/abs/2505.04417.

Paul Hagemann, Sophie Mildenberger, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. *SIAM Journal on Mathematics of Data Science*, 7(3):1337–1366, 2025.

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

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video Diffusion Models. In *Neural Information Processing Systems*, 2022b. URL https://mlanthology.org/neurips/2022/ho2022neurips-video/.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9538–9563. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/kerrigan23a.html.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985. PMLR, 2023.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025.

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.

Christopher J Paciorek and Mark J Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Michael L Stein. The screening effect in kriging. *The Annals of Statistics*, 30(1):298–323, 2002.

Biao Zhang and Peter Wonka. Functional diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4723–4732. IEEE, 2024.

A Additional Methodological Details

A.1 Training and Sampling Algorithms

Algorithms 1 and 2 give the complete training and sampling procedures of Section 3. The two stages of training share no data path: the residual denoiser is trained on the true global A_0, not on the output of the global stage.

Algorithm 1 Training

Require: Context set X_m; query access to F_0

Fix the numerical encoder $\hat{S}_L$ on X_m: least squares (basis) or evaluation on $X_k \subset X_m$ (anchor)

Stage 1: Global denoising.

repeat

  Sample F_0; $A_0 \leftarrow \hat{S}_L F_0(X_m)$

  Sample $A_1 \sim \nu_A$, $t \sim \mathrm{Unif}[0,1]$; $A_t \leftarrow \alpha_t A_0 + \sigma_t A_1$

  Gradient step on $\|D_\phi(A_t, t) - A_0\|^2$

until convergence

Stage 2: Residual denoising.

repeat

  Sample F_0; $A_0 \leftarrow \hat{S}_L F_0(X_m)$; $U_0(X_m) \leftarrow F_0(X_m) - (\mathcal{R}A_0)(X_m)$

  Sample queries $Q = \{q_j\}_{j=1}^{|Q|}$, $q_j \overset{\text{i.i.d.}}{\sim} \bar{\pi}_m$, and evaluate $U_0(q) = F_0(q) - (\mathcal{R}A_0)(q)$, $q \in Q$

  Sample $U_1 \sim \nu_U$ and evaluate it on $X_m \cup Q$; sample $t \sim \mathrm{Unif}[0,1]$ and set $U_t \leftarrow \alpha_t U_0 + \sigma_t U_1$ on $X_m \cup Q$

  Gradient step on

$$\frac{1}{|Q|} \sum_{q \in Q} |D_\theta(t, A_0, U_t(q), \ell(U_t; q), q) - U_0(q)|^2,$$

  with $l(U_t; q)$ formed from $U_t(X_m)$

until convergence

Algorithm 2 Sampling

Require: Queries $\mathcal{Q} \subseteq \chi$ (arbitrary), time grid $1 = t_N > \cdots > t_1 > t_0 = \epsilon$

Sample $A_{t_N} \sim \nu_A$

for $j = N, \ldots, 1$ **do**

  $A_{t_{j-1}} \leftarrow \alpha_{t_{j-1}} D_\phi(A_{t_j}, t_j) + \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}} \big[A_{t_j} - \alpha_{t_j} D_\phi(A_{t_j}, t_j)\big]$

end for

$\hat{A}_0 \leftarrow A_{t_0}$

Sample U_{t_N} on $\mathcal{Q} \cup X_m$ from ν_U, independent of A_{t_N}

for $j = N, \ldots, 1$ **do**

  $l_j(\cdot) \leftarrow l(U_t; \cdot)$ from $U_{t_j}(X_m)$

  for all $z \in X_m \cup \mathcal{Q}$ **do**

    $\widehat{D}_z \leftarrow D_\theta\big(t_j, \hat{A}_0, U_{t_j}(z), l(U_{t_j}; z), z\big)$

    $U_{t_{j-1}}(z) \leftarrow \alpha_{t_{j-1}} \widehat{D}_z + \frac{\sigma_{t_{j-1}}}{\sigma_{t_j}} \big[U_{t_j}(z) - \alpha_{t_j} \widehat{D}_z\big]$

  end for

end for

$\hat{U}_0 \leftarrow (I - \mathcal{R}S_L) U_{t_0}$

return $\mathcal{R}\hat{A}_0 + \hat{U}_0$ on $\mathcal{Q}$

Early stopping. Both reverse chains stop at $t_0 = \epsilon > 0$, so $\hat{A}_0$ and $\hat{U}_0$ denote the stopped states A_{t_0} and U_{t_0}, not the states at $t = 0$. The residual bound of Theorem 2 compares the law of the stopped residual reverse chain with the data law and includes the stopping error $\epsilon_{\text{stop}} = (1 - \alpha_\epsilon)[\mathbb{E}\|U_0\|^2_{L^2(\pi)}]^{1/2} + \sigma_\epsilon[\mathbb{E}\|U_1\|^2_{L^2(\pi)}]^{1/2}$.

Projection. Algorithm 2 projects once at the end. The image implementation (Appendix B.2) instead applies $(I - \mathcal{R}S_L)$ to the x_0-prediction $\widehat{D}$ at every step, which keeps the chain in V_U throughout. When ν_U is supported on V_U and D_θ maps into V_U, the two coincide. Otherwise, the per-step projection removes the component of the network output outside V_U before it is propagated.

A.2 Selection of global and residual diffusion noises

Nothing in Section 4 requires ν_A, ν_U to be pushforwards of a common μ_1. It suffices that each stage admits a prescribed Gaussian terminal law, independent of its corresponding target. The shared construction makes $\mathcal{R}A_t + U_t = F_t$ hold for every t, which is what lets the hierarchical and simultaneous denoisers be compared on the same noisy function. Also note that A_1 and U_1 should be drawn independently at training and generation, even though 2 splits a single F_1.

A.3 Global Encoders

This section gives the two encoders of Section 3 in full. The basis-based encoder is the one used in the experiments, and the anchor-based encoder reads the global state off a subset of the context points.

Basis-based encoder. Let $\{\psi_i\}_{i\le L} \subset \mathcal{H}$ be orthonormal in $\mathcal{K} = L^2(\chi)$. Take $A_0 = (\langle F_0, \psi_i\rangle_{\mathcal{K}})_{i\le L}$ and $\mathcal{R}a = \sum_{i\le L} a_i\psi_i$, so $S_L\mathcal{R} = I_L$. In practice the coefficients are recovered from $F_0(X_m)$ by least squares, which gives a numerical encoder $\hat{S}_L$ with $\hat{S}_L\mathcal{R} = I_L$.

Anchor-based encoder. Select anchors $X_k \subset X_m$, $k \ll m$, and take $A_0 = F_0(X_k)$, so that $L = k$. Since $\mathcal{H}$ is an RKHS with kernel $K(\cdot,\cdot)$, the minimum-norm interpolant of the anchor values is

$$\mathcal{R}(a)(x) = \sum_{i=1}^{k} w_i(x)a_i, \quad w_i(x) = \sum_{j=1}^{k} K(x, x_j)(\mathbf{K}^{-1})_{ji}, \tag{16}$$

where $\mathbf{K} = (K(x_i, x_j))_{i,j\le k}$ is the Gram matrix of the anchors. The interpolant passes through the anchor values, so $S_L\mathcal{R} = I_k$.

B Additional Experimental Results and Details

B.1 Controlled Gaussian Process Mixture Diffusion

B.1.1 Benchmark

We use the real tensor Fourier basis on $[0,1]^2$, generated from

$$\phi_0(x) = 1, \qquad \phi_{k,c}(x) = \sqrt{2}\cos(2\pi kx), \qquad \phi_{k,s}(x) = \sqrt{2}\sin(2\pi kx), \quad 1 \le k \le k_{\max},$$

via $\psi_{k_1,k_2}(x_1,x_2) = \phi_{k_1}(x_1)\phi_{k_2}(x_2)$. With $k_{\max} = 14$, this gives $M = 29^2 = 841$ basis functions. The first $L = 25$ bases form the global subspace and the remaining $K = M - L = 816$ bases form the residual subspace. The global state follows a four-component multivariate Student-t mixture with weights $(0.34, 0.28, 0.22, 0.16)$ and three degrees of freedom. We refer to the four mixture components as modes $z \in \{0,1,2,3\}$. Each mode z has mean $\mu_z = 3.2d_z$ for a fixed unit direction $d_z \in R^{25}$ and scale

$$S_z = 0.45(A_zA_z^\top + 0.35I), \qquad A_z(i,j) \sim \mathcal{N}(0, 1/L).$$

A draw is

$$A_0 = \mu_z + \frac{L_z\varepsilon}{\sqrt{G}}, \qquad L_zL_z^\top = S_z, \quad \varepsilon \sim \mathcal{N}(0,I), \quad G = \chi^2_3/3.$$

The conditional residual coefficients have mean

$$H(a) = \tanh(aW_1 + b)W_2 \in \mathbb{R}^{816},$$

where $W_1 \in \mathbb{R}^{25\times 48}$, $b_1 \in \mathbb{R}^{48}$, $W_2 \in \mathbb{R}^{48\times 816}$ are fixed throughout the experiment. We then define $H_L(a)(x) = \psi_U(x)^\top H(a)$ as the conditional mean field in the residual subspace.

Each global state a is assigned to its nearest mixture mode,

$$z(a) = \arg\min_{z\in\{0,1,2,3\}} \|a - \mu_z\|_2.$$

The four base covariance regimes associated with these modes are:

mode	ℓ_0	anisotropy	spatial freq.	phase	ν
0	0.085	1.0	1.0	0	$3/2$
1	0.055	2.5	1.5	$\pi/3$	$1/2$
2	0.130	1.6	0.5	$2\pi/3$	$5/2$
3	0.070	3.5	2.0	π	$3/2$

Let $s_1(a)$ and $s_2(a)$ denote the projectionsm of a onto two fixed directions in the global state space. We perturb the mode-specific base parameters by

$$\ell_0(a) = \ell_0 e^{0.25 s_1(a)}, \qquad \mathrm{aniso}(a) = \mathrm{aniso}_0\, e^{0.20 s_1(a) s_2(a)}, \qquad \phi(a) = \phi_0 + 0.6 s_2(a).$$

For $x = (x_1, x_2)$, define

$$s_a(x) = \sin(2\pi f x_1 + \phi(a)) \cos(2\pi f x_2 - \phi(a)),$$

$$\ell_1(x; a) = \ell_0(a) e^{0.45 s_a(x)}, \qquad \ell_2(x; a) = \frac{\ell_1(x; a)}{\mathrm{aniso}(a)}, \qquad \vartheta(x; a) = 0.9\pi s_a(x).$$

With R_ϑ the planar rotation matrix,

$$\Sigma_a(x) = R_{\vartheta(x;a)}\, \mathrm{diag}\big(\ell_1(x; a)^2, \ell_2(x; a)^2\big) R_{\vartheta(x;a)}^\top.$$

The Paciorek–Schervish covariance is

$$C_a(x, y) = \frac{|\Sigma_a(x)|^{1/4} |\Sigma_a(y)|^{1/4}}{|\frac{1}{2}(\Sigma_a(x) + \Sigma_a(y))|^{1/2}} M_{\nu_z}\left(\sqrt{(x-y)^\top [\tfrac{1}{2}(\Sigma_a(x) + \Sigma_a(y))]^{-1} (x-y)} \right).$$

We use the closed-form Matérn kernels $M_{1/2}(r) = e^{-r}$, $M_{3/2}(r) = (1 + \sqrt{3}r)e^{-\sqrt{3}r}$, and $M_{5/2}(r) = (1 + \sqrt{5}r + 5r^2/3)e^{-\sqrt{5}r}$.

The kernel is projected onto the residual Fourier subspace using quadrature, and the resulting coefficient covariance is trace normalized. Consequently, the conditional spatial covariance used in the experiment is the finite-rank projection $\psi_U(x)^\top \Sigma_0(a) \psi_U(y)$, rather than the unprojected Paciorek Schervish kernel. This distinction matters for the finite-rank saturation visible in Figure 1.

Forward residual noise is the projection of a stationary isotropic Matérn-$3/2$ covariance kernel, which is trace normalized as well. We build the bridge as

$$c_t = \alpha_t c_0 + \sigma_t c_1, \qquad \Sigma_t(a) = \alpha_t^2 \Sigma_0(a) + \sigma_t^2 \Sigma_1.$$

For any linear observation $Z = P c_t$, the affine Bayes predictor of $U_0(q) = \psi_U(q)^T c_0$ is

$$d_P^*(q) = m_q(a) + \alpha_t \psi_U(q)^T \Sigma_0(a) P^T (P \Sigma_t(a) P^T)^\dagger \{Z - \alpha_t P H(a)\}.$$

Choosing P as the full coefficient identity, the sampled design, the radius-r neighborhood, or the compressed local observation gives d^*_{Bayes}, d^*_{full}, d^*_r, and d^*_{local} exactly. All denoising errors except trained neural network denoiser are therefore accurate in closed form.

B.1.2 Additional Results

Assumption Validation. Figure 10 tests Assumption 3 directly. The tail mass $T(s)$ of the kriging coefficients decays at least as fast as the covariance envelope $\rho_0(s)$ in every mode, and the resulting screening constant Λ stays of order $O(1)$, growing only slowly with n_N under refinement. The rightmost panel shows why: the scaled weights $n_N w_i$ collapse onto a fixed radial shape that is negligible beyond half the radius, so the outer points of N carry little weight whatever the design is. Figure 6 separates the two norms of w_N that appear in Proposition 8: the Euclidean norm $n_N \|w_N\|_2^2$ grows under refinement because neighboring weights alternate in sign, but the energy norm $\|w_N\|^2_{\Sigma_{N|q}}$ saturates, which is what the compression bound uses, and then the compression error Δ_{comp} saturates with it.

The information ladder and its time dependence. Figure 7 plots the conditional variances $V_{\text{Bayes}} \le V_m \le V_r \le V_{\text{loc}} \le V_{\text{query}}$ against t and confirms the ordering of Theorem 1 at every time, with the gaps widest at intermediate t and for modes 1 and 3. Figure 11 shows the corresponding error terms: Δ_{loc} and Δ_{comp} peak at intermediate t and vanish at both ends, Λ decreases as $t \to 1$, so the factor $(1 + \Lambda)^2$ of Theorem 6 is harmless at high noise, and the compressed local statistic retains a substantial fraction of the Bayes headroom at every t, least for the strongly anisotropic mode 3.

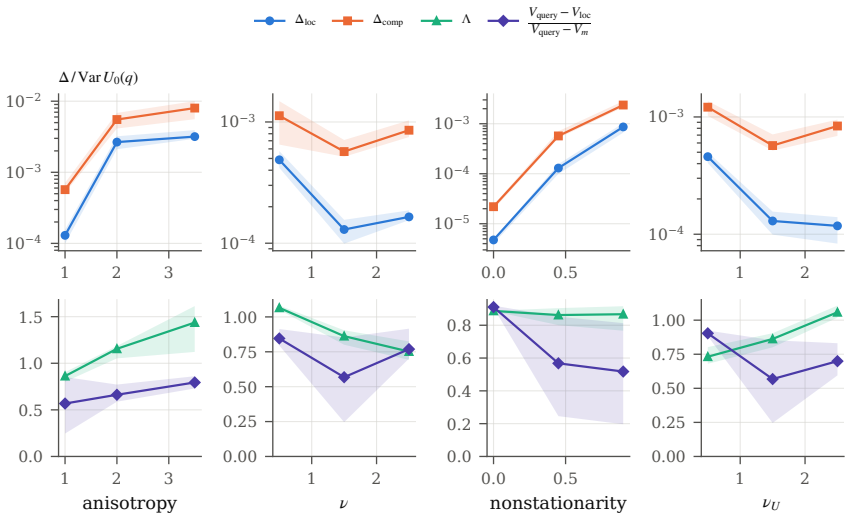


Figure 3: **Sensitivity of localization and compression to the data and noise covariance parameters.** Each column varies one factor in anisotropy, residual smoothness ν, nonstationarity, or noise smoothness ν_U, while keeping all other parameters fixed at the reference configuration. The top row shows normalized localization and compression errors, and the bottom row shows the screening constant Λ and the retained fraction of local predictive information. The approximation behavior remains stable across a broad range of perturbations.

Robustness and the polynomial degree. Figure 3 varies one covariance parameter at a time around the reference configuration: anisotropy, residual smoothness ν, nonstationarity amplitude, and noise smoothness ν_U. Localization and compression errors grow with anisotropy and roughness, as the constants in Theorem 6 and Proposition 8 predict, but Λ remains $O(1)$ and the retained fraction of local predictive information stays above one half throughout. Figure 4 varies the local polynomial regression degree p. Raising p from 0 to 1 changes little: the two curves nearly coincide in every mode, because the linear term learns the local slope, which is an odd function of $x - q$ and carries almost no information about the value at the centre of a nearly symmetric neighborhood. Raising p from 1 to 2 adds the curvature terms and lowers κ_p substantially, with the largest gain in the smooth, nearly isotropic mode 2. The same figure shows that on a fixed design the compression error is not monotone in r, since r influences both the neighborhood and the bandwidth of ℓ, so a rate in $h = r$ is only meaningful under refinement with r/h_X fixed.

Denoiser hierarchy at fixed time. Figure 5 compares seven denoisers at the same (a, t, q), from the prior mean to the exact Bayes rule. The compressed statistic $\ell + U_t(q)$ closes most of the gap between query-only conditioning and the raw neighborhood, while ℓ alone is worse than $U_t(q)$ alone. This indicates that the query value is the dominant observation and the neighborhood refines it. Figure 8 shows where the remaining errors sit: both Δ_{loc} and Δ_{comp} follow the local length-scale field of each nonstationary mode rather than the design grid, and are largest where the correlation length is short relative to r. Figure 9 compares one realization per mode at $t = 0.6$; the compressed local prediction is visually indistinguishable from the full context one, with a relative MSE within 0.011 in modes 0–2 and 0.05 above it in the strongly anisotropic mode 3.

B.2 Functional Diffusion on Natural Image Generation

Data and hierarchy. CIFAR-10 training images (50,000, $32 \times 32 \times 3$, scaled to $[-1, 1]$, with horizontal-flip augmentation) are treated as functions F_0 on the pixel grid X_m, $m = 1024$. The encoder S_L is a two-level orthogonal `db4` wavelet transform with periodic extension. The global state A_0 is the $8 \times 8 \times 3$ low-pass block ($L = 192$), standardized per coefficient, and $\mathcal{R}$ reconstructs from this block alone, so $S_L \mathcal{R} = I_L$ and $U_0 = (I - \mathcal{R} S_L) F_0$ is exact. The forward noise is Gaussian in the wavelet domain, with each detail band scaled to the empirical standard deviation of that band. Its global and residual blocks define ν_A and ν_U respectively, which are sampled independently.

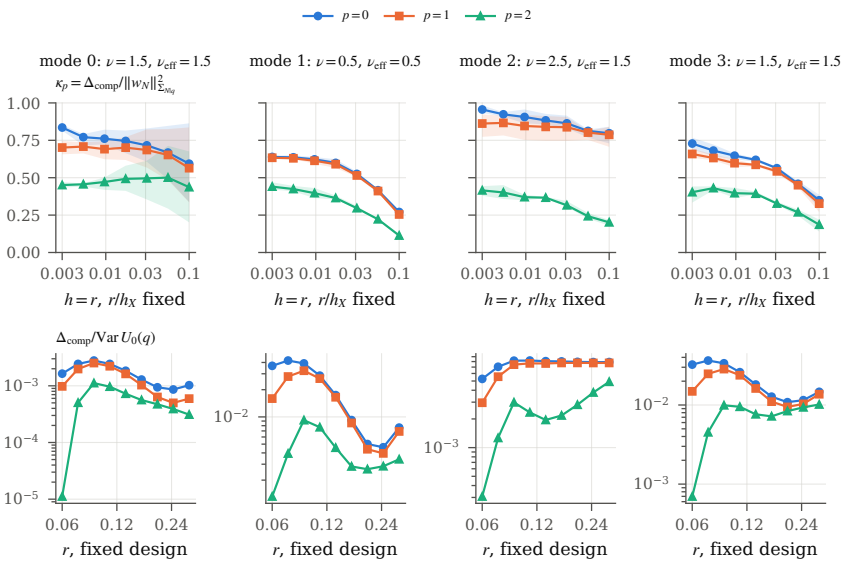


Figure 4: **Effect of local polynomial degree and local scale on compression.** The top row shows the normalized compression ratio κ_p as the local scale $h = r$ varies with r/h_X fixed. The bottom row shows $\Delta_{\text{comp}}/\text{Var}(U_0(q))$ as the localization radius varies for a fixed design X_m, again with $h = r$. It is not monotone because r influences both number of neighboring contexts N and bandwidth. Enlarging N gives the oracle more information, which raises Δ_{comp} until N covers the range of the kriging weights w_N, whereas widening the window lets ℓ uses the points whose w_N is large with higher weights, so this lowers it. Higher order local statistics generally retain more neighborhood information, with the improvement depending on the covariance smoothness.

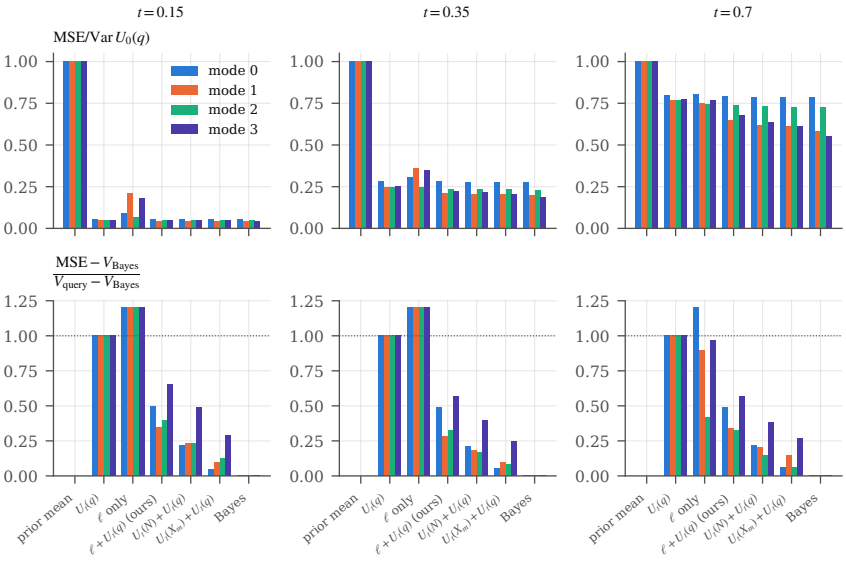


Figure 5: **Information hierarchy of different denoisers at three diffusion times.** The top row compares normalized MSE, while the bottom row reports the relative remaining Bayes gap. The compressed statistic $\ell + U_t(q)$ retains much of the benefit of the uncompressed local neighborhood, and reduces MSE greatly compared to query-only and local regression only conditioning.

Networks. The global denoiser D_ϕ is a U-Net on the 8×8 global state. The regularized network has half the width and dropout 0.1. The class-conditional network has dropout 0.1 and adds a label embedding to its time embedding. It is trained with the label dropped for 10% of the images and

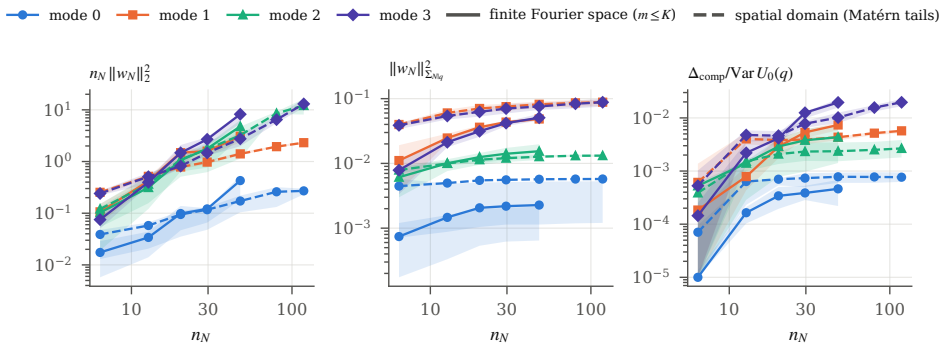


Figure 6: **Growth of the kriging weights under design refinement.** As the design is refined with r fixed, number of neighboring contexts n_N grows. The Euclidean norm $n_N\|w_N\|_2^2$ (left) increases because neighboring weights alternate in sign, whereas the energy norm $\|w_N\|^2_{\Sigma_{N|q}}$ (middle), which is the quantity entering the compression bound, increases in the early stage while saturating later, and so does Δ_{comp} (right). The finite Fourier space (solid, $m \le K$) and the spatial domain (dashed) agree on the three quantities broadly.

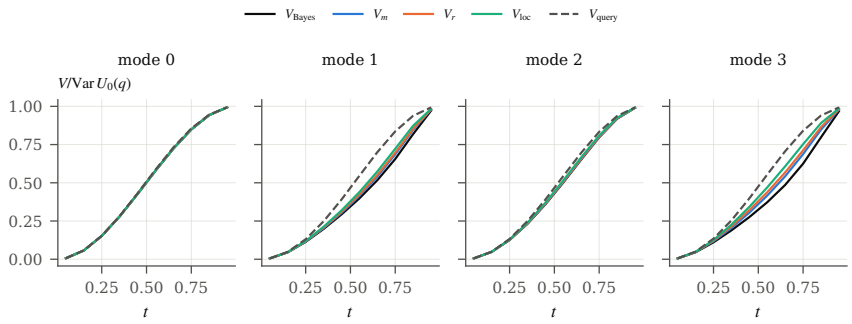


Figure 7: **Information ladder along the diffusion time.** Conditional variances $V_{\text{Bayes}} \le V_m \le V_r \le V_{\text{loc}} \le V_{\text{query}}$, normalized by $\text{Var}\, U_0(q)$, for the four covariance modes. The ordering in Theorem 1 holds at every t. All levels vanish as $t \to 0$ and approach the prior variance as $t \to 1$. The gaps are widest at intermediate t and for the rough or strongly anisotropic modes 1 and 3.

sampled without guidance. The residual denoiser D_θ has a trunk that maps (U_t, G_0) to a feature map and a per-query MLP head that reads the feature map at q, $U_t(q)$, $l(U_t; q)$, Fourier features of q and a time embedding. The full-context trunk is a U-Net with attention. The local trunk is a stack of r blocks, each a 3×3 convolution followed by a 1×1 convolution with per-pixel normalization, so its receptive field is exactly the $(2r+1)^2$ window. G_0 enters through a separate branch, which may be global since G_0 is a function of A_0. At $r = 0$ the stack is empty and the window enters through l alone. The statistic l is the vector of degree-2 local polynomial regression coefficients with a Wendland weight at three different scales. The simultaneous and monolithic baselines have the same architecture as the residual denoiser. The simultaneous model is conditioned on the noisy global state A_t, and the monolithic model denoises the whole noisy field F_t without a global state. Both stages are trained with AdamW and the loss weight $|B(t)| \propto 1/\sigma_t$ of Theorem 2, the global stage with ε-prediction and the residual stage with x_0-prediction.

Sampling and Evaluation. Both chains use 250 DDIM steps, and the residual x_0-prediction is projected by $I - \mathcal{R}S_L$ at every step, so it has zero global state. The gen and oracle arms share the terminal noise of each sample, so every comparison is paired. FID uses 10,000 samples against the training set. The thumbnail FID compares $\mathcal{R}\hat{A}_0$ with $\mathcal{R}A_0$ of the training images. FD_A is the Fréchet distance between generated and training global states, reported as $\sqrt{\text{FD}_A}$ in percent of their RMS. Real test images give 1.97 and 3.6% on these two statistics.

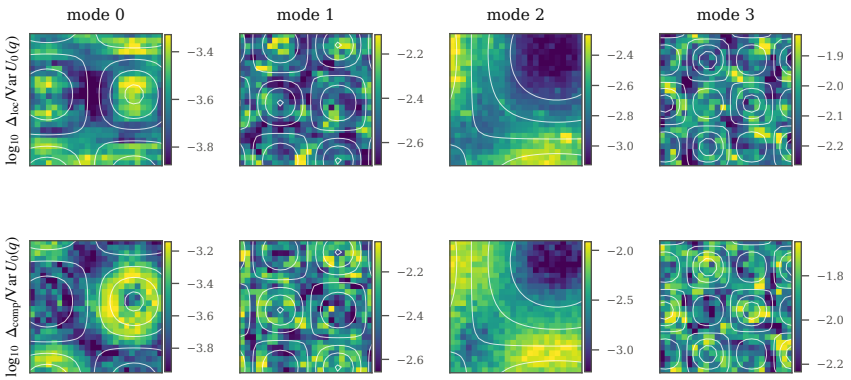


Figure 8: **Spatial distribution of localization and compression errors.** It shows the maps of $\log_{10} \Delta_{\text{loc}}/\text{Var}\, U_0(q)$ (top) and $\log_{10} \Delta_{\text{comp}}/\text{Var}\, U_0(q)$ (bottom) over query locations q, with contours of the local length-scale field of each nonstationary mode overlaid. The spatial variation of both errors follows the nonstationary covariance geometry rather than the context design grid, showing that approximation difficulty depends on the local correlation structure, with larger errors tending to occur where the correlation length is short relative to r and h.

Table 2: **Conditioning augmentation.** FID-10k with the base global stage. The no aug and aug columns are the base residual trained for 200 more epochs,, without and with augmentation ($\sigma \sim U[0, \frac{1}{2}]$). σ_s is the noise level of the global state at sampling.

residual	base	no aug	aug				
epochs	2500	2700	2700				
σ_s			0	0.1	0.2	0.3	0.5
gen	40.7	42.5	**30.8**	31.6	36.3	40.7	48.8
oracle	14.3	15.1	**12.3**	16.3	22.0	27.4	37.1
gap	26.4	27.4	18.5	15.3	14.3	13.3	11.6

Conditioning augmentation. Table 2 is the ablation behind finding (i). With the same 200 extra epochs, training without augmentation leaves the gap at 27.4, while training with it lowers the gap to 18.5 and the oracle FID from 15.1 to 12.3. The oracle arm uses held-out global states, so part of this gain is generalization. In contrast, augmentation at sampling ($\sigma_s > 0$) raises both FIDs monotonically in σ_s, because the residual is then generated from a noisy global state. The gap narrows only because the oracle FID rises faster.

Global-Stage Memorization. Table 3 supports finding (ii). Training the base global net longer or wider lowers its training loss but raises the thumbnail FID, and its held-out error grows relative to its training error, a sign of memorization. The generated global states are not copies of training ones. Without dropout, the thumbnail FID is lowest early in training and then rises, while dropout keeps it more stable. We use the half-width net with dropout 0.1 as the regularized global stage. It does not overfit, its thumbnail FID is close to the best, and it was chosen on training data only.

Locality. Table 4 supports finding (iii). The full-context and local residual networks are trained alike and differ only in how much of the noisy residual they see. With longer training, the full-context network's FID rises while its training loss falls, as in the global stage, but the local network's FID stays flat. The local network then becomes better given the true global state as well, not only end to end. A network that sees the whole field can memorize the training residuals, while a small window limits what it can memorize.

Samples. Figure 12 compares samples of the hierarchical, simultaneous and monolithic models. Each column shares one noise draw across all rows. The top rows are generated from the true global state of the test image in the first row. The middle rows show the generated global state $\mathcal{R}\hat{A}_0$ of the

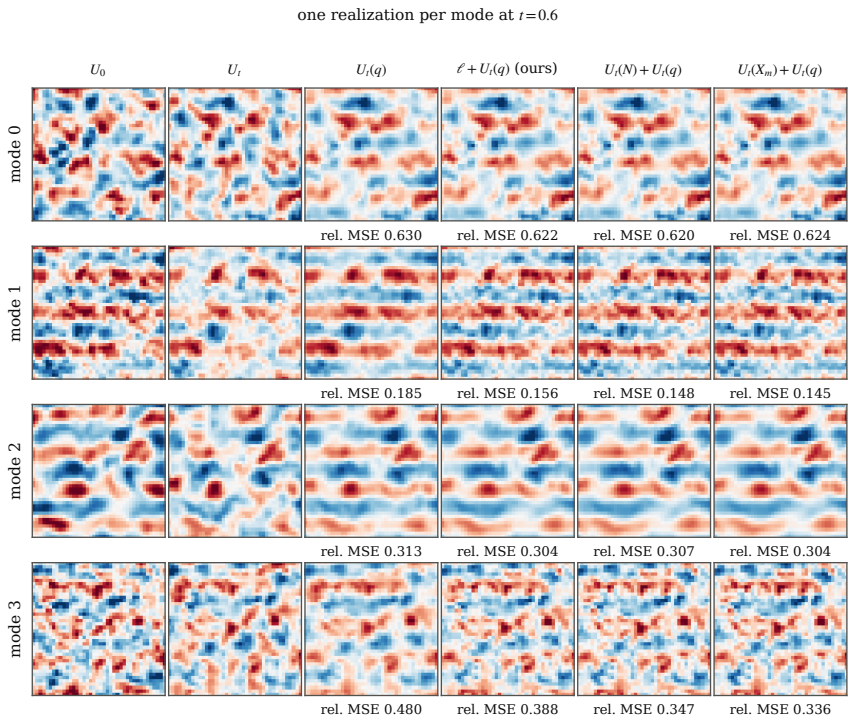


Figure 9: **One realization per mode at** $t = 0.6$**.** For each mode one residual U_0 and one noisy function $U_t = \alpha_t U_0 + \sigma_t U_1$ are drawn. Unlike the reverse chains of Figure 2, no generation is run. The remaining panels show the posterior mean of U_0 computed from U_t under different information sets: the query-only $U_t(q)$, the compressed local statistic $\ell + U_t(q)$ (ours), the raw neighborhood $U_t(N) + U_t(q)$, and the full context $U_t(X_m) + U_t(q)$, with the relative MSE against U_0 below each panel. The compressed predictor is visually indistinguishable from the full context one. Its relative MSE matches to the full context within 0.01 in modes 0–2 and exceeds it by 0.05 (0.388 against 0.336) in the strongly anisotropic mode 3, while still recovering two thirds of the gap between query-only and full context conditioning.

Table 3: **Global-Stage Memorization.** Thumbnail FID and error ratio (held-out over training denoising error) of the global network. Top: the base network (width 128, no dropout) trained longer or made wider. Bottom: thumbnail FID during training. The regularized global stage is the network of width 64 with dropout 0.1 after 2500 epochs. Real test images give a thumbnail FID of 1.97.

width	epochs	error ratio	thumbnail FID
128	2500	1.67	6.71
128	3500	1.99	7.87
192	2500	2.64	10.07

		thumbnail FID at epoch				error ratio
width	dropout	250	750	1500	2500	at 2500
128	0	**3.07**	4.18	5.78	6.61	1.69
128	0.1	**3.23**	3.26	3.77	4.13	1.26
128	0.3	3.64	**3.17**	3.22	3.40	1.10
64	0.1	4.06	3.36	**3.22**	3.26	1.05

regularized global stage, the samples built on it and the monolithic samples. The bottom rows do the same for the class-conditional global stage.

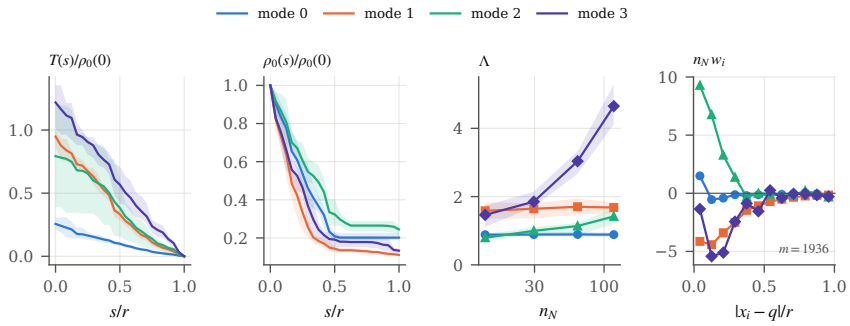


Figure 10: **Screening condition and kriging weight profiles.** Left two panels: the tail mass of the kriging coefficients $T(s)$ and the covariance envelope $\rho_0(s)$, both normalized by $\rho_0(0)$, as functions of s/r; Assumption 3 requires $T(s) \le \Lambda \rho_0(s)$. Third panel: the screening constant Λ against n_N under design refinement, which stays moderate and grows slowly. Right: the scaled weights $n_N w_i$ against $|x_i - q|/r$ at $m = 1936$, which collapse onto a mode dependent radial shape: largest in magnitude near the query, negligible beyond about half the radius, and positive near the query for the smooth mode 2 but negative for the rougher modes 1 and 3.

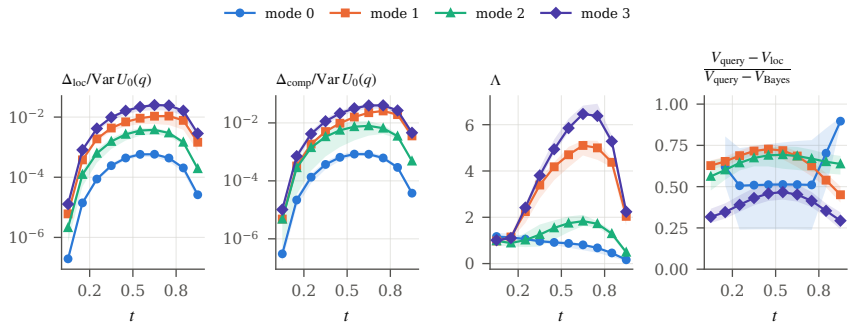


Figure 11: **Localization, compression, and screening along the diffusion time.** Normalized Δ_{loc} and Δ_{comp} (left two panels) peak at intermediate t and vanish at both ends, where either the noise or the signal dominates. The screening constant Λ (third panel) stays $O(1)$ and decreases as $t \to 1$, so the factor $(1 + \Lambda)^2$ in Theorem 6 does not blow up at high noise. The right panel shows the fraction of the Bayes headroom $V_{\mathrm{query}} - V_{\mathrm{Bayes}}$ retained by the compressed local statistic, between 0.3 and 0.9 across modes and times.

B.3 Signed Distance Fields of MNIST and Fashion-MNIST

Data. Each image (28×28; 60,000 training and 10,000 test images for both MNIST and Fashion-MNIST) is thresholded at 0.3 of its $[0, 1]$ intensity and converted to the signed distance to its boundary, clipped at $d_{\max} = 8$ pixels and divided by $d_{\max}$. For the 64×64 setting, images are resampled bilinearly to 64^2 before thresholding, with $d_{\max} = 18.3$ pixels. This tests how the local statistic and the model scale with the grid on the same shapes without adding new detail. The global state is the first $L = 64$ two-dimensional orthonormal DCT coefficients in nested zig-zag order, a basis encoder in the sense of Section 3, with $\mathcal{R}$ the inverse transform on those modes, so the residual $U_0 = F_0 - \mathcal{R}A_0$ is exact. The residual noise is white noise projected onto the orthogonal complement of these 64 modes.

Models. The global denoiser is a three-layer MLP on the normalized A_0 with time and class embeddings (0.35M parameters). The residual denoiser is a pointwise MLP (0.12M parameters). At each query q it receives the six coefficients of a Gaussian-weighted degree-2 polynomial fit to the noisy residual over a radius-3 window (radius 7 at 64^2), the value, gradient and Laplacian of the coarse field $\mathcal{R}A_0$ at q, the coordinates of q, and embeddings of A_0, time and class label. The full-context baseline is an MLP on the flattened residual field with the same conditioning (2.5M

Table 4: **Locality.** FID-10k of the full-context and local residual networks over training, with the base global stage.

residual	full			loc $r = 8$		
epochs	1000	1500	2000	1000	1500	2000
gen	35.2	37.7	39.7	32.6	32.7	32.6
oracle	17.9	20.3	22.0	19.6	19.6	19.6

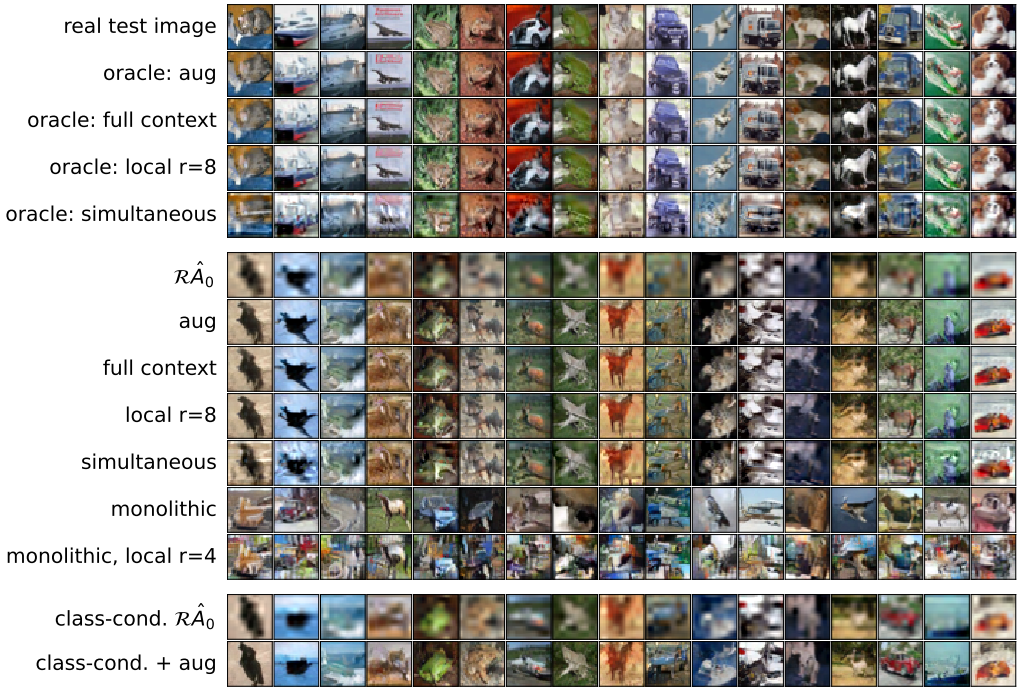


Figure 12: **CIFAR-10 samples.**

parameters at 28^2, 7.6M at 64^2). All models are class-conditioned with classifier-free guidance and are trained for 500 epochs with Adam. Both residual models sample from the same global state, label and terminal noise.

Results. The local model has the lower denoising error at every noise level on both 28^2 datasets, by 2–7$\times$ at low noise, and on the 64^2 grid except at $t = 0.96$, where the residual's conditional range is longest (Table 5). On eikonal error the full-context model is closer to the data at 28^2 and the local model closer at 64^2, where its size is unchanged while the baseline's triples. On topology neither is consistently closer. On Fashion-MNIST both models generate fewer fragmented shapes than the data (single component fraction 0.780 and 0.848 against 0.727); the local model is closer to the data, so the shortfall is not caused by locality. The ranking is the same at 60 and 200 epochs. Figure 13 shows one real and one generated field per class for both models.

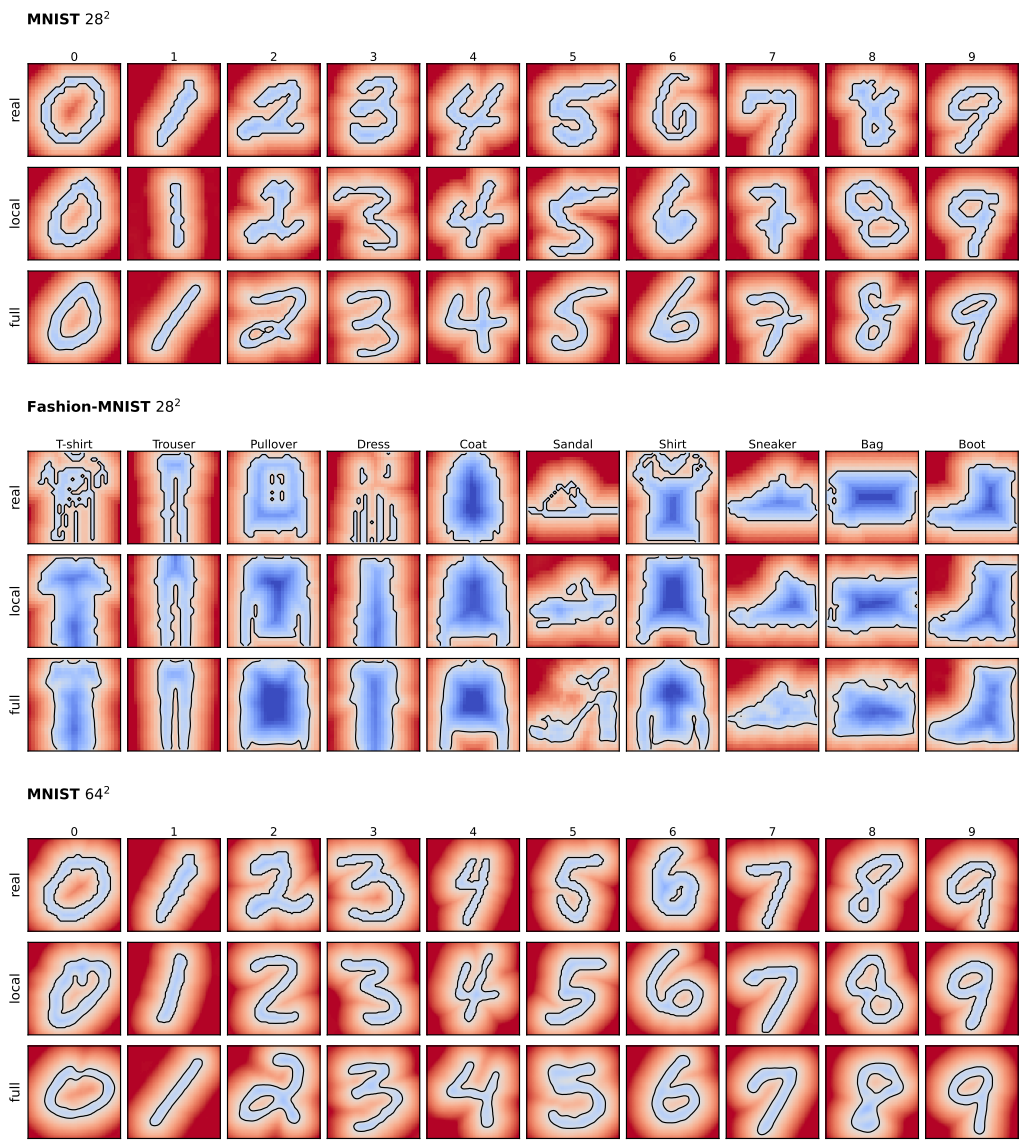


Figure 13: **Signed distance fields.** One field per class. Rows from top to bottom: real, local, full. The latter two are from the same global state, label and terminal noise. Color is the signed distance, the black curve is its zero level set.

C Assumptions and Auxiliary Statements

This appendix states the assumptions used by the results of Section 4 and the auxiliary statements used in their proofs.

Assumption 4 (Local Regression Stability)**.** *There exists $C_{\mathcal{L}} < \infty$, uniform over the admissible m, r and X_m, such that for all context values $y, y' \in \mathbb{R}^m$,*

$$\|l(y;\cdot) - l(y';\cdot)\|_{L^2(\pi,\mathbb{R}^{D_P})} \le C_{\mathcal{L}}\|y - y'\|_m \quad \textit{and} \quad \|l(y;X_m) - l(y';X_m)\|_m \le C_{\mathcal{L}}\|y - y'\|_m,$$

where $\|y\|_m^2 = \frac{|\chi|}{m}\sum_{i=1}^m |y_i|^2$.

Assumption 5 (Denoiser Stability)**.** *For all $t \in [\epsilon, 1]$ and $q \in \chi$, the learned denoiser satisfies*

$$|D_\theta(t,a,u,l,q) - D_\theta(t,a',u',l',q)| \le L_A(t)\|a - a'\| + L_u(t)|u - u'| + L_l(t)\|l - l'\|, \quad (17)$$

with L_A, L_u and L_l integrable on $[\epsilon, 1]$.

Table 5: **Signed Distance Fields.** Held-out x_0-MSE at time t, single-component fraction and eikonal error $\mathbb{E}\,||\nabla d| - 1|$ of 500 generated fields, compared with 2,000 real fields (data).

		x_0-MSE			single	eikonal	
Dataset	model	τ=0.1	0.5	0.96	comp.	error	params
MNIST 28^2	local	**0.015**	**0.043**	**0.249**	0.922	0.233	0.12M
	full	0.102	0.171	0.287	0.938	0.196	2.5M
	data	–	–	–	0.972	0.201	–
Fashion 28^2	local	**0.112**	**0.161**	**0.485**	0.780	0.254	0.12M
	full	0.209	0.306	0.583	0.848	0.229	2.5M
	data	–	–	–	0.727	0.218	–
MNIST 64^2	local	**0.046**	**0.061**	0.222	0.958	0.133	0.12M
	full	0.100	0.114	**0.207**	0.974	0.142	7.6M
	data	–	–	–	0.980	0.106	–

Assumption 6 (Regularity)**.** *The schedule (α_t, σ_t) is continuously differentiable on $[\epsilon, 1]$ with $\sigma_t > 0$. The moments $\mathbb{E}\|A_0\|^2$, $\mathbb{E}\|U_0\|^2_{L^2(\pi)}$ and $\mathbb{E}\|U_1\|^2_{L^2(\pi)}$ are finite.*

Assumption 6 makes $A(t)$, $B(t)$ and ϵ_{stop} finite. With Assumptions 4 and 5, the learned residual flow of Theorem 2 is Lipschitz in the context values, so it has a unique solution on $[\epsilon, 1]$.

Assumption 7 (Non-degenerate Residual Noise)**.** *For every design X_m and query q considered, the residual noise covariance C_U restricted to $\{q\} \cup X_m$ satisfies $\lambda_{\min}(C_U) > 0$.*

Assumption 8 (Learning and Approximation Rates)**.** *Let N be the number of training functions, let $\mathbb{E}$ denote the expectation over the training set, and let $N_* = N \min\{|\chi|/(r + l_c)^d, |Q|\}$. (i) The law μ_A and the global stage satisfy the assumptions of Oko et al. (2023), with Besov smoothness θ on $[-1, 1]^L$. (ii) Uniformly in $t \in [\epsilon, 1]$, the learning error averaged over $q \sim \bar{\pi}_m$ satisfies*

$$\mathbb{E}\,\bar{\Delta}_{\text{learn}}(t) \lesssim N_*^{-2\theta'/(2\theta'+L+D_p+d+1)} + N^{-2\theta'/(2\theta'+L)}.$$

(iii) Uniformly in $t \in [\epsilon, 1]$ and $q \in \chi$, the discretization error satisfies $\Delta_{\text{disc}}^{1/2} \lesssim m^{-s/d}$.

Part (ii) is the regression rate for a function of smoothness θ' of the $L + D_p + d + 1$ inputs (a, u, l, q), adjusted to how the training queries are drawn. Two queries of one function are nearly independent only when their windows of radius r are more than l_c apart, so each function contributes about $\min\{|\chi|/(r + l_c)^d, |Q|\}$ samples, which gives N_*. All queries of one function share the same A_0, so the dependence on a is learned from only N values, which gives the second term. For part (iii), when U_0 and U_t are jointly Gaussian given A_0, as in Proposition 7, d^*_{Bayes} is an affine functional of U_t, and Δ_{disc} is the error of approximating it by the values of U_t on $\{q\} \cup X_m$. The rate $m^{-s/d}$ is the scattered-data approximation rate on a quasi-uniform design when this functional has Sobolev smoothness s. When C_U is proportional to C_a, d^*_{Bayes} depends on U_t only through $U_t(q)$, and $\Delta_{\text{disc}} = 0$.

Theorem 2, Proposition 3 and Theorem 4 use Assumptions 4–6. Theorem 6 and Proposition 12 use Assumptions 1–3 and 7. Theorem 10 uses all of them. Its localization and compression terms follow from Theorem 6 and Proposition 9 with $h = r$.

C.1 Localization Error and the Correlation Ratio

Proposition 11. *Let $V_{\text{full}} = \mathbb{E}[\text{Var}(U_0(q) \mid A_0, U_t(q), U_t(X_m))]$ and suppose $\bar{\eta}(r) < 1$. Then*

$$\mathbb{E}[d^*_{\text{full}} - d^*_r]^2 = \frac{\bar{\eta}^2(r)}{1 - \bar{\eta}^2(r)}\, V_{\text{full}}.$$

Thus, when $V_{\text{full}} > 0$, the localization error vanishes exactly when $\bar{\eta}(r) \to 0$.

C.2 Polynomial Covariance Decay

Proposition 12. *Under the assumptions of Theorem 6, with the polynomial envelope* $\rho_0(u) = c_0(1 + u/l_c)^{-a}$, $a > d/2$, *in place of the exponential one,*

$$\mathbb{E}[d^*_{\text{full}} - d^*_r]^2 \lesssim \frac{\alpha_t^2}{\sigma_t^2} \cdot \frac{m}{|\chi|}\, r^{d-2a}.$$

C.3 Kriging Weights and the Local Polynomial Statistic

Fix A_0, write $N = N(q, r)$ and $n_h = |N|$, and center all variables given A_0. Conditionally on A_0, let $s_{qq} = \mathrm{Var}(U_t(q))$, $s_{qN} = s_{Nq}^T = \mathrm{Cov}(U_t(q), U_t(N))$, $\Sigma_N = \mathrm{Cov}(U_t(N))$, $\gamma_q = \mathrm{Cov}(U_0(q), U_t(q))$ and $\sigma_N = \mathrm{Cov}(U_0(q), U_t(N))$.

Proposition 13. *Under Assumptions 2 and 7,* $d^*_r = w_q U_t(q) + w_N^T U_t(N)$ *with*

$$w_N = \Sigma_N^{-1}\sigma_N - \Sigma_N^{-1} s_{Nq} w_q, \qquad w_q = \frac{\gamma_q - s_{qN}\Sigma_N^{-1}\sigma_N}{s_{qq} - s_{qN}\Sigma_N^{-1} s_{Nq}}. \tag{18}$$

Let $\mathcal{P}_p$ be the polynomials of degree at most p on $\mathbb{R}^d$, and $D_p = \dim \mathcal{P}_p$. For multi-indices α with $|\alpha| \le p$, let $V = (V_{i\alpha}) = ((x_i - q)^\alpha)_{x_i \in N}$ be the design matrix and $W = \mathrm{diag}(W_h(x_i - q))_{x_i \in N}$ the weight matrix, and suppose $V^T W V$ is invertible. The local statistic is $l(U_t; q) = \mathcal{L}\, U_t(N)$, with the local polynomial coefficient map $\mathcal{L} = (V^T W V)^{-1} V^T W$.

Lemma 1. $\mathrm{ran}(\mathcal{L}^T) = \{(W_h(x_i - q)\varphi(x_i - q))_{x_i \in N} : \varphi \in \mathcal{P}_p\}$.

Proof. Since $\mathcal{L}^T = WV(V^T W V)^{-1}$ and $(V^T W V)^{-1}$ is invertible, $\mathrm{ran}(\mathcal{L}^T) = \mathrm{ran}(WV)$. For $c \in \mathbb{R}^{D_p}$, the entries of WVc are $W_h(x_i - q)\varphi(x_i - q)$ with $\varphi(x) = \sum_{|\alpha| \le p} c_\alpha x^\alpha$, and every $\varphi \in \mathcal{P}_p$ has this form. □

D Proofs

D.1 Proof of Theorem 1

Proof. First we show that

$$\sigma(A_0, U_t(q), l(U_t; q)) \subseteq \sigma(A_0, U_t(q), U_t(N(q, r))) \subseteq \sigma(A_0, U_t(q), U_t(X_m)) \subseteq \sigma(A_0, U_t).$$

The first inclusion is the hypothesis that $l(U_t; q)$ is measurable with respect to $U_t(N(q, r))$. The second holds since $N(q, r) \subseteq X_m$, and the third since point evaluation is bounded on $\mathcal{H}$. Denote these σ-algebras by $\mathcal{G}_{\text{local}} \subseteq \mathcal{G}_r \subseteq \mathcal{G}_{\text{full}} \subseteq \mathcal{G}_{\text{Bayes}}$, so that d^*_{local}, d^*_r, d^*_{full} and d^*_{Bayes} are the conditional expectations of $U_0(q)$ given each of them. Since D_θ is a function of $(A_0, U_t(q), l(U_t; q))$, it is $\mathcal{G}_{\text{local}}$-measurable, and

$$d^*_{\text{Bayes}} - D_\theta = (d^*_{\text{Bayes}} - d^*_{\text{full}}) + (d^*_{\text{full}} - d^*_r) + (d^*_r - d^*_{\text{local}}) + (d^*_{\text{local}} - D_\theta).$$

For nested $\mathcal{G}_1 \subseteq \mathcal{G}_2$ and $Y \in L^2$, the tower property shows that $\mathbb{E}[Y \mid \mathcal{G}_2] - \mathbb{E}[Y \mid \mathcal{G}_1]$ is orthogonal to every square-integrable $\mathcal{G}_1$-measurable variable. Each increment above is therefore orthogonal to all the increments to its right, and expanding the square gives $\mathbb{E}[D_\theta - d^*_{\text{Bayes}}]^2 = \Delta_{\text{disc}} + \Delta_{\text{loc}} + \Delta_{\text{comp}} + \Delta_{\text{learn}}$. In the same way, $U_0(q) - d^*_{\text{Bayes}}$ is orthogonal to the $\mathcal{G}_{\text{Bayes}}$-measurable $D_\theta - d^*_{\text{Bayes}}$, which gives the second identity. □

D.2 Proof of Theorem 2

Proof. The final function space Wasserstein distance is measured with respect to continuous queries $q \sim \pi$, while the context dynamics are evaluated on X_m. Throughout, Wasserstein distances between residual laws are taken in $L^2(\pi)$. To control both terms, let

$$\pi_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}, \quad \overline{\pi}_m := \lambda\pi + (1 - \lambda)\pi_m, \quad 0 < \lambda < 1.$$

Write $D_\theta(t,q) = D_\theta\big(t, A_0, U_t(q), l(U_t;q), q\big)$, with $l(U_t;q)$ formed from $U_t(X_m)$, and define the continuous query denoising, grid denoising, and mixed denoising errors by

$$\eta_\pi(t)^2 = \mathbb{E}_{q\sim\pi}|D_\theta(t,q) - d^*_{\text{Bayes}}(t,q)|^2,$$

$$\eta_{\pi_m}(t)^2 = \mathbb{E}_{q\sim\pi_m}|D_\theta(t,q) - d^*_{\text{Bayes}}(t,q)|^2 = \frac{1}{m}\sum_{i=1}^m \mathbb{E}|D_\theta(t,x_i) - d^*_{\text{Bayes}}(t,x_i)|^2,$$

and

$$\eta_{\text{mix}}(t)^2 := \mathbb{E}_{q\sim\overline{\pi}_m}|D_\theta(t,q) - d^*_{\text{Bayes}}(t,q)|^2.$$

By linearity,

$$\eta_{\text{mix}}(t)^2 = \lambda\eta_\pi(t)^2 + (1-\lambda)\eta_{\pi_m}(t)^2.$$

The grid error in the norm $\|\cdot\|_m$ is

$$\eta_m(t)^2 = \frac{|\chi|}{m}\sum_{i=1}^m \mathbb{E}|D_\theta(t,x_i) - d^*_{\text{Bayes}}(t,x_i)|^2 = |\chi|\eta_{\pi_m}(t)^2.$$

With

$$C_\lambda := \left[\max\left\{\frac{1}{\lambda}, \frac{|\chi|}{1-\lambda}\right\}\right]^{1/2},$$

we have $1 \le C_\lambda^2\lambda$ and $|\chi| \le C_\lambda^2(1-\lambda)$, so comparing coefficients gives

$$\eta_\pi(t)^2 + \eta_m(t)^2 = \eta_\pi(t)^2 + |\chi|\eta_{\pi_m}(t)^2 \le C_\lambda^2\left(\lambda\eta_\pi(t)^2 + (1-\lambda)\eta_{\pi_m}(t)^2\right) = C_\lambda^2\eta_{\text{mix}}(t)^2.$$

Now fix $A_0 = a$, and consider a generic forward function interpolation $U_t = \alpha_t U_0 + \sigma_t U_1$. Differentiating a sample path gives

$$\dot{U}_t = \dot{\alpha}_t U_0 + \dot{\sigma}_t U_1.$$

The velocity field in reverse process is

$$v^*(t,u) = \mathbb{E}[\dot{\alpha}_t U_0 + \dot{\sigma}_t U_1 \mid U_t = u] \tag{19}$$

$$= \dot{\alpha}_t\mathbb{E}[U_0 \mid U_t = u] + \frac{\dot{\sigma}_t}{\sigma_t}\left(u - \alpha_t\mathbb{E}[U_0 \mid U_t = u]\right) \tag{20}$$

$$= A(t)u + B(t)\mathbb{E}[U_0 \mid U_t = u], \tag{21}$$

where $A(t) = \frac{\dot{\sigma}_t}{\sigma_t}, B(t) := \dot{\alpha}_t - \alpha_t\frac{\dot{\sigma}_t}{\sigma_t}$.

Let U_t^* be the Bayes optimal trajectory and $\hat{U}_t$ be the learned trajectory, using the same terminal noise at $t=1$. Then the probability flow of U_t^* is

$$\partial_t U_t^*(q) = A(t)U_t^*(q) + B(t)d^*_{\text{Bayes}}(t,q),$$

and similarly

$$\partial_t\widehat{U}_t(q) = A(t)\widehat{U}_t(q) + B(t)\widehat{D}_\theta(t,q), \qquad \widehat{D}_\theta(t,q) := D_\theta\big(t,a,\widehat{U}_t(q), l(\widehat{Y}_t;q), q\big),$$

where

$$Y_t^* = U_t^*(X_m), \qquad \hat{Y}_t = \hat{U}_t(X_m).$$

Set

$$e_t^U := U_t^* - \widehat{U}_t, \qquad e_t^Y := Y_t^* - \widehat{Y}_t.$$

Let $D^*_\theta(t,q) := D_\theta\big(t,a,U_t^*(q), l(Y_t^*;q), q\big)$ and $\xi_t(q) = d^*_{\text{Bayes}}(t,q) - D^*_\theta(t,q)$, then

$$d^*_{\text{Bayes}}(t,q) - \widehat{D}_\theta(t,q) = [d^*_{\text{Bayes}}(t,q) - D^*_\theta(t,q)] + [D^*_\theta(t,q) - \widehat{D}_\theta(t,q)]. \tag{22}$$

The first term is the denoising error $\xi_t(q)$. The second term is controlled by Assumptions 4 and 5,

$$\|D^*_\theta(t,\cdot) - \widehat{D}_\theta(t,\cdot)\|_{L^2(\pi)} \le L_u(t)\|e_t^U\|_{L^2(\pi)} + L_l(t)C_{\mathcal{L}}\|e_t^Y\|_m. \tag{23}$$

The difference equation is

$$\partial_t e_t^U(q) = A(t)e_t^U(q) + B(t)\xi_t(q) + B(t)[D^*_\theta(t,q) - \widehat{D}_\theta(t,q)].$$

Integrating it backward from $t = 1$ to t gives

$$\|e_t^U\|_{L^2(\pi)} \le \int_t^1 (|A(s)| + |B(s)|L_u(s))\|e_s^U\|_{L^2(\pi)}ds \tag{24}$$

$$+ \int_t^1 |B(s)|L_l(s)C_{\mathcal{L}}\|e_s^Y\|_m ds \tag{25}$$

$$+ \int_t^1 |B(s)|\|\xi_s\|_{L^2(\pi)}ds. \tag{26}$$

Similarly, the difference e_t^Y can be bounded by

$$\|e_t^Y\|_m \le \int_t^1 [|A(s)| + |B(s)|(L_u(s) + L_l(s)C_{\mathcal{L}})]\|e_s^Y\|_m ds + \int_t^1 |B(s)|\|\xi_s(X_m)\|_m ds.$$

Let

$$E_t = \max\left\{\|e_t^U\|_{L^2(\pi)}, \|e_t^Y\|_m\right\},$$

and

$$\Xi_t = \max\left\{\|\xi_t(\cdot)\|_{L^2(\pi)}, \left(\frac{|\chi|}{m}\sum_{i=1}^m |\xi_t(x_i)|^2\right)^{1/2}\right\}.$$

With $L_{\text{ode}}(t) = |A(t)| + |B(t)|\left(L_u(t) + L_l(t)C_{\mathcal{L}}\right)$, the query and context inequalities imply

$$E_t \le \int_t^1 L_{\text{ode}}(s)E_s ds + \int_t^1 |B(s)|\Xi_s ds.$$

The two trajectories start from the same terminal noise, so $E_1 = 0$. The Gronwall inequality gives

$$E_\epsilon \le \int_\epsilon^1 \exp\left(\int_\epsilon^t L_{\text{ode}}(s)ds\right)|B(t)|\Xi_t dt.$$

Since $\|U_\epsilon^* - \widehat{U}_\epsilon\|_{L^2(\pi)} \le E_\epsilon$, by Minkowski's integral inequality

$$\left(\mathbb{E}\|U_\epsilon^* - \widehat{U}_\epsilon\|_{L^2(\pi)}^2\right)^{1/2} \le \int_\epsilon^1 \exp\left(\int_\epsilon^t L_{\text{ode}}(s)ds\right)|B(t)|\left(\mathbb{E}\Xi_t^2\right)^{1/2}dt.$$

The Bayes trajectory U_t^* has the law of U_t given $A_0 = a$, and $\Xi_t^2 \le \|\xi_t\|_{L^2(\pi)}^2 + \frac{|\chi|}{m}\sum_i |\xi_t(x_i)|^2$. Hence, by the mixed sampling, $\left(\mathbb{E}\Xi_t^2\right)^{1/2} \le (\eta_\pi(t)^2 + \eta_m(t)^2)^{1/2} \le C_\lambda \eta_{\text{mix}}(t)$, with the errors taken conditionally on $A_0 = a$. The pair $(U_\epsilon^*, \widehat{U}_\epsilon)$ is a coupling of $\mu_{U_\epsilon^*|a}$ and $\mu_{\widehat{U}|a}$, the law of $\widehat{U}_\epsilon$, therefore

$$\mathcal{W}_2\left(\mu_{U_\epsilon^*|a}, \mu_{\widehat{U}|a}\right) \le \left(\mathbb{E}\left[\|U_\epsilon^* - \widehat{U}_\epsilon\|_{L^2(\pi)}^2 \mid a\right]\right)^{1/2} \le C_\lambda \int_\epsilon^1 g_\epsilon(t)\eta_{\text{mix}}(t)dt, \tag{27}$$

where $g_\epsilon(t) := |B(t)|\exp\left(\int_\epsilon^t L_{\text{ode}}(s)ds\right)$.

Next we bound the error caused by early stopping. Since $\mu_{U_\epsilon^*|a} = \mu_{U_\epsilon|a}$, where $U_\epsilon = \alpha_\epsilon U_0 + \sigma_\epsilon U_1$, we have

$$\mathcal{W}_2\left(\mu_{U_0|a}, \mu_{U_\epsilon^*|a}\right) \le \left(\mathbb{E}\left[\|U_0 - U_\epsilon\|_{L^2(\pi)}^2 \mid a\right]\right)^{1/2}$$

$$\le (1 - \alpha_\epsilon)\left(\mathbb{E}[\|U_0\|_{L^2(\pi)}^2 \mid a]\right)^{1/2} + \sigma_\epsilon\left(\mathbb{E}\|U_1\|_{L^2(\pi)}^2\right)^{1/2}.$$

By the triangle inequality, $\mathcal{W}_2(\mu_{U_0|a}, \mu_{\widehat{U}|a})$ is at most the sum of the two bounds. Let

$$\epsilon_{\text{stop}} := (1 - \alpha_\epsilon)[\mathbb{E}\|U_0\|_{L^2(\pi)}^2]^{\frac{1}{2}} + \sigma_\epsilon[\mathbb{E}\|U_1\|_{L^2(\pi)}^2]^{\frac{1}{2}}.$$

Averaging over A_0 with Minkowski's inequality, and noting that g_ϵ does not depend on a, gives

$$\left[\mathbb{E}_{A_0}\mathcal{W}_2^2\left(\mu_{U_0|A_0}, \mu_{\widehat{U}|A_0}\right)\right]^{1/2} \le C_\lambda \int_\epsilon^1 g_\epsilon(t)\eta_{\text{mix}}(t)dt + \epsilon_{\text{stop}}.$$

□

D.3 Proof of Proposition 3

Proof. Let $\mu_{\hat{U}|a}$ denote the law of the learned residual output given $A_0 = a$, as in Theorem 2. We show that the map $a \mapsto \mu_{\hat{U}|a}$ is Lipschitz in Wasserstein distance.

Fix a, a' and run two copies of the learned residual sampler with the same terminal noise. Denote the two query trajectories by $\widehat{U}_t^a$ and $\widehat{U}_t^{a'}$, and their context states by $\widehat{Y}_t^a = \widehat{U}_t^a(X_m), \widehat{Y}_t^{a'} = \widehat{U}_t^{a'}(X_m)$. Let

$$\Delta_t = \widehat{U}_t^a - \widehat{U}_t^{a'}, \quad \delta_t := \widehat{Y}_t^a - \widehat{Y}_t^{a'}.$$

The context dynamics satisfy

$$\partial_t \widehat{Y}_t^a = A(t)\widehat{Y}_t^a + B(t) D_\theta^{\mathrm{gr}}(t, \widehat{Y}_t^a, a),$$

where $D_\theta^{\mathrm{gr}}(t, Y, a) = \big(D_\theta(t, a, Y_i, l(Y; x_i), x_i)\big)_{i \le m}$ is the learned denoiser on the grid, and analogously for a'. Hence

$$\partial_t \delta_t = A(t)\delta_t + B(t)\left[D_\theta^{\mathrm{gr}}(t, \widehat{Y}_t^a, a) - D_\theta^{\mathrm{gr}}(t, \widehat{Y}_t^{a'}, a')\right].$$

By Assumptions 4 and 5,

$$\left\| D_\theta^{\mathrm{gr}}(t, \widehat{Y}_t^a, a) - D_\theta^{\mathrm{gr}}(t, \widehat{Y}_t^{a'}, a') \right\|_m \le (L_u(t) + L_l(t) C_{\mathcal{L}}) \, \|\delta_t\|_m + c_\chi L_A(t) \|a - a'\|,$$

where $c_\chi := \max\{1, \sqrt{|\chi|}\}$, since a vector with entries bounded by c has $\|\cdot\|_m$ norm at most $\sqrt{|\chi|}\, c$. Integrating backward from 1 to t therefore gives

$$\|\delta_t\|_m \le \int_t^1 L_{\mathrm{ode}}(s) \|\delta_s\|_m \, ds + c_\chi \|a - a'\| \int_t^1 |B(s)| L_A(s) \, ds,$$

where

$$L_{\mathrm{ode}}(t) = |A(t)| + |B(t)| \, (L_u(t) + L_l(t) C_{\mathcal{L}}).$$

For the continuous query trajectories,

$$\partial_t \widehat{U}_t^a(q) = A(t)\widehat{U}_t^a(q) + B(t) D_\theta\Big(t, a, \widehat{U}_t^a(q), l(\widehat{Y}_t^a; q), q\Big),$$

and similarly for a'. Therefore

$$\begin{aligned} \|\Delta_t\|_{L^2(\pi)} \le & \int_t^1 (|A(s)| + |B(s)| L_u(s)) \, \|\Delta_s\|_{L^2(\pi)} \, ds \\ & + \int_t^1 |B(s)| L_l(s) C_{\mathcal{L}} \|\delta_s\|_m \, ds \\ & + \|a - a'\| \int_t^1 |B(s)| L_A(s) \, ds. \end{aligned}$$

Let $E_t = \max\left\{\|\Delta_t\|_{L^2(\pi)}, \|\delta_t\|_m\right\}$. Since $c_\chi \ge 1$, the two inequalities give

$$E_t \le \int_t^1 L_{\mathrm{ode}}(\tau) E_\tau \, d\tau + c_\chi \|a - a'\| \int_t^1 |B(\tau)| L_A(\tau) \, d\tau.$$

Since $E_1 = 0$, the Gronwall inequality gives

$$E_\epsilon \le c_\chi \|a - a'\| \int_\epsilon^1 |B(t)| L_A(t) \exp\left(\int_\epsilon^t L_{\mathrm{ode}}(\tau) \, d\tau\right) dt.$$

The synchronous construction $(\widehat{U}_\epsilon^a, \widehat{U}_\epsilon^{a'})$ is a valid coupling of $\mu_{\hat{U}|a}$ and $\mu_{\hat{U}|a'}$, hence

$$\begin{aligned} \mathcal{W}_2(\mu_{\hat{U}|a}, \mu_{\hat{U}|a'}) & \le \left(\mathbb{E}\|\widehat{U}_\epsilon^a - \widehat{U}_\epsilon^{a'}\|_{L^2(\pi)}^2\right)^{1/2} \\ & \le \left(\mathbb{E} E_\epsilon^2\right)^{1/2} \\ & \le c_\chi \left[\int_\epsilon^1 |B(t)| L_A(t) \exp\left(\int_\epsilon^t L_{\mathrm{ode}}(s) \, ds\right) dt\right] \|a - a'\|. \end{aligned}$$

Therefore, we can take

$$L_{\text{gen}(\theta)} = c_\chi \int_\epsilon^1 |B(t)| L_A(t) \exp\left(\int_\epsilon^t L_{\text{ode}}(s)\, ds\right) dt,$$

so that $\mathcal{W}_2(\mu_{\hat{U}|a}, \mu_{\hat{U}|a'}) \le L_{\text{gen}(\theta)} \|a - a'\|$. □

D.4 Proof of Theorem 4

Proof. All Wasserstein distances between function laws are taken in $L^2(\pi)$, and $\|\mathcal{R}\|$ is the operator norm of $\mathcal{R} : \mathbb{R}^L \to L^2(\pi)$. Let $\mu_{\bar{F}}$ be the law of $\bar{F}_0 = \mathcal{R}A_0 + \hat{U}_0(A_0)$, where $A_0 \sim \mu_A$ and $\hat{U}_0(a) \sim \mu_{\hat{U}|a}$ is the learned residual given the true global state. By the triangle inequality $\mathcal{W}_2(\mu_F, \mu_{\hat{F}}) \le \mathcal{W}_2(\mu_F, \mu_{\bar{F}}) + \mathcal{W}_2(\mu_{\bar{F}}, \mu_{\hat{F}})$.

For the first term, couple $F_0 = \mathcal{R}A_0 + U_0$ and $\bar{F}_0$ through the same A_0 and an optimal coupling of $\mu_{U_0|A_0}$ and $\mu_{\hat{U}|A_0}$, chosen measurably in A_0. Then $F_0 - \bar{F}_0 = U_0 - \hat{U}_0(A_0)$, so

$$\mathcal{W}_2^2(\mu_F, \mu_{\bar{F}}) \le \mathbb{E}_{A_0 \sim \mu_A}[\mathcal{W}_2^2(\mu_{U_0|A_0}, \mu_{\hat{U}|A_0})].$$

To bound the second term, we let

$$\pi_A^* = \arg\min_{\pi_A \in \Pi(\mu_A, \mu_{\hat{A}})} \int_{\mathbb{R}^L \times \mathbb{R}^L} \|a - a'\|^2 d\pi_A(a, a').$$

For each pair $A_0 = a$, $\hat{A}_0 = a'$, let

$$\pi^*_{\hat{U}|a,a'} = \arg\min_{\pi_{\hat{U}|a,a'} \in \Pi(\mu_{\hat{U}|a}, \mu_{\hat{U}|a'})} \int_{V_U \times V_U} \|u - v\|^2_{L^2(\pi)} \pi_{\hat{U}|a,a'}(du, dv).$$

Consider the coupling $\pi(d\bar{F}_0, d\hat{F}_0) = d\pi_A^*(A_0, \hat{A}_0)\, \pi^*_{\hat{U}|A_0,\hat{A}_0}(d\hat{U}(A_0), d\hat{U}(\hat{A}_0))$ of $\mu_{\bar{F}}$ and $\mu_{\hat{F}}$, then

$$\begin{aligned}
\mathcal{W}_2(\mu_{\bar{F}}, \mu_{\hat{F}}) &\le [\mathbb{E}_{\bar{F}_0, \hat{F}_0 \sim \pi} \|\bar{F}_0 - \hat{F}_0\|^2_{L^2(\pi)}]^{\frac{1}{2}} && (28)\\
&\le [\mathbb{E}_{A_0, \hat{A}_0 \sim \pi_A^*} \|\mathcal{R}(A_0 - \hat{A}_0)\|^2_{L^2(\pi)}]^{\frac{1}{2}} \\
&\quad + [\mathbb{E}_{A_0, \hat{A}_0 \sim \pi_A^*} \mathbb{E}_{U, \hat{U} \sim \pi^*_{\hat{U}|A_0,\hat{A}_0}} \|U - \hat{U}\|^2_{L^2(\pi)}]^{\frac{1}{2}} && (29)\\
&\le \|\mathcal{R}\| \mathcal{W}_2(\mu_A, \mu_{\hat{A}}) + [\mathbb{E}_{A_0, \hat{A}_0 \sim \pi_A^*} \mathcal{W}_2^2(\mu_{\hat{U}|A_0}, \mu_{\hat{U}|\hat{A}_0})]^{\frac{1}{2}} && (30)\\
&\le \|\mathcal{R}\| \mathcal{W}_2(\mu_A, \mu_{\hat{A}}) + L_{\text{gen}(\theta)} [\mathbb{E}_{A_0, \hat{A}_0 \sim \pi_A^*} \|A_0 - \hat{A}_0\|^2]^{\frac{1}{2}} && (31)\\
&= (\|\mathcal{R}\| + L_{\text{gen}(\theta)}) \mathcal{W}_2(\mu_A, \mu_{\hat{A}}), && (32)
\end{aligned}$$

where the fourth inequality is Proposition 3. Therefore,

$$\mathcal{W}_2(\mu_F, \mu_{\hat{F}}) \le (\|\mathcal{R}\| + L_{\text{gen}(\theta)}) \mathcal{W}_2(\mu_A, \mu_{\hat{A}}) + [\mathbb{E}_{A_0 \sim \mu_A} \mathcal{W}_2^2(\mu_{U_0|A_0}, \mu_{\hat{U}|A_0})]^{\frac{1}{2}}.$$

□

D.5 Proof of Theorem 5

Proof. Let $Y = U_0(q)$, $Z_I = U_t(I)$, $Z_J = U_t(J)$ and $Z = (Z_I, Z_J)$, so that d^*_{full} conditions on (A_0, Z) and d^*_r on (A_0, Z_I). By Assumption 7, $\Sigma \succeq \sigma_t^2 C_U$ is invertible. Joint Gaussianity gives $d^*_{\text{full}} = \mathbb{E}[Y \mid A_0] + \sigma_q^T \Sigma^{-1}(Z - \mathbb{E}[Z \mid A_0])$, thus $\text{Var}(d^*_{\text{full}} \mid A_0) = \sigma_q^T \Sigma^{-1} \Sigma \Sigma^{-1} \sigma_q = \sigma_q^T \Sigma^{-1} \sigma_q$. Similarly, $d^*_r = \mathbb{E}[Y \mid A_0] + \sigma_I^T \Sigma_{II}^{-1}(Z_I - \mathbb{E}[Z_I \mid A_0])$ and $\text{Var}(d^*_r \mid A_0) = \sigma_I^T \Sigma_{II}^{-1} \sigma_I$. By nesting, $d^*_{\text{full}} - d^*_r \perp d^*_r$ given A_0, so

$$\mathbb{E}[(d^*_{\text{full}} - d^*_r)^2 \mid A_0] = \text{Var}(d^*_{\text{full}} \mid A_0) - \text{Var}(d^*_r \mid A_0) = \sigma_q^T \Sigma^{-1} \sigma_q - \sigma_I^T \Sigma_{II}^{-1} \sigma_I.$$

Let $S = \Sigma_{JJ} - \Sigma_{JI} \Sigma_{II}^{-1} \Sigma_{IJ}$ be the Schur complement, then the Schur decomposition of Σ is

$$\Sigma = \begin{bmatrix} I & 0 \\ \Sigma_{JI} \Sigma_{II}^{-1} & I \end{bmatrix} \begin{bmatrix} \Sigma_{II} & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & \Sigma_{II}^{-1} \Sigma_{IJ} \\ 0 & I \end{bmatrix}.$$

Therefore

$$\Sigma^{-1} = \begin{bmatrix} I & -\Sigma_{II}^{-1}\Sigma_{IJ} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{II}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{JI}\Sigma_{II}^{-1} & I \end{bmatrix}.$$

Since

$$\begin{bmatrix} I & 0 \\ -\Sigma_{JI}\Sigma_{II}^{-1} & I \end{bmatrix} \begin{bmatrix} \sigma_I \\ \sigma_J \end{bmatrix} = \begin{bmatrix} \sigma_I \\ \sigma_J - \Sigma_{JI}\Sigma_{II}^{-1}\sigma_I \end{bmatrix} = \begin{bmatrix} \sigma_I \\ w \end{bmatrix},$$

we have

$$\sigma_q^T \Sigma^{-1} \sigma_q = \begin{bmatrix} \sigma_I \\ w \end{bmatrix}^T \begin{bmatrix} \Sigma_{II}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} \sigma_I \\ w \end{bmatrix} = \sigma_I^T \Sigma_{II}^{-1} \sigma_I + w^T S^{-1} w.$$

Hence $\mathbb{E}[(d^*_{\text{full}} - d^*_r)^2 \mid A_0] = w^T S^{-1} w$, and averaging over A_0 gives the unconditional form. □

D.6 Proof of Theorem 6

Proof. By Theorem 5, $\mathbb{E}[(d^*_{\text{full}} - d^*_r)^2 \mid A_0] = w^T S^{-1} w$, with $S = \Sigma_{JJ} - \Sigma_{JI}\Sigma_{II}^{-1}\Sigma_{IJ}$ and $w = \sigma_J - \Sigma_{JI}\Sigma_{II}^{-1}\sigma_I$. Since U_1 is independent of U_0, $\Sigma = \alpha_t^2 \mathrm{Cov}(U_0(I \cup J) \mid A_0) + \sigma_t^2 C_U \succeq \sigma_t^2 C_U$ on $I \cup J = \{q\} \cup X_m$. With $S^{-1} = (\Sigma^{-1})_{JJ}$ and Assumption 7,

$$\begin{aligned} w^T S^{-1} w &\le \frac{\|w\|^2}{\lambda_{\min}(S)} = \|w\|^2 \lambda_{\max}(S^{-1}) = \|w\|^2 \lambda_{\max}((\Sigma^{-1})_{JJ}) \\ &\le \|w\|^2 \lambda_{\max}(\Sigma^{-1}) = \frac{\|w\|^2}{\lambda_{\min}(\Sigma)} \le \frac{\|w\|^2}{\sigma_t^2 \lambda_{\min}(C_U)}. \end{aligned}$$

Next we bound $\|w\|$. Let $\bar{u} := \Sigma_{II}^{-1}\sigma_I$ be the local kriging weight vector. Note that $\|w\|^2 = \sum_{x \in J} |w_x|^2$ with

$$w_x = \sigma_x - \Sigma_{xI}\Sigma_{II}^{-1}\sigma_I = \mathrm{Cov}(U_t(x), U_0(q) - d^*_r \mid A_0), \quad x \in J.$$

By Assumption 1, $|\mathrm{Cov}(U_t(x), U_t(y) \mid A_0)| \le \alpha_t^2 |\mathrm{Cov}(U_0(x), U_0(y) \mid A_0)| + \sigma_t^2 |C_U(x,y)| \lesssim \rho_0(\|x-y\|)$, since $\alpha_t^2 + \sigma_t^2$ is bounded.

Take $x \in J$ and set $D = \|x - q\| > r$. The first term of w_x is $\sigma_x = \mathrm{Cov}(U_t(x), U_0(q) \mid A_0) = \alpha_t \mathrm{Cov}(U_0(x), U_0(q) \mid A_0)$, so $|\sigma_x| \le |\alpha_t| c_0 e^{-D/l_c} \le |\alpha_t| c_0 e^{-D/(2l_c)}$. To bound the second term $\Sigma_{xI}\bar{u}$, set $s_D = \min\{D/2, r\}$ and split the local points into

$$I_{\text{near}}(x) = \{y \in I : \|y - q\| \le s_D\}, \qquad I_{\text{tail}}(x) = \{y \in I : \|y - q\| > s_D\}.$$

Accordingly we have

$$\Sigma_{xI}\bar{u} = \sum_{y \in I_{\text{near}}(x)} \mathrm{Cov}(U_t(x), U_t(y) \mid A_0)\bar{u}_y + \sum_{y \in I_{\text{tail}}(x)} \mathrm{Cov}(U_t(x), U_t(y) \mid A_0)\bar{u}_y.$$

For $y \in I_{\text{near}}(x)$, we have $\|x - y\| \ge D/2$, so by Assumption 3 with $s = 0$,

$$\Big| \sum_{y \in I_{\text{near}}(x)} \mathrm{Cov}(U_t(x), U_t(y) \mid A_0)\bar{u}_y \Big| \lesssim \rho_0(D/2)\, T(0) \le \Lambda |\alpha_t| c_0 e^{-D/(2l_c)}.$$

To bound the tail mass, if $D > 2r$, then $s_D = r$ and every $y \in I$ satisfies $\|y - q\| \le r$, hence $I_{\text{tail}}(x) = \emptyset$, so it suffices to consider only $r < D \le 2r$. In this case $s_D = D/2$. Points in $I_{\text{tail}}(x)$ can be arbitrarily close to x, so covariance decay alone does not give a valid bound. Instead, Assumption 3 gives

$$\Big| \sum_{y \in I_{\text{tail}}(x)} \mathrm{Cov}(U_t(x), U_t(y) \mid A_0)\bar{u}_y \Big| \lesssim c_0 \sum_{\|y-q\| > D/2} |\bar{u}_y| \le c_0 T(D/2) \le \Lambda |\alpha_t| c_0 e^{-D/(2l_c)}.$$

Combining terms together,

$$\begin{aligned} |w_x| &\le |\sigma_x| + \Big| \sum_{y \in I} \mathrm{Cov}(U_t(x), U_t(y) | A_0)\bar{u}_y \Big| && (33) \\ &\lesssim |\alpha_t| c_0 e^{-D/(2l_c)} + 2\Lambda |\alpha_t| c_0 e^{-D/(2l_c)} && (34) \\ &\lesssim (1 + \Lambda) |\alpha_t| c_0 e^{-D/(2l_c)}. && (35) \end{aligned}$$

Thus $|w_x|^2 \lesssim (1+\Lambda)^2 \alpha_t^2 c_0^2 e^{-D/l_c}$. Summing over $x \in J$,

$$\|w\|_2^2 \lesssim (1+\Lambda)^2 \alpha_t^2 c_0^2 \sum_{x \in J} e^{-\|x-q\|/l_c}.$$

The sum is compared with an integral through quasi-uniformity. The separation distance q_X and the fill distance of X_m are both of order $(|\chi|/m)^{1/d}$, and the fill distance is at most l_c, so $q_X \lesssim l_c$. The balls $B(x, q_X/2)$, $x \in J$, are disjoint and lie in $\{y : \|y-q\| \geq r - q_X/2\}$, and $\|y-q\| \leq \|x-q\| + q_X/2$ on $B(x, q_X/2)$. Averaging $e^{-\|x-q\|/l_c} \leq e^{q_X/(2l_c)} e^{-\|y-q\|/l_c}$ over each ball and substituting $u = \|y-q\| + q_X/2$ gives

$$\sum_{x \in J} e^{-\|x-q\|/l_c} \leq \frac{e^{q_X/l_c}}{|B(0, q_X/2)|} \int_{\|y-q\| \geq r - q_X/2} e^{-(\|y-q\|+q_X/2)/l_c}\, dy \lesssim \frac{m}{|\chi|} \int_r^\infty e^{-u/l_c} u^{d-1} du.$$

For $r \geq l_c$, the exponential tail integral satisfies

$$\int_r^\infty e^{-u/l_c} u^{d-1} du \lesssim l_c r^{d-1} e^{-r/l_c}.$$

Substituting the integral gives

$$\|w\|_2^2 \lesssim (1+\Lambda)^2 \alpha_t^2 c_0^2 \frac{m l_c}{|\chi|} r^{d-1} e^{-r/l_c}.$$

Therefore,

$$\mathbb{E}[(d^*_{\text{full}} - d^*_r)^2 \mid A_0] \lesssim \frac{\alpha_t^2 (1+\Lambda)^2}{\sigma_t^2} \cdot \frac{c_0^2 m l_c r^{d-1}}{\lambda_{\min}(C_U)|\chi|} e^{-r/l_c}.$$

The bound does not depend on A_0, so it also holds for $\mathbb{E}[d^*_{\text{full}} - d^*_r]^2$. □

D.7 Proof of Proposition 7

Proof. (i) Given $A_0 = a$, U_0 is Gaussian and $U_1 \sim \nu_U = \mathcal{N}(0, C_U)$ is an independent centered Gaussian (Appendix A.2), so (U_0, U_1) and its linear image (U_0, U_t) are jointly Gaussian. Since $\mathcal{H}$ is an RKHS, point evaluation is continuous on $\mathcal{H}$, so $(U_0(q), U_t(q), U_t(X_m))$ is jointly Gaussian given $A_0 = a$ for every $q \in \chi$ and $t > 0$, which is Assumption 2.

(ii) This is trivial. □

D.8 Proof of Proposition 8

Proof. Write $X = U_t(N)$, $Y = U_0(q)$, $Z = U_t(q)$ and $v = (Z, X^T)^T$, centered given A_0 as in Appendix C.3. Let $\Sigma_v = \mathrm{Cov}(v \mid A_0) = \begin{pmatrix} s_{qq} & s_{qN} \\ s_{Nq} & \Sigma_N \end{pmatrix}$ and $T = \begin{pmatrix} 1 & 0 \\ 0 & \mathcal{L} \end{pmatrix}$, then the local statistic together with the query value is

$$Tv = \begin{pmatrix} Z \\ \mathcal{L}X \end{pmatrix}.$$

By Lemma 1,

$$\mathrm{ran}(\mathcal{L}^T) = \{W_h \varphi : \varphi \in \mathcal{P}_p\}.$$

Observe that $d^*_r = \bar{w}^T v$, with $\bar{w} = (w_q, w_N^T)^T$ from Proposition 13, is the projection of $Y = U_0(q)$ onto the Gaussian linear space generated by v, while the compressed predictor d^*_{local} is the projection onto the smaller subspace generated by Tv. Therefore, the compression error is the squared covariance induced distance from the raw regression coefficient to the subspace of regression coefficients representable by the compressed statistic,

$$\Delta_{\text{comp}} = \min_{z \in \mathrm{ran}(T^T)} \|\bar{w} - z\|^2_{\Sigma_v}.$$

For any $(c, \mathcal{L}^T a) \in \mathrm{ran}(T^T)$, let $\delta_q = w_q - c$, $\delta_N = w_N - \mathcal{L}^T a$, then

$$\min_{z \in \mathrm{ran}(T^T)} \|\bar{w} - z\|^2_{\Sigma_v} = \min_{\delta_q, \delta_N} [s_{qq}\delta_q^2 + 2\delta_q s_{qN}\delta_N + \delta_N^T \Sigma_N \delta_N] \tag{36}$$

$$= \min_{\delta_N} \delta_N^T \Big(\Sigma_N - \frac{s_{Nq}s_{qN}}{s_{qq}}\Big)\delta_N \tag{37}$$

$$= \min_{a \in \mathbb{R}^{D_p}} \|w_N - \mathcal{L}^T a\|^2_{\Sigma_{N|q}}, \tag{38}$$

since c is free and $\Sigma_{N|q} = \Sigma_N - s_{Nq}s_{qN}/s_{qq}$. Since $\{\mathcal{L}^T a : a \in \mathbb{R}^{D_p}\} = \{W_h\varphi : \varphi \in \mathcal{P}_p\}$, we have

$$\Delta_{\text{comp}} = \min_{\varphi \in \mathcal{P}_p} \|w_N - W_h\varphi\|^2_{\Sigma_{N|q}}.$$

□

D.9 Proof of Proposition 9

Proof. The fitted value on N with local regression is $\hat{X} = Vl(U_t; q)$. Let $\hat{d} = w_q Z + w_N^T \hat{X}$, then it is a predictor measurable with respect to (A_0, Z, l). Let $\Delta_X = X - \hat{X}$, by optimality of d^*_{local},

$$\Delta_{\text{comp}} = \mathbb{E}[d_r^* - d^*_{\text{local}}]^2 \le \mathbb{E}[d_r^* - \hat{d}]^2 = \mathbb{E}|w_N^T \Delta_X|^2 \le \|w_N\|_2^2 \mathbb{E}\|\Delta_X\|_2^2.$$

By discrete quasi-optimality

$$\|\Delta_X\|_2^2 \le C_{\text{qo}} \inf_{\varphi \in \mathcal{P}_p} \sum_{x_i \in N} |U_t(x_i) - \varphi(x_i)|^2.$$

A sampling inequality converts the discrete design residual to the continuous one: for any $\varphi \in \mathcal{P}_p$,

$$\frac{1}{n_h} \sum_{x_i \in N} |U_t(x_i) - \varphi(x_i)|^2 \le C_{\text{samp}} \frac{\|U_t - \varphi\|^2_{L^2(B(q,h))}}{|B(q,h)|}.$$

Since $U_t \in \mathcal{H} = H^s(\chi)$, the Bramble-Hilbert Lemma gives

$$\inf_{\varphi \in \mathcal{P}_p} \|U_t - \varphi\|_{L^2(B(q,h))} \le C_{d,p} h^k |U_t|_{H^k(B(q,h))}, \quad k \le \min\{p+1, s\}.$$

Combining the three inequalities gives

$$\|\Delta_X\|_2^2 \le C n_h h^{2k} \frac{|U_t|^2_{H^k(B(q,h))}}{|B(q,h)|}.$$

Since U_t restricted to $B(q,h)$ is a stationary random field with spectral density $\widehat{C}_t(\xi)$ and finite spectral moment $S_k(t) = \int_{\mathbb{R}^d} |\xi|^{2k} \widehat{C}_t(\xi) d\xi$, we have $\mathbb{E}|U_t|^2_{H^k(B(q,h))} \le C|B(q,h)|S_k(t)$. Taking expectation gives

$$\mathbb{E}\|\Delta_X\|_2^2 \le C n_h h^{2k} S_k(t),$$

and therefore

$$\Delta_{\text{comp}} \le C_{d,k} \|w_N\|_2^2 \, n_h h^{2k} S_k(t),$$

with $C_{d,k}$ depending on d, k, C_{qo} and C_{samp} only. □

D.10 Proof of Theorem 10

We prove the following precise form of Theorem 10. Let $N_* = N \min\{|\chi|/(r + l_c)^d, |Q|\}$ as in Assumption 8. Fix a quasi-uniform design X_m with fill distance at most l_c, a radius $r = h \ge l_c$, a degree p and a stopping time $\epsilon > 0$. Suppose Assumptions 1–8 hold with $\rho_0(u) = c_0 e^{-u/l_c}$, and the hypotheses of Proposition 9 hold at every $q \in \chi$ and $t \in [\epsilon, 1]$ with $\sup_t S_k(t) < \infty$. Then for any $\delta > 0$,

$$\begin{aligned}\mathbb{E}\mathcal{W}_2(\mu_F, \mu_{\hat{F}}) \lesssim (\|\mathcal{R}\| + L_{\text{gen}(\theta)}) N^{-\frac{\theta+1-\delta}{2(2\theta+L)}} + N_*^{-\frac{\theta'}{2\theta'+L+D_p+d+1}} + N^{-\frac{\theta'}{2\theta'+L}} \\ + m^{-\frac{s}{d}} + \Big(\frac{m}{\lambda_{\min}(C_U)}\Big)^{\frac{1}{2}} r^{\frac{d-1}{2}} e^{-\frac{r}{2l_c}} + \sqrt{m}\, r^{k+\frac{d}{2}} + \epsilon_{\text{stop}},\end{aligned} \tag{39}$$

where $\mathbb{E}$ is over the training set and the constant does not depend on N, m or r.

Proof. Throughout, $\lesssim$ hides constants that may depend on d, k, δ, λ, $|\chi|$, ϵ, the schedule, Λ, c_0, l_c, C_{qo}, C_{samp} and $\sup_{t\in[\epsilon,1]} S_k(t)$, but not on N, m or r. Assumptions 4 and 5 hold for every trained network, so $L_{\mathrm{gen}(\theta)}$ and g_ϵ are deterministic, and $\mathbb{E}$ denotes the expectation over the training set. Write $G_\epsilon = \int_\epsilon^1 g_\epsilon(t)\,dt$, which is finite by Assumption 6.

Step 1: structure. Theorems 4 and 2 give

$$\mathcal{W}_2(\mu_F, \mu_{\hat{F}}) \le (\|\mathcal{R}\| + L_{\mathrm{gen}(\theta)})\mathcal{W}_2(\mu_A, \mu_{\hat{A}}) + C_\lambda \int_\epsilon^1 g_\epsilon(t)\eta_{\mathrm{mix}}(t)\,dt + \epsilon_{\mathrm{stop}}. \tag{40}$$

Theorem 1 holds at every (t, q). Averaging its first identity over $q \sim \bar{\pi}_m$, and writing a bar for this average, gives $\eta_{\mathrm{mix}}(t)^2 = \bar{\Delta}_{\mathrm{disc}}(t) + \bar{\Delta}_{\mathrm{loc}}(t) + \bar{\Delta}_{\mathrm{comp}}(t) + \bar{\Delta}_{\mathrm{learn}}(t)$, and since $\sqrt{a+b+c+d} \le \sqrt{a} + \sqrt{b} + \sqrt{c} + \sqrt{d}$ for nonnegative numbers,

$$\eta_{\mathrm{mix}}(t) \le \bar{\Delta}_{\mathrm{disc}}^{1/2}(t) + \bar{\Delta}_{\mathrm{loc}}^{1/2}(t) + \bar{\Delta}_{\mathrm{comp}}^{1/2}(t) + \bar{\Delta}_{\mathrm{learn}}^{1/2}(t). \tag{41}$$

The first three terms do not depend on the trained network. We bound each term uniformly in $t \in [\epsilon, 1]$ and integrate against g_ϵ.

Step 2: discretization. By Assumption 8(iii), $\bar{\Delta}_{\mathrm{disc}}^{1/2}(t) \lesssim m^{-s/d}$ uniformly in t, so

$$\int_\epsilon^1 g_\epsilon \bar{\Delta}_{\mathrm{disc}}^{1/2}\,dt \lesssim G_\epsilon m^{-s/d}.$$

Step 3: localization. Theorem 6 holds at every q with the same constants, so

$$\bar{\Delta}_{\mathrm{loc}}^{1/2}(t) \lesssim \frac{|\alpha_t|}{\sigma_t}(1+\Lambda)c_0\Big(\frac{m\,l_c}{\lambda_{\min}(C_U)|\chi|}\Big)^{1/2} r^{\frac{d-1}{2}} e^{-\frac{r}{2l_c}}.$$

By Assumption 6, α_t and σ_t are continuous on $[\epsilon, 1]$ with $\sigma_t > 0$, so $M_\epsilon = \sup_{t\in[\epsilon,1]} |\alpha_t|/\sigma_t$ is finite, and

$$\int_\epsilon^1 g_\epsilon \bar{\Delta}_{\mathrm{loc}}^{1/2}\,dt \lesssim G_\epsilon M_\epsilon \Big(\frac{m}{\lambda_{\min}(C_U)}\Big)^{1/2} r^{\frac{d-1}{2}} e^{-\frac{r}{2l_c}}.$$

Step 4: compression. Proposition 9 with $h = r$ gives $\Delta_{\mathrm{comp}} \le C_{d,k}\|w_N\|_2^2\, n_h\, r^{2k} S_k(t)$ at every q. Here w_N is the restriction to N of the kriging vector $\bar{u} = \Sigma_{II}^{-1}\sigma_I$ (Proposition 13), so Assumption 3 with $s = 0$ gives $\|w_N\|_2 \le \|w_N\|_1 \le T(0) \le \Lambda|\alpha_t|$. Since $r \ge l_c$ is at least the fill distance, quasi-uniformity gives $n_h \lesssim (m/|\chi|)r^d$. Hence

$$\bar{\Delta}_{\mathrm{comp}}^{1/2}(t) \lesssim \Lambda|\alpha_t|\, S_k(t)^{1/2}\Big(\frac{m}{|\chi|}\Big)^{1/2} r^{k+\frac{d}{2}}, \qquad \int_\epsilon^1 g_\epsilon \bar{\Delta}_{\mathrm{comp}}^{1/2}\,dt \lesssim G_\epsilon\,\sqrt{m}\, r^{k+\frac{d}{2}},$$

using $\sup_t |\alpha_t| < \infty$ and $\sup_t S_k(t) < \infty$.

Step 5: learning. By Tonelli's theorem and Jensen's inequality,

$$\begin{aligned}\mathbb{E}\int_\epsilon^1 g_\epsilon(t)\bar{\Delta}_{\mathrm{learn}}^{1/2}(t)\,dt &= \int_\epsilon^1 g_\epsilon(t)\,\mathbb{E}\big[\bar{\Delta}_{\mathrm{learn}}^{1/2}(t)\big]dt \le \int_\epsilon^1 g_\epsilon(t)\big[\mathbb{E}\bar{\Delta}_{\mathrm{learn}}(t)\big]^{1/2}dt \\ &\lesssim G_\epsilon\Big(N_*^{-\frac{\theta'}{2\theta'+L+D_p+d+1}} + N^{-\frac{\theta'}{2\theta'+L}}\Big),\end{aligned}$$

by Assumption 8(ii) and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$.

Step 6: global stage. By Assumption 8(i) and Oko et al. (2023), for any $\delta > 0$,

$$\mathbb{E}\,\mathcal{W}_1(\mu_A, \mu_{\hat{A}}) \lesssim N^{-\frac{\theta+1-\delta}{2\theta+L}}.$$

On $[-1,1]^L$ every coupling satisfies $\|a - a'\| \le 2\sqrt{L}$, so $\mathcal{W}_2^2 \le 2\sqrt{L}\,\mathcal{W}_1$, and Jensen's inequality gives

$$\mathbb{E}\mathcal{W}_2(\mu_A, \mu_{\hat{A}}) \le \big(2\sqrt{L}\,\mathbb{E}\,\mathcal{W}_1(\mu_A, \mu_{\hat{A}})\big)^{1/2} \lesssim N^{-\frac{\theta+1-\delta}{2(2\theta+L)}}.$$

Taking the expectation of (40), inserting (41) and Steps 2 to 6 gives

$$\begin{aligned}\mathbb{E}\mathcal{W}_2(\mu_F, \mu_{\hat{F}}) \lesssim{}& (\|\mathcal{R}\| + L_{\mathrm{gen}(\theta)})N^{-\frac{\theta+1-\delta}{2(2\theta+L)}} + N_*^{-\frac{\theta'}{2\theta'+L+D_p+d+1}} + N^{-\frac{\theta'}{2\theta'+L}} \\ &+ m^{-\frac{s}{d}} + \Big(\frac{m}{\lambda_{\min}(C_U)}\Big)^{\frac{1}{2}} r^{\frac{d-1}{2}} e^{-\frac{r}{2l_c}} + \sqrt{m}\, r^{k+\frac{d}{2}} + \epsilon_{\mathrm{stop}},\end{aligned}$$

which is (39). Only the first three terms depend on N. □

D.11 Proof of Proposition 11

Proof. Write $Y = U_0(q)$, $C = (A_0, U_t(I))$, $V = U_t(J)$ and $\mathcal{G}_C = \sigma(C)$. Then $d_r^* = \mathbb{E}[Y \mid \mathcal{G}_C]$ and $d_{\text{full}}^* = \mathbb{E}[Y \mid V, C]$. Let $Z = Y - \mathbb{E}[Y \mid \mathcal{G}_C]$, then $\mathbb{E}[Z \mid \mathcal{G}_C] = 0$. Since $\mathbb{E}[Y \mid \mathcal{G}_C]$ is $\sigma(V, C)$-measurable, $d_{\text{full}}^* - d_r^* = \mathbb{E}[Z \mid V, C] =: W$. By the tower property $\mathbb{E}[W \mid \mathcal{G}_C] = \mathbb{E}[Z \mid \mathcal{G}_C] = 0$, so $\mathbb{E}[W^2 \mid \mathcal{G}_C] = \mathrm{Var}(W \mid \mathcal{G}_C)$. Note that $W = \mathbb{E}[Y \mid V, C] - \mathbb{E}[Y \mid \mathcal{G}_C]$ differs from $\mathbb{E}[Y \mid V, C]$ by a $\mathcal{G}_C$-measurable term, which does not change the conditional variance, so $\mathrm{Var}(W \mid \mathcal{G}_C) = \mathrm{Var}(\mathbb{E}[Y \mid V, C] \mid \mathcal{G}_C)$. Therefore, $\mathbb{E}[(d_{\text{full}}^* - d_r^*)^2 \mid \mathcal{G}_C] = \mathrm{Var}(\mathbb{E}[Y \mid V, C] \mid \mathcal{G}_C)$. Integrating over C and using the definition of $\bar{\eta}(r)$, we have

$$\mathbb{E}[d_{\text{full}}^* - d_r^*]^2 = \mathbb{E}[\mathrm{Var}(\mathbb{E}[Y \mid V, C] \mid \mathcal{G}_C)] = \bar{\eta}^2(r)\, \mathbb{E}[\mathrm{Var}(Y \mid C)].$$

Write $V_r = \mathbb{E}[\mathrm{Var}(Y \mid \mathcal{G}_C)]$. By the law of total variance $\mathrm{Var}(Y \mid \mathcal{G}_C) = \mathbb{E}[\mathrm{Var}(Y \mid V, C) \mid \mathcal{G}_C] + \mathrm{Var}(\mathbb{E}[Y \mid V, C] \mid \mathcal{G}_C)$, so $V_r = V_{\text{full}} + \mathbb{E}[d_{\text{full}}^* - d_r^*]^2$. Combining the two identities, $\mathbb{E}[d_{\text{full}}^* - d_r^*]^2 = \bar{\eta}^2(r)\big(V_{\text{full}} + \mathbb{E}[d_{\text{full}}^* - d_r^*]^2\big)$, and since $\bar{\eta}(r) < 1$,

$$\mathbb{E}[d_{\text{full}}^* - d_r^*]^2 = \frac{\bar{\eta}^2(r)}{1 - \bar{\eta}^2(r)} V_{\text{full}}.$$

□

D.12 Proof of Proposition 12

Proof. The proof of Theorem 6 uses the envelope only through $\rho_0(D/2)$ and the final sum. Repeating it with $\rho_0(u) = c_0(1 + u/l_c)^{-a}$, the same three bounds give

$$|w_x| \lesssim (1 + \Lambda)|\alpha_t|\, \rho_0(D/2) \le (1 + \Lambda)|\alpha_t|\, c_0(2l_c)^a D^{-a},$$

since $1 + D/(2l_c) \ge D/(2l_c)$. Sum over $x \in J$ and compare with an integral as in Theorem 6. On $B(x, q_X/2)$ we have $\|y - q\| \le \|x - q\| + q_X/2 \le 2\|x - q\|$, so $D^{-2a} \le 2^{2a}\|y - q\|^{-2a}$ there, and the balls lie in $\{\|y - q\| \ge r/2\}$. Hence

$$\|w\|_2^2 \lesssim (1 + \Lambda)^2 \alpha_t^2 c_0^2 (2l_c)^{2a} \frac{m}{|\chi|} \int_{r/2}^{\infty} u^{d-1-2a}\, du = (1 + \Lambda)^2 \alpha_t^2 c_0^2 (2l_c)^{2a} \frac{m}{|\chi|} \cdot \frac{2^{2a-d} r^{d-2a}}{2a - d},$$

which is finite since $a > d/2$. Dividing by $\sigma_t^2 \lambda_{\min}(C_U)$ as in Theorem 6 gives

$$\mathbb{E}[(d_{\text{full}}^* - d_r^*)^2 \mid A_0] \lesssim \frac{\alpha_t^2 (1 + \Lambda)^2}{\sigma_t^2} \cdot \frac{c_0^2 (2l_c)^{2a}\, m}{(2a - d)\lambda_{\min}(C_U)|\chi|}\, r^{d-2a},$$

which, after averaging over A_0, is the stated bound with the constants made explicit. □

D.13 Proof of Proposition 13

Proof. Write $X = U_t(N)$, $Y = U_0(q)$, $Z = U_t(q)$ and $v = \begin{pmatrix} Z \\ X \end{pmatrix}$, centered given $A_0 = a$. Under conditional joint Gaussianity (Assumption 2),

$$d_r^* = w_q Z + w_N^T X = \mathbb{E}[Y \mid v, A_0 = a] = \mathrm{Cov}(Y, v \mid a)\mathrm{Cov}(v \mid a)^{-1} v.$$

With

$$\mathrm{Cov}(v \mid a) = \begin{pmatrix} s_{qq} & s_{qN} \\ s_{Nq} & \Sigma_N \end{pmatrix},$$

define the Schur complement $V_q = s_{qq} - s_{qN}\Sigma_N^{-1} s_{Nq}$. By Assumption 7, $\mathrm{Cov}(v \mid a) \succeq \sigma_t^2 C_U$ is positive definite, so Σ_N and V_q are invertible, and the block inverse is

$$\begin{pmatrix} s_{qq} & s_{qN} \\ s_{Nq} & \Sigma_N \end{pmatrix}^{-1} = \begin{pmatrix} V_q^{-1} & -V_q^{-1} s_{qN} \Sigma_N^{-1} \\ -\Sigma_N^{-1} s_{Nq} V_q^{-1} & \Sigma_N^{-1} + \Sigma_N^{-1} s_{Nq} V_q^{-1} s_{qN} \Sigma_N^{-1} \end{pmatrix}.$$

The regression coefficient vector is

$$\begin{pmatrix} w_q \\ w_N \end{pmatrix} = \mathrm{Cov}(v \mid a)^{-1} \begin{pmatrix} \gamma_q \\ \sigma_N \end{pmatrix},$$

so

$$w_q = V_q^{-1}\gamma_q - V_q^{-1}s_{qN}\Sigma_N^{-1}\sigma_N = \frac{\gamma_q - s_{qN}\Sigma_N^{-1}\sigma_N}{s_{qq} - s_{qN}\Sigma_N^{-1}s_{Nq}},$$

and

$$w_N = -\Sigma_N^{-1}s_{Nq}V_q^{-1}\gamma_q + (\Sigma_N^{-1} + \Sigma_N^{-1}s_{Nq}V_q^{-1}s_{qN}\Sigma_N^{-1})\sigma_N = \Sigma_N^{-1}\sigma_N - \Sigma_N^{-1}s_{Nq}w_q.$$

□