DIFFUSION MODELS BREAK THE CURSE OF DIMENSIONALITY FOR MIXED-SMOOTH BESOV DENSITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models learn a probability law by adding noise, learning how to denoise, and then running the noising process backward. In high dimensions, classical nonparametric theory predicts a curse: the sample-size exponent worsens with the ambient dimension. We show that this curse can be provably broken when the density has mixed smoothness, meaning that simultaneous high-frequency variation across many coordinates is strongly limited, the structure exploited by sparse grids. On the torus, for densities in a mixed Besov ball, we prove a total-variation minimax rate with the one-dimensional polynomial exponent $n^{-s/(2s+1)}$ and only logarithmic dependence on the dimension. We then show that a diffusion estimator using a sparse-grid score approximator attains this rate, up to standard algorithmic polylogarithms, under explicit score regularity and multiplier assumptions; for $s > 1/2$ the multiplier condition holds on the whole class, so the rate is unconditional there. A controlled study reproduces the predicted scaling. Our analysis also reveals a new Wasserstein geometry: transport is governed by the finest coordinate scale, leaving sharp diffusion W_1 rates an important open direction.

1 INTRODUCTION

Score-based diffusion models (Ho et al., 2020; Song et al., 2021; Karras et al., 2022) have become a dominant paradigm for learning high-dimensional distributions from i.i.d. samples, generating strikingly high-fidelity outputs across image, audio, and scientific domains. From a statistical perspective, a central question is whether such models are *sample-efficient distribution estimators* under *strong* losses such as total variation (TV) and Wasserstein-1 (W_1) (Villani, 2009).

Oko et al. (2023) placed diffusion within classical minimax theory by showing it is *minimax optimal* over isotropic Besov-type classes. But isotropic smoothness inherits the *curse of dimensionality*: rates scale like $n^{-s/(2s+d)}$ (Stone, 1982; Tsybakov, 2009), requiring astronomical samples at moderate d. This creates a tension: diffusion succeeds empirically in regimes where worst-case isotropic theory predicts severe statistical hardness.

One way forward is to identify additional structure beyond isotropic smoothness that can fundamentally change statistical complexity. A well-developed but underutilized structural assumption in high-dimensional approximation theory is *dominating mixed smoothness* (Vybiral, 2006; Dung et al., 2018). Unlike isotropic smoothness—which measures regularity uniformly in all directions—mixed smoothness captures the idea that the object is smooth in *coordinate-wise interactions*, allowing high-dimensional functions and densities to remain approximable without paying an exponential price in d. Mixed-smooth function classes underpin sparse-grid and hyperbolic-cross methods, where approximation error depends on the dimension only *weakly* (often through polylogarithmic factors) (Smolyak, 1963; Zenger, 1990; Bungartz & Griebel, 2004; Barthelmann et al., 2000). The assumption is natural when coordinates are meaningful variables with limited high-order coupling: product densities, their mixtures and densities built from smooth low-order interactions are typical members (Section 2). In learning theory, mixed-smooth Besov structure has also been shown to mitigate the curse for deep networks in supervised settings (Suzuki, 2019). These developments suggest an intriguing possibility: perhaps diffusion models can inherit the same "curse-breaking" phenomenon when the *target density* admits mixed-smooth regularity.

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.

This paper asks: *can score-based diffusion models break the polynomial curse of dimensionality under mixed smoothness?* We answer yes for total variation loss. Building on the minimax diffusion framework of Oko et al. (2023), we introduce a *mixed-smooth Besov* density model and show that the minimax rate has the one-dimensional polynomial exponent, with d entering only through sparse-grid logarithmic powers and constants. A sparse-grid (Smolyak) score parameterization (Smolyak, 1963; Bungartz & Griebel, 2004) attains this TV rate, up to algorithmic polylogarithms, under an explicit score/multiplier regularity condition. For W_1, we retain the modular score-to-distribution interface but separate matching upper bounds from the mixed transport lower-bound geometry.

Positioning. The table lists the closest results with the loss for each row. Oko et al. (2023, Section 6) already pointed to mixed or anisotropic smoothness as a way around the curse; a quantitative diffusion theorem further needs a lower bound with the mixed logarithmic factor, control of the score $\nabla p_t/p_t$ in the mixed class (closed under multiplication only for $s > 1/2$), and a random-design analysis of sparse-grid score matching. For direct hyperbolic-cross density estimation, the TV-endpoint lower bound of Ding et al. (2021) carries no logarithmic factor; Theorem 1 shows that $(\log n)^{s(d-1)/(2s+1)}$ is necessary, and the diffusion estimator of Theorem 3 attains it. Related end-to-end analyses treat other generative dynamics: Zhou & Liu (2025) prove Wasserstein-2 guarantees for flow matching with neural velocity fields, and Wakasugi & Suzuki (2025) bound the KL error of discrete-state diffusion without a polynomial cost in the number of states, via spectral decay and continuous embeddings. We instead treat continuous densities with dominating mixed smoothness at the TV minimax scale; Appendix M.5 compares these works and details the steps beyond Oko et al. (2023).

Reference	Class	Task / estimator	Rate up to logs (loss)
Stone (1982)	isotropic Hölder	regression	$n^{-s/(2s+d)}$ (L^2)
Suzuki (2019)	mixed-smooth Besov	regression, deep ReLU NN	$n^{-s/(2s+1)}$ (L^2)
Ding et al. (2021)	mixed-smooth Sobolev	density, hyperbolic-cross series	$n^{-s/(2s+1)}$ (TV)
Oko et al. (2023)	isotropic Besov	density, diffusion + NN score	$n^{-s/(2s+d)}$ (TV)
This paper	**mixed-smooth Besov**	**density, diffusion + sparse grid**	$n^{-s/(2s+1)}(\log n)^{\frac{s(d-1)}{2s+1}}$ (TV)

1.1 Contributions

- **Mixed-smooth Besov density model.** We formalize dominating mixed smoothness for density estimation on the torus, connecting it to tensor-product wavelets and sparse-grid index sets (Vybiral, 2006; Dung et al., 2018; Bungartz & Griebel, 2004).
- **Curse-breaking TV minimax lower bound** (Theorem 1). Over $B^{s,\mathrm{mix}}_{2,2}$ densities, the TV lower bound is $n^{-s/(2s+1)}(\log n)^{s(d-1)/(2s+1)}$, replacing the isotropic exponent $s/(2s+d)$ by the one-dimensional exponent $s/(2s+1)$.
- **A matching sparse-grid diffusion upper bound** (Theorem 3). Under the explicit score-regularity condition in Appendix F, sparse-grid denoising score matching attains the TV lower bound up to algorithmic polylogarithms. For W_1, we prove a balanced-block lower bound that exposes the anisotropic transport scale and state a conservative upper guarantee; matching rates remain open. The score-regularity condition holds for every $s > 1/2$ (Appendix J.1).
- **A modular proof architecture.** Our upper bound combines a deterministic score-to-distribution reduction (Theorem 2, restating Oko et al. (2023) in module-agnostic form) with new approximation/generalization bounds for sparse-grid score matching—a reusable interface for other score classes.

2 Mixed-Smooth Besov Model and Preliminaries

This section defines the mixed-smooth Besov class and the sparse-grid tools that make *mixed smoothness* operational; diffusion objects (noising, scores, denoising) appear in Section 4.

2.1 Setup and notation

We observe i.i.d. samples $X_1, \dots, X_n \sim p_0$ on the periodic torus $\mathbb{T}^d \equiv (\mathbb{R}/\mathbb{Z})^d \simeq [0,1]^d$ (Dung et al., 2018; Bungartz & Griebel, 2004). We write $\|f\|_{L^p}$ for the $L^p(\mathbb{T}^d)$ norm and $\langle f, g\rangle = \int fg\,dx$

for the L^2 inner product. The model class consists of bounded densities; throughout we assume $0 < \underline{p} < 1 < \overline{p} < \infty$, since otherwise the class is empty or degenerate. Define

$$\mathcal{P}(\underline{p}, \overline{p}) := \Big\{ p \in L^1(\mathbb{T}^d) : p \geq 0,\ \int p\, dx = 1,\ \underline{p} \leq p \leq \overline{p} \text{ a.e.} \Big\}, \tag{1}$$

where $\underline{p} > 0$ keeps $\nabla \log p_t$ regular.

We measure error in total variation $\mathrm{TV}(p, q) := \frac{1}{2}\|p - q\|_{L^1(\mathbb{T}^d)}$ and Wasserstein-1 $W_1(p, q) := \sup_{\mathrm{Lip}(f)\leq 1} \int_{\mathbb{T}^d} f\,(p - q)\, dx$ (Kantorovich–Rubinstein duality (Villani, 2009)), with Lipschitz constant w.r.t. the geodesic distance on $\mathbb{T}^d$.

2.2 Mixed-smooth Besov classes via tensor-product wavelets

We characterize mixed smoothness via wavelets, which reveal both a clean norm structure and the sparse-grid index sets driving curse-breaking rates (Vybiral, 2006; Dung et al., 2018).

Throughout, $\{\psi_{j,k}\}_{j\geq 0}$ denotes a sufficiently regular periodic orthonormal wavelet system on $\mathbb{T}$ with zero mean. To obtain a complete tensor basis, include the constant scaling function by setting $\eta_{-1,0} \equiv 1,\ \eta_{j,k} = \psi_{j,k}\ (j \geq 0)$.

For $j \in \mathbb{N}_{-1}^d := \{-1, 0, 1, \ldots\}^d$, define $|j|_+ = \sum_{m=1}^d \max(j_m, 0)$ and

$$\mathcal{K}_j := \prod_{m=1}^d \mathcal{K}_{j_m}^{(1)}, \qquad \mathcal{K}_{-1}^{(1)} := \{0\}, \quad \mathcal{K}_\ell^{(1)} := \{0, \ldots, 2^\ell - 1\}\ (\ell \geq 0).$$

The tensor functions are $\Psi_{j,k}(x) = \prod_{m=1}^d \eta_{j_m,k_m}(x_m),\ |\mathcal{K}_j| = 2^{|j|_+}$. The $j_m = -1$ factors are essential: they include constants, main effects, and lower-order interactions. Any $f \in L^2(\mathbb{T}^d)$ admits

$$f = \bar{f} + \sum_{j \in \mathbb{N}_{-1}^d \setminus \{-\mathbf{1}\}} \sum_{k \in \mathcal{K}_j} \theta_{j,k}(f) \Psi_{j,k}, \qquad \theta_{j,k}(f) = \langle f, \Psi_{j,k} \rangle. \tag{2}$$

Mixed-smooth Besov norm (wavelet form). For $s > 0$, $B_{2,2}^{s,\mathrm{mix}}(\mathbb{T}^d)$ is characterized by (Vybiral, 2006; Dung et al., 2018; Triebel, 2006)

$$\|f\|_{B_{2,2}^{s,\mathrm{mix}}}^2 := |\bar{f}|^2 + \sum_{j \in \mathbb{N}_{-1}^d \setminus \{-\mathbf{1}\}} 2^{2s|j|_+} \sum_{k \in \mathcal{K}_j} \theta_{j,k}(f)^2 < \infty. \tag{3}$$

The signature is ℓ_1 aggregation through $|j|_+$: refining many coordinates jointly is expensive while inactive coordinates ($j_m = -1$) carry no resolution cost. This is what enables sparse-grid approximation.

Mixed-smooth density class. We define $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p}) := \Big\{ p \in \mathcal{P}(\underline{p}, \overline{p}) \ :\ \|p\|_{B_{2,2}^{s,\mathrm{mix}}} \leq R \Big\}$.

All minimax statements are over $p_0 \in \mathcal{B}^{s,\mathrm{mix}}$; we take $R > 1$ so that the constant density has a nontrivial neighborhood inside the Besov ball.

Examples and scope of the model. (i) A product $\prod_m p_m(x_m)$ has $\|p\|_{B_{2,2}^{s,\mathrm{mix}}} = \prod_m \|p_m\|_{B_{2,2}^s(\mathbb{T})}$: the smoothness index does not deteriorate with d, while the radius is a product of d norms. (ii) The class is convex, so it contains finite mixtures of its members. (iii) $1 + \sum_{|S|\leq q} g_S(x_S)$, with zero-mean mixed-smooth terms in at most q coordinates, is a member of radius at most $1 + \sum_S \|g_S\|_{B_{2,2}^{s,\mathrm{mix}}}$ whenever it satisfies the envelope. (iv) For $s > 1/2$ the space is an algebra (Appendix J), so Gibbs densities with finitely many bounded, positive, mixed-smooth low-order clique potentials are members. The torus removes boundary effects from the wavelets and the heat flow; the lower bound transfers to $[0, 1]^d$ with interior perturbations, whereas an upper bound there needs boundary-adapted sparse grids and a reflected forward process. The envelope $p_0 \geq \underline{p}$ controls $1/p_t$, hence the score, as $t \downarrow 0$ (Oko et al., 2023); in the Hilbert case $p = q = 2$ projection and least squares close cleanly, while $p < 2$ would require nonlinear thresholding. Angular variables such as phases or torsion angles live on the torus. In practice mixed smoothness can be checked only at finite resolution:

Appendix M.3 gives noise-corrected estimates of the shell energies $\sum_{|j|_+=l}\sum_k \theta_{j,k}^2$ with error bars, a plausibility check rather than a membership test. Appendix O applies a finite-frequency analogue to hourly wind-direction records and finds substantial cross-station interactions; temporal dependence, quantization and the frequency cutoff limit its interpretation.

2.3 Sparse-grid (Smolyak) truncation and effective complexity

Sparse-grid truncation chooses which multi-levels j to keep in (2). Isotropic truncation $\|j\|_\infty \le L$ activates $\asymp 2^{Ld}$ basis functions; mixed smoothness instead suggests the hyperbolic-cross (Smolyak) index set $\mathcal{J}_L^{\mathrm{HC}} = \{j \in \mathbb{N}_{-1}^d : |j|_+ \le L\}$, spanning $\mathcal{V}_L^{\mathrm{SG}} := \mathrm{span}\{\Psi_{j,k} : j \in \mathcal{J}_L^{\mathrm{HC}},\ k \in \mathcal{K}_j\}$, with the classical count $\dim(\mathcal{V}_L^{\mathrm{SG}}) \asymp 2^L L^{d-1}$ (Smolyak, 1963; Bungartz & Griebel, 2004).

The exponential factor is 2^L rather than 2^{Ld}; with statistically optimal $L \asymp \log n$, the residual L^{d-1} becomes polylogarithmic in n.

Intuition $d = 2$**.** A 2D tensor wavelet with $j = (3, -1)$ is fine in x_1 and constant in x_2, while $j = (2, 1)$ is a genuine interaction. At $L = 3$, hyperbolic truncation $|j|_+ \le 3$ keeps all main effects and thin rectangles such as $(3, -1)$, $(-1, 3)$, $(2, 1)$, but excludes joint-fine corners such as $(3, 3)$: sharp per-coordinate variation is permitted, joint high-order interactions are suppressed.

Approximation behavior. Let Π_L be the L^2-projector onto $\mathcal{V}_L^{\mathrm{SG}}$. For $f \in B_{2,2}^{s,\mathrm{mix}}$ (Dung et al., 2018; Bungartz & Griebel, 2004), $\|f - \Pi_L f\|_{L^2} \lesssim 2^{-sL}\,\|f\|_{B_{2,2}^{s,\mathrm{mix}}}$.

For the Hilbert ball $B_{2,2}^{s,\mathrm{mix}}$ there is no sharp extra $L^{(d-1)/2}$ bias factor: the logarithms enter through the effective dimension $2^L L^{d-1}$, exponentially smaller than the isotropic 2^{Ld}.

3 Minimax Lower Bounds under Mixed Smoothness

We establish the information-theoretic TV lower bound over $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$, plus a balanced-block W_1 lower bound revealing the anisotropic transport geometry. The proof is a standard Fano/Assouad reduction (Tsybakov, 2009; Yu, 1997); the mixed-smooth ingredient is that the hard TV alternatives live on a hyperbolic shell with $m_L \asymp 2^L L^{d-1}$ degrees of freedom rather than the isotropic 2^{Ld} (Smolyak, 1963; Bungartz & Griebel, 2004; Dung et al., 2018).

For a loss d on $\mathcal{P}(\mathbb{T}^d)$, recall $\mathfrak{R}_n(\mathsf{d}; \mathcal{F}) := \inf_{\widehat{p}} \sup_{p_0 \in \mathcal{F}} \mathbb{E}_{p_0^{\otimes n}}[\mathsf{d}(\widehat{p}, p_0)]$, where the infimum is over all estimators $\widehat{p} = \widehat{p}(X_1, \dots, X_n)$.

Hard family. Fix $L \ge 1$ and let $\mathcal{J}_{=L} := \{j \in \mathbb{N}_{-1}^d : |j|_+ = L\}$, $\Lambda_L := \{(j,k) : j \in \mathcal{J}_{=L},\ k \in \mathcal{K}_j\}$. Then $m_L := |\Lambda_L| \asymp 2^L L^{d-1}$. For signs $\theta \in \{\pm 1\}^{m_L}$ define $p_\theta(x) = 1 + \delta \sum_{\lambda \in \Lambda_L} \theta_\lambda \Psi_\lambda(x)$. All perturbation terms have zero mean, so $\int p_\theta = 1$. Appendix B verifies positivity, Besov membership, metric separation, and KL control.

Theorem 1 (Minimax lower bounds under mixed smoothness)**.** *Fix $s > 0$, fixed d, an envelope $0 < \underline{p} < 1 < \overline{p} < \infty$, and a radius $R > 1$ large enough to contain a nontrivial neighborhood of the constant density. There exist constants $c, c' > 0$ depending only on the fixed problem parameters and the wavelet system such that, for all sufficiently large n,*

$$\mathfrak{R}_n(\mathrm{TV}; \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})) \ge c\, n^{-\frac{s}{2s+1}} (\log n)^{\frac{s(d-1)}{2s+1}}, \tag{4}$$

$$\mathfrak{R}_n(W_1; \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})) \ge c'\, n^{-\frac{s+1/d}{2s+1}}. \tag{5}$$

The W_1 statement uses a balanced tensor block and is not claimed to be the sharp minimax rate; a matching diffusion upper bound remains open.

Proof sketch. For TV, choose a Varshamov–Gilbert packing on Λ_L. The key L^1 separation is proved in Appendix B by L^2–L^4 interpolation and the $p = 4$ Littlewood–Paley inequality, avoiding any endpoint L^1 square-function equivalence. The $B_{2,2}^{s,\mathrm{mix}}$ radius imposes $\delta \lesssim R 2^{-sL} m_L^{-1/2}$, while Fano requires $\delta \lesssim n^{-1/2}$. Balancing gives $2^{(2s+1)L} L^{d-1} \asymp n$. The TV separation is $\delta\sqrt{m_L} \asymp 2^{-sL}$, hence $2^{-sL} \asymp n^{-s/(2s+1)} (\log n)^{s(d-1)/(2s+1)}$.

For W_1, Appendix B uses a balanced block with $j_m \approx L/d$. Kantorovich–Rubinstein duality sees the finest side length $2^{-\|j\|_\infty} \approx 2^{-L/d}$, not the mixed scale $2^{-|j|_+} = 2^{-L}$, giving $n^{-(s+1/d)/(2s+1)}$.

Effective dimension and curse-breaking. The TV exponent $-s/(2s+1)$ is the one-dimensional nonparametric exponent; d appears only through $(\log n)^{s(d-1)/(2s+1)}$. By contrast, isotropic Besov/Hölder classes give cursed strong-loss rates with d in the polynomial exponent, e.g. $n^{-s/(2s+d)}$ in TV (Donoho et al., 1996; Oko et al., 2023). Mixed smoothness thus removes d from the polynomial exponent of the information-theoretic limit; Section 5 shows that diffusion attains it, while the matching W_1 theory remains a transport-geometric problem.

4 From Score Estimation Error to Distribution Estimation Error

We give a *deterministic* reduction: integrated score MSE along noise levels controls TV and W_1 between the diffusion sampler's output and the target law. The reduction follows Oko et al. (2023) but is restated in a modular form decoupling the score class from the distribution-error bound—Section 5 plugs sparse-grid score estimators into this interface.

4.1 Heat-flow parameterization and standing assumptions

Forward heat flow on the torus. Let $X \sim \mathsf{P}_0$ have density p_0 on $\mathbb{T}^d$. For $t > 0$, let $K_t^{\mathbb{T}}$ be the periodic heat kernel and define $Y_t = (X + \sqrt{t}Z) \bmod 1$, $Z \sim \mathcal{N}(0, I_d)$, and $Z \perp X$, equivalently $Y_t \mid X = x$ has density $K_t^{\mathbb{T}}(\cdot - x)$. The marginal density is $p_t = p_0 * K_t^{\mathbb{T}}$ and the score is $\boldsymbol{s}_t = \nabla \log p_t$ (in bold, to distinguish it from the smoothness index s). This is the variance-exploding/heat-flow parameterization; variance-preserving diffusions reduce to it by a time change (Song et al., 2021; Oko et al., 2023).

Why denoising score matching targets $\boldsymbol{s}_t$. On $\mathbb{T}^d$ the denoising target is the periodic kernel score

$$\tau_t(x, y) := \nabla_y \log K_t^{\mathbb{T}}(y - x). \tag{6}$$

By differentiating $p_t(y) = \int K_t^{\mathbb{T}}(y - x)p_0(x)dx$ under the integral,

$$\mathbb{E}[\tau_t(X, Y_t) \mid Y_t = y] = \nabla \log p_t(y) = \boldsymbol{s}_t(y). \tag{7}$$

In the Euclidean, non-wrapped Gaussian model this reduces to the familiar target $-Z/\sqrt{t}$; the periodic form avoids an ambiguity caused by wrapping.

Standing assumptions (for the reductions). Fix $0 < \underline{t} < \overline{t} < \infty$ (the oracle-initialized analysis may take $\overline{t} \lesssim 1$). We assume standard regularity on $t \in [\underline{t}, \overline{t}]$: uniform p_t bounds and Girsanov integrability of the drift mismatch, as used in Appendix E.

Error decomposition. Let $\mathsf{P}_{\underline{t}}$ be the law of the (ideal) reverse sampler endpoint at noise level $\underline{t}$ (so it equals the forward law of $X_{\underline{t}}$), and let $\mathsf{P}_{\underline{t}}^{\widehat{\boldsymbol{s}}}$ be the endpoint law of the learned *continuous-time* reverse sampler started from the *correct* initial law $p_{\overline{t}}$ but driven by $\widehat{\boldsymbol{s}}$, where $\widehat{\boldsymbol{s}}_t$ is any estimate of $\boldsymbol{s}_t$ computed from the data (in Theorem 3, the sparse-grid least-squares estimator $\widehat{\boldsymbol{s}}_{t,L}$ of Section 5). As in Oko et al. (2023), we separate the total estimation/sampling error into

$$\mathsf{d}(\widehat{\mathsf{P}}, \mathsf{P}_0) \;\le\; \underbrace{\mathsf{d}(\mathsf{P}_0, \mathsf{P}_{\underline{t}})}_{\text{early stopping / bias}} \;+\; \underbrace{\mathsf{d}(\mathsf{P}_{\underline{t}}, \mathsf{P}_{\underline{t}}^{\widehat{\boldsymbol{s}}})}_{\text{score error propagation}} \;+\; \underbrace{\mathsf{d}(\mathsf{P}_{\underline{t}}^{\widehat{\boldsymbol{s}}}, \widehat{\mathsf{P}})}_{\text{initialization / discretization}} . \tag{8}$$

Section 5 and Appendix H choose $(\underline{t}, \overline{t})$, initialization, and discretization so that the first and third terms are lower order. The technical heart is the second term, controlled below.

4.2 The integrated score error functional

Define the pointwise (population) score MSE at noise level t as $\mathcal{E}(t) = \mathbb{E}_{X \sim p_t}\big[\|\widehat{\boldsymbol{s}}_t(X) - \boldsymbol{s}_t(X)\|_2^2\big]$, and its integrated version on $[\underline{t}, \overline{t}]$ as

$$\mathcal{E}_{[\underline{t}, \overline{t}]} \;:=\; \int_{\underline{t}}^{\overline{t}} \mathcal{E}(t)\, dt. \tag{9}$$

This is exactly the population quantity targeted by denoising score matching through (7) (Hyvärinen, 2005; Vincent, 2011; Song et al., 2021).

4.3 From integrated score error to distribution error

The chain Girsanov $\to$ data processing $\to$ Pinsker (for TV) and a synchronous-coupling time-localization argument (for W_1) convert $\mathcal{E}_{[\underline{t},\bar{t}]}$ into endpoint-distribution bounds. Full lemmas in Appendix E; the assembled result is below, and Section 5 feeds in a concrete bound on $\mathcal{E}_{[\underline{t},\bar{t}]}$ to obtain a matching TV upper bound and a conservative W_1 bound.

Theorem 2 (Score-to-distribution reduction for TV and W_1)**.** *Under the standing assumptions above, for the learned reverse sampler started from the correct initial law $p_{\bar{t}}$, the endpoint laws satisfy:*

$$\mathbb{E}\big[\mathrm{TV}(\mathsf{P}^{\widehat{\boldsymbol{s}}}_{\underline{t}}, \mathsf{P}_{\underline{t}})\big] \;\lesssim\; \mathbb{E}\left[\left(\int_{\underline{t}}^{\bar{t}} \mathbb{E}_{X\sim p_t}\big[\|\widehat{\boldsymbol{s}}_t(X) - \boldsymbol{s}_t(X)\|_2^2\big]\; dt\right)^{1/2}\right], \tag{10}$$

$$\mathbb{E}\big[W_1(\mathsf{P}^{\widehat{\boldsymbol{s}}}_{\underline{t}}, \mathsf{P}_{\underline{t}})\big] \;\lesssim\; \widetilde{C}\sum_{k=1}^{K}\left(\sqrt{t_{k-1}}\cdot\mathbb{E}\left[\int_{t_{k-1}}^{t_k} \mathbb{E}_{X\sim p_t}\big[\|\widehat{\boldsymbol{s}}_t(X) - \boldsymbol{s}_t(X)\|_2^2\big]\; dt\right]^{1/2}\right), \tag{11}$$

for every partition $\underline{t} = t_0 < \cdots < t_K = \bar{t}$ with $t_k \le 2t_{k-1}$, where $\widetilde{C} = 2^{C_{\mathrm{lip}}} C_{\mathrm{stab}}$ under the W_1-only stability condition of Lemma 15, which (10) *and* (12) *do not use. In particular, Jensen yields the simpler TV corollary*

$$\mathbb{E}\big[\mathrm{TV}(\mathsf{P}^{\widehat{\boldsymbol{s}}}_{\underline{t}}, \mathsf{P}_{\underline{t}})\big] \;\lesssim\; \left(\int_{\underline{t}}^{\bar{t}} \mathbb{E}[\mathcal{E}(t)]\, dt\right)^{1/2}. \tag{12}$$

Why this is a reduction, not a learning bound. Theorem 2 is a *deterministic, module-agnostic inequality*: it converts *any* integrated score MSE, however achieved (sparse grids, neural networks, anything else), into a distribution-error bound, and it depends on no sample size, model class or algorithm. Section 5 certifies $\mathcal{E}_{[\underline{t},\bar{t}]}$ using sparse grids under mixed smoothness; a neural-network analysis would slot into the same interface. The upper bounds of Theorem 3 use only the TV part.

5 Sparse-Grid Score Approximation and Generalization Bounds

Theorem 2 reduces distribution estimation to controlling integrated score MSE. Under the mixed-smooth model, sparse-grid score parameterization gives approximation error 2^{-sL} with $m_L \asymp 2^L L^{d-1}$ active modes, and empirical denoising score matching gives an $O(m_L/(nt))$ generalization term. Balancing these yields the TV rate matching Theorem 1 up to algorithmic polylogarithms.

5.1 Sparse-grid hypothesis class for vector scores

Sparse grid is not an alternative to diffusion. The score-based diffusion framework (forward noising, score learning, reverse-time sampling) is unchanged throughout this section: only the choice of score parameterization changes. Practice uses neural networks for this step; we use sparse grids, which are tractable and inherit the approximation geometry of mixed smoothness.

Recall the tensor-product wavelet basis $\{\Psi_{j,k}\}$ and sparse-grid space $\mathcal{V}_L^{\mathrm{SG}}$ from Section 2, where the hyperbolic-cross index set is

$$\mathcal{J}_L^{\mathrm{HC}} = \{j \in \mathbb{N}_{-1}^d :\; |j|_+ \le L\}, \qquad \mathcal{V}_L^{\mathrm{SG}} = \mathrm{span}\{\Psi_{j,k} :\; j \in \mathcal{J}_L^{\mathrm{HC}},\; k \in \mathcal{K}_j\}.$$

Let

$$\Lambda_L := \{(j,k) :\; j \in \mathcal{J}_L^{\mathrm{HC}},\; k \in \mathcal{K}_j\}, \qquad m_L := |\Lambda_L| \asymp 2^L L^{d-1}. \tag{13}$$

We approximate each component of the score vector field by $\mathcal{V}_L^{\mathrm{SG}}$. For a fixed noise level t, define the sparse-grid score class $\mathcal{S}_{t,L} := \Big\{s : \mathbb{T}^d \to \mathbb{R}^d \;:\; s_r(x) = \sum_{\lambda\in\Lambda_L} a_{r,\lambda}\, \Psi_\lambda(x) \text{ for } r \in [d]\Big\}$. We write Π_L^{SG} for the L^2-orthogonal projector onto $\mathcal{V}_L^{\mathrm{SG}}$ and apply it component-wise to vector fields.

Why sparse grids. The key quantitative distinction from isotropic truncations is (13): m_L scales like 2^L up to a polynomial factor in L, rather than 2^{Ld}.

5.2 Empirical denoising score matching on sparse grids

We estimate $\boldsymbol{s}_t$ from data via denoising score matching. For a fixed t, draw $Y_{i,t}$ from the periodic heat kernel $K_t^{\mathbb{T}}(\cdot - X_i)$ and set $U_{i,t} = \tau_t(X_i, Y_{i,t})$ as in (6); across levels we use one draw $Z_i \sim \mathcal{N}(0, I_d)$ per sample, $Y_{i,t} = (X_i + \sqrt{t}\, Z_i) \bmod 1$. Then (7) gives $\mathbb{E}[U_{i,t} \mid Y_{i,t} = y] = \boldsymbol{s}_t(y)$, so $\boldsymbol{s}_t$ is the population regression function for $(Y_{i,t}, U_{i,t})$. In Euclidean/no-wrap experiments this reduces to the usual target $-Z_i/\sqrt{t}$. We define the empirical risk

$$\widehat{\mathcal{L}}_t(s) := \frac{1}{n}\sum_{i=1}^{n} \|s(Y_{i,t}) - U_{i,t}\|_2^2, \qquad s \in \mathcal{S}_{t,L}, \tag{14}$$

and let $\widehat{\boldsymbol{s}}_{t,L} \in \arg\min_{s \in \mathcal{S}_{t,L}} \widehat{\mathcal{L}}_t(s)$. Because $\mathcal{S}_{t,L}$ is a finite-dimensional linear space, $\widehat{\boldsymbol{s}}_{t,L}$ is a (vector-valued) least-squares estimator over $d\, m_L$ coefficients. We take the minimum-norm solution, unique when the empirical Gram matrix $\widehat{G}_t$ is invertible.

5.3 Balancing approximation and estimation across noise levels

Combining sparse-grid approximation of $\boldsymbol{s}_t$ under mixed smoothness with a fixed-t ERM generalization bound (proofs in Appendix F) yields the per-level bound

$$\mathbb{E}[\mathcal{E}(t)\mathbf{1}_{\mathcal{G}_t^{(1/4)}}] = \mathbb{E}\Big[\mathbb{E}_{X\sim p_t}\big[\|\widehat{\boldsymbol{s}}_{t,L}(X) - \boldsymbol{s}_t(X)\|_2^2\big]\mathbf{1}_{\mathcal{G}_t^{(1/4)}}\Big] \;\lesssim\; \frac{1}{t}\Big(2^{-2sL} + \frac{d\, m_L}{n}\Big). \tag{15}$$

Here $\mathcal{G}_t^{(1/4)} := \{\lambda_{\min}(\widehat{G}_t) \geq \underline{p}/4\}$ (Appendix J.2). Appendix J.3 gives one computable event $\mathcal{C}_n \subseteq \bigcap_{t\in[\underline{t},\overline{t}]} \mathcal{G}_t^{(1/4)}$ with $\mathbb{P}(\mathcal{C}_n^c) = O(n^{-1})$; integrating on it (recall (9)) gives

$$\mathbb{E}\Big[\mathbf{1}_{\mathcal{C}_n}\int_{\underline{t}}^{\overline{t}} \mathcal{E}(t)\,dt\Big] \;\lesssim\; \log\Big(\frac{\overline{t}}{\underline{t}}\Big)\Big(2^{-2sL} + \frac{d\, m_L}{n}\Big). \tag{16}$$

The bracket in (16) is minimized by balancing 2^{-2sL} with $m_L/n \asymp 2^L L^{d-1}/n$, hence $2^{(2s+1)L}L^{d-1} \asymp n$. Thus $L^\star = \frac{1}{2s+1}\log_2 n + O(\log\log n)$ and $m_{L^\star} \asymp n^{1/(2s+1)}(\log n)^{2s(d-1)/(2s+1)}$. Substituting yields the integrated score MSE bound

$$\mathbb{E}\Big[\mathbf{1}_{\mathcal{C}_n}\int_{\underline{t}}^{\overline{t}} \mathcal{E}(t)\,dt\Big] \;\lesssim\; \log\Big(\frac{\overline{t}}{\underline{t}}\Big)\, n^{-\frac{2s}{2s+1}}\,(\log n)^{\frac{2s(d-1)}{2s+1}}, \tag{17}$$

where the polynomial exponent in n is independent of d; fixed-d constants and the displayed sparse-grid logarithms carry the remaining dimension dependence.

5.4 Upper bound and minimax optimality

Plugging the integrated score MSE rate (17) into Theorem 2 on $\mathcal{C}_n$, with TV ≤ 1 off $\mathcal{C}_n$, yields the TV upper bound. On the compact torus, the displayed W_1 upper bound follows conservatively from $W_1 \leq \operatorname{diam}(\mathbb{T}^d)\,\mathrm{TV}$; exploiting the time-localized W_1 weights sharply is left open.

Theorem 3 (Sparse-grid diffusion upper bound)**.** *Let $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$ and assume the multiplier/score-regularity condition in Appendix F. Let $\widehat{\mathsf{P}}$ be the continuous-time diffusion sampler with sparse-grid score estimator at the balanced resolution $L^\star$ and the certified output rule of Appendix J.3, started from $p_{\overline{t}}$ or from an initializer whose error is lower order. Under standard early-stopping and discretization controls, for n sufficiently large,*

$$\mathbb{E}[\mathrm{TV}(\widehat{\mathsf{P}}, \mathsf{P}_0)] \lesssim n^{-s/(2s+1)}(\log n)^{\frac{s(d-1)}{2s+1}}\,\mathrm{polylog}(n), \tag{18}$$

$$\mathbb{E}[W_1(\widehat{\mathsf{P}}, \mathsf{P}_0)] \lesssim n^{-s/(2s+1)}(\log n)^{\frac{s(d-1)}{2s+1}}\,\mathrm{polylog}(n). \tag{19}$$

The W_1 line is the conservative consequence of the TV bound on the torus; matching the balanced-block lower exponent requires a refined time-localized transport analysis and is not claimed here.

Appendix J.3 gives these controls, the uniform output when $\mathcal{C}_n$ fails, and an Euler scheme.

Matching the TV lower bound. When the score-regularity condition is available uniformly, the TV exponent and logarithmic factor in Theorem 3 match the lower bound in Theorem 1, up to the algorithmic polylogarithms from time discretization and endpoint control.

Corollary 1 (TV minimax rate under mixed smoothness)**.** *If the score-regularity condition in Appendix F holds uniformly over the model class, then the TV minimax risk satisfies, up to algorithmic polylogarithms,*

$$\mathfrak{R}_n(\mathrm{TV}; \mathcal{B}^{s,\mathrm{mix}}) = \widetilde{\Theta}\Big(n^{-s/(2s+1)}(\log n)^{\frac{s(d-1)}{2s+1}}\Big),$$

and is attained by the sparse-grid diffusion estimator of Theorem 3. Without a uniform multiplier theorem, the upper bound should be read over the score-regular subclass. For W_1 the diffusion bound gives only the bracket $n^{-(s+1/d)/(2s+1)} \lesssim \mathfrak{R}_n(W_1) \lesssim n^{-s/(2s+1)}(\log n)^{s(d-1)/(2s+1)}\mathrm{polylog}(n)$.

Proposition J.1 verifies this condition uniformly for every $s > 1/2$; for $s \le 1/2$ it is assumed. Appendix M.1 shows that it holds for genuinely rough densities (positive step products, outside $H^{1/2}_{\mathrm{mix}}$, for $0 < s < 1/2$) and for nonconstant densities at $s = 1/2$, and Proposition M.5 shows that a weaker, score-specific approximation input suffices for Theorem 3. Appendix N shows that this input holds for every product density when $0 < s < 1$ (hence on the whole class when $d = 1$), whereas the multiplier condition itself fails for some rough densities, and that the minimax rate itself needs no score input.

Curse-breaking, made formal. Mixed smoothness reduces score degrees of freedom from 2^{Ld} to $m_L \asymp 2^L L^{d-1}$, and Theorem 2 converts this into a TV rate whose polynomial exponent is independent of d; the n-dependent price is the sparse-grid logarithm, with constants depending on fixed d. This is the precise sense in which sparse-grid diffusion breaks the polynomial curse.

6 Discussion and Limitations

Relation to isotropic Besov minimax diffusion theory. Oko et al. (2023) established minimax optimality of diffusion estimators over classical *isotropic* Besov/Hölder classes, where strong-loss rates scale like $n^{-s/(2s+d)}$ (TV) and $n^{-(s+1)/(2s+d)}$ (W_1), so d appears in the exponent. Our contribution identifies a structurally distinct regime—dominating mixed smoothness—where the TV minimax limit has a one-dimensional polynomial exponent because hyperbolic-cross representations have effective size $m_L \asymp 2^L L^{d-1}$ rather than 2^{Ld} (Smolyak, 1963; Bungartz & Griebel, 2004; Vybiral, 2006). Under the score/multiplier condition in Appendix F, diffusion transfers this reduced score complexity into TV guarantees via Theorem 2, the reduction of Oko et al. (2023); the mixed-specific steps are the lower bound with its logarithmic factor, score regularity in the mixed class, and sparse-grid score estimation. Appendix D gives W_1 benchmarks and an autonomous-sampler perturbative closure for $s > 2/d$, but the unconditional all-s matching W_1 theory remains open.

The sparse-grid vs neural network gap. A natural concern is that practical diffusion uses neural networks for score estimation, whereas our upper bound (Section 5) uses sparse grids. *The gap is precisely localized*: Theorem 1 (lower bound) and Theorem 2 (score-to-distribution reduction) are independent of the score class and apply verbatim to any NN-based estimator; only Theorem 3 (the algorithmic upper bound) uses sparse grids. A neural version needs a generalization bound for network score matching (a time-indexed regression whose variance grows like t^{-1}), after which Theorem 2 applies unchanged. In supervised learning over mixed-smooth Besov classes, Suzuki (2019) showed that deep ReLU networks recover the sparse-grid rates—we expect a similar trajectory here. Computationally, one noise level of our estimator is a least-squares problem with m_L features: forming the Gram matrix costs $O(nm_L^2)$ and solving $O(m_L^3 + dm_L^2)$, with $O(m_L^2)$ memory, repeated for each fitted level, whereas a time-conditioned network shares its parameters across levels at a cost set by its architecture and training; we claim no speed advantage. Each coefficient is indexed by output, active coordinates, scale and location, so the fitted score shows which interactions and scales it uses (a representation-level, not causal, reading); neural parameterizations can learn features without a prescribed sparse-grid basis, but their performance outside coordinate-aligned mixed structure is not assessed here (Appendix M.4).

Other open directions. Beyond the NN gap: nonperiodic domains need boundary-adapted wavelets (Cohen et al., 1993; Triebel, 2008); allowing p_0 to vanish locally requires refined score analysis near

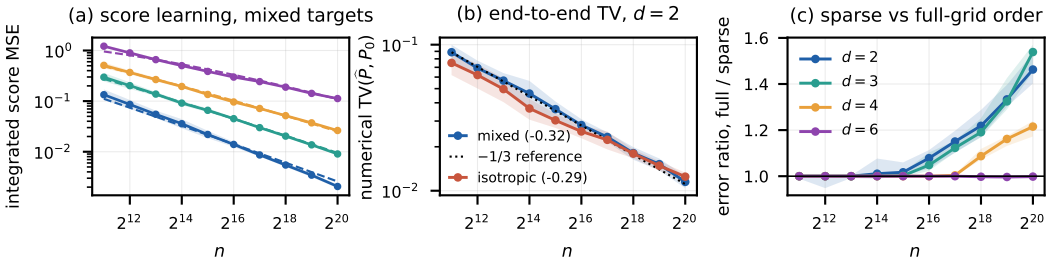

Figure 1: Controlled study. (a) Integrated score error for mixed targets (16-seed geometric mean, ± 1 s.d. in log space; colours as in (c)); dashed: $n^{-2/3}(\log n)^{2(d-1)/3}$ with a fitted constant (post hoc). (b) Numerical TV in $d = 2$ for mixed and equal-smoothness isotropic targets; legend: fitted slopes. (c) Full-grid (sup-norm) to hyperbolic-cross error ratio with matched candidate budgets.

zero; closing the low-smoothness W_1 nonlinear-remainder gap requires a new mixed transport or dynamics-specific bilinear argument; and the computational-statistical tradeoff under discretized samplers remains open (Song et al., 2021; Kloeden & Platen, 1992).

7 Experiments

Controlled study. We use product-density mixtures ($s = 1$), held-out selection including the null model, $n = 2^{11}, \ldots, 2^{20}$, and 16 seeds per cell (Appendix L). Table 1 reports, for each d, the per-seed log–log slope of the integrated score error against n (mean $\pm$ sample s.d. across seeds) and, at $n = 2^{20}$, the median over seeds of the ratio of the full-grid to the hyperbolic-cross integrated error under the same candidate feature budgets and held-out selection rule, together with the number of seeds in which the sparse-grid error is lower. Integrated score error decreases across all four dimensions (Figure 1a), and at the endpoint the hyperbolic-cross ordering has lower error than the full-grid ordering on the same mixed targets in all 16 seeds for $d = 2, 3, 4$ (Figure 1c); in $d = 6$ the two orderings share their first 861 features, which cover most selected models, so no ratio is reported. In $d = 2$, the numerical Fokker–Planck solution gives a TV slope of -0.32 ± 0.02, alongside the $-1/3$ reference (Figure 1b). The registered secondary outcome, the TV bound of the decomposition (8) evaluated numerically, decreases with n on average in every cell, with mean slopes -0.32, -0.26, -0.22 and -0.17 for $d = 2, 3, 4, 6$; it is an estimate, not a certified bound. The prespecified equal-smoothness comparison and subsequent matched-marginal control are fully reported in Appendix L, including validity checks and density-envelope differences.

Table 1: Core results on mixed targets (16 seeds); definitions in the text.

d	slope	full/sparse	sparse lower
2	-0.67 ± 0.03	1.46	16/16
3	-0.56 ± 0.02	1.54	16/16
4	-0.47 ± 0.01	1.22	16/16
6	-0.38 ± 0.007	—	—

8 Conclusion

This work shows that diffusion estimation can exploit mixed smoothness rather than suffer from ambient dimension. For densities in mixed Besov classes, sparse-grid score models separate the statistical difficulty into a one-dimensional smoothness exponent and logarithmic sparse-grid factors. The lower bound and the matching total-variation construction, obtained under the score-regularity condition of Appendix F, identify the hyperbolic shell as the information bottleneck, giving a rate whose polynomial sample exponent is independent of d; the dimension enters only through logarithmic powers and constants. This is the paper's central message: diffusion models can inherit the same curse-breaking structure long known in sparse-grid approximation, but only when the score class respects mixed regularity. The Wasserstein analysis reveals a distinct transport geometry, where balanced tensor scales matter more than total frequency. Future work should turn these structured score estimators into neural architectures, prove a sharp unconditional W_1 diffusion upper bound, and test whether mixed-smooth inductive bias improves high-dimensional generative modeling beyond synthetic benchmarks.

REPRODUCIBILITY STATEMENT

All assumptions are stated where they are used; the score-regularity condition is stated in Appendix F, with examples and a score-specific alternative in Appendix M. Proofs are in the appendices; Appendix J contains the supplementary derivations referenced in the main text. The code, protocols and per-seed results of the controlled study in Appendix L, together with the scripts, raw-file hashes and outputs of the weather-data diagnostic in Appendix O, will be made public after publication of the paper. No separate code or data supplement accompanies this submission.

AI USE STATEMENT

In preparing this manuscript we used generative AI tools as follows. ChatGPT (OpenAI) was used to organize feedback on earlier drafts, to plan the revision, to specify bounded checks of the arguments, to recompute the weather-data summaries of Appendix O independently from the raw files, and to draft candidate proof steps for Appendices J, M, K and N, which were then checked and written up with Claude. Claude (Anthropic) was used to edit the text for readability, to assist in writing proofs (the supplementary derivations in Appendices J, M, K and N), to download and analyze the public weather data of Appendix O, to check cited sources, to audit archived material during revision, and to design, implement and run the controlled study in Appendix L, including the generation of its synthetic data, and to help interpret its results. We did not use generative AI tools for translation or for formulating the main results (Theorems 1–3 and their proofs); AI-assisted formulation of hypotheses was limited to the auxiliary statements of Appendices J, M, K and N described above. All AI-drafted proof steps were checked by the authors, and the code of Appendix L was reviewed by the authors. The authors take responsibility for the final content of this work, including all text, mathematical claims and code produced with the aid of these tools.

REFERENCES

Volker Barthelmann, Erich Novak, and Klaus Ritter. High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics*, 12(4):273–288, 2000.

Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004. doi: 10.1017/S0962492904000182.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1(1):54–81, 1993. ISSN 1063-5203. doi: https://doi.org/10.1006/acha.1993.1005. URL https://www.sciencedirect.com/science/article/pii/S1063520383710055.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian distributions. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12096–12107, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/334da4cbb76302f37bd2e9d86f558869-Paper-Conference.pdf.

Liang Ding, Lu Zou, Wenjia Wang, Shahin Shahrampour, and Rui Tuo. High-dimensional nonparametric density estimation in mixed smooth sobolev spaces, 2021. URL https://arxiv.org/abs/2006.03696.

David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2242660.

Dinh Dung, Vladimir N. Temlyakov, and Tino Ullrich. *Hyperbolic Cross Approximation*. Advanced Courses in Mathematics – CRM Barcelona. Birkhäuser, 2018. doi: 10.1007/978-3-319-92240-9.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24240–24253. Curran Associates, Inc., 2022. doi: 10.52202/068431-1760. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/994545b2308bbbbc97e3e687ea9e464f-Paper-Conference.pdf.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Stochastic Modelling and Applied Probability*. Springer, 1992. doi: 10.1007/978-3-662-12616-5.

Nick Lindemulder, Martin Meyries, and Mark Veraar. Complex interpolation with dirichlet boundary conditions on the half line. *Mathematische Nachrichten*, 291(16):2435–2456, 2018. doi: https://doi.org/10.1002/mana.201700204. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mana.201700204.

NOAA National Centers for Environmental Information. Global surface hourly, 2001. URL https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database. ISD-Lite subset; 2024 observations from stations 725090-14739, 725059-14702, 725098-54704, and 725088-54733; accessed September 26, 2026.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/oko23a.html.

s.A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet Math. Dokl.*, 4:240–243, 01 1963.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969.

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

Hans Triebel. *Theory of Function Spaces III*, volume 100 of *Monographs in Mathematics*. Birkh"auser, 2006. doi: 10.1007/3-7643-7582-5.

Hans Triebel. *Function Spaces and Wavelets on Domains*, volume 7 of *EMS Tracts in Mathematics*. European Mathematical Society, 2008. ISBN 9783037190197.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

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

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

Jan Vybiral. *Function spaces with dominating mixed smoothness*. 2006. URL http://eudml.org/doc/285996.

Shintaro Wakasugi and Taiji Suzuki. State size independent statistical error bound for discrete diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 138908–138943. Curran Associates, Inc., 2025. doi: 10.52202/085713-4637. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/caabce23491e5418f4c9ea348359fcd2-Paper-Conference.pdf.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pp. 423–435, 1997.

Christoph Zenger. Sparse grids. Technical report, Oct 1990. URL http://www5.in.tum.de/pub/Zenger_SparseGrids.pdf. URL: http://www5.in.tum.de/pub/Zenger$_{S}parseGrids.pdf < br > series : SFB - Bericht < br >$.

Zhengyu Zhou and Weiwei Liu. An error analysis of flow matching for deep generative modeling. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78903–78932. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhou25l.html.

APPENDIX

Throughout the appendix, constants may depend on fixed parameters $(s, d, R, \underline{p}, \overline{p})$ and on the chosen wavelet system, but not on n, L.

A BACKGROUND: TENSOR-PRODUCT WAVELETS ON THE TORUS

Let $\{\psi_{j,k}\}_{j\geq 0}$ be a sufficiently regular periodic orthonormal wavelet system on $\mathbb{T}$, and set $\eta_{-1,0} \equiv 1$, $\eta_{j,k} = \psi_{j,k}$ for $j \geq 0$. For $j \in \mathbb{N}_{-1}^d := \{-1, 0, 1, \ldots\}^d$, define $|j|_+ = \sum_m \max(j_m, 0)$ and $\Psi_{j,k} = \prod_m \eta_{j_m,k_m}$. The family consisting of the constant and all nonconstant tensor products is an orthonormal basis of $L^2(\mathbb{T}^d)$; the nonconstant elements have zero mean and satisfy $\|\Psi_{j,k}\|_\infty \lesssim 2^{|j|_+/2}$ with bounded overlap at each fixed j.

A.1 ONE-DIMENSIONAL WAVELET PRIMER (HAAR MENTAL PICTURE)

For the Haar wavelet, $\psi(x) = \mathbf{1}_{[0,1/2)}(x) - \mathbf{1}_{[1/2,1)}(x)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ after periodization. Truncation at level L gives the usual dyadic piecewise-constant approximation. The smoother Daubechies systems used for Besov characterizations obey the same multiresolution bookkeeping, with higher regularity and vanishing moments.

B LOWER BOUND DETAILS: POSITIVITY, SEPARATION, AND KL CONTROL

B.1 HYPERBOLIC-SHELL COUNTING

Lemma 1 (Hyperbolic-shell geometry)**.** *For $L \geq 1$, define $\mathcal{J}_{=L} = \{j \in \mathbb{N}_{-1}^d : |j|_+ = L\}$ and $\Lambda_L = \{(j,k) : j \in \mathcal{J}_{=L}, k \in \mathcal{K}_j\}$. Then*

$$|\Lambda_L| =: m_L \asymp 2^L L^{d-1}.$$

Proof. If r coordinates are active, there are $\binom{d}{r}$ active sets and $\binom{L+r-1}{r-1}$ ways to distribute total level L among them; each such j has 2^L translations. Thus

$$m_L = 2^L \sum_{r=1}^{d} \binom{d}{r}\binom{L+r-1}{r-1} \asymp 2^L L^{d-1}.$$

□

B.2 FEASIBLE PERTURBATIONS

For $\theta \in \{\pm 1\}^{m_L}$, set

$$g_\theta = \sum_{\lambda\in\Lambda_L} \theta_\lambda \Psi_\lambda, \qquad p_\theta = 1 + \delta g_\theta.$$

Lemma 2 (Crude L^∞ bound and positivity)**.** *For every sign vector, $\|g_\theta\|_\infty \lesssim 2^{L/2}L^{d-1}$. Hence, for any fixed envelope $0 < \underline{p} < 1 < \overline{p}$, $p_\theta \in [\underline{p}, \overline{p}]$ if $\delta \leq c_p 2^{-L/2}L^{-(d-1)}$, where c_p depends only on the envelope. The balanced choice used in the TV lower bound satisfies this condition for all large L after reducing the absolute constant.*

Proof. At fixed j, bounded overlap gives $\|\sum_k \theta_{j,k}\Psi_{j,k}\|_\infty \lesssim 2^{L/2}$. Summing over $O(L^{d-1})$ resolution vectors yields the displayed bound. At the TV balance, $\delta \lesssim 2^{-sL}2^{-L/2}L^{-(d-1)/2}$, which is smaller than $2^{-L/2}L^{-(d-1)}$ for large L because $s > 0$. □

Lemma 3 (Besov control for hyperbolic-shell perturbations)**.** *For $p_\theta = 1 + \delta g_\theta$,*

$$\|p_\theta\|^2_{B^{s,\mathrm{mix}}_{2,2}} \lesssim 1 + \delta^2 m_L 2^{2sL}.$$

Thus p_θ lies in the radius-R ball when $\delta \lesssim (R-1)_+ 2^{-sL} m_L^{-1/2}$, or more simply when R is fixed large enough to absorb the constant baseline and the perturbation radius.

Proof. The perturbation coefficients are exactly $\pm\delta$ on the shell $|j|_+ = L$ and zero elsewhere. Substitution into (3) gives the claim. □

B.3 Metric separation and KL

Lemma 4 (L^2, L^4, and TV separation)**.** *For two sign vectors with Hamming distance $h = h(\theta, \theta')$,*

$$\|p_\theta - p_{\theta'}\|_2^2 = 4\delta^2 h, \qquad \|p_\theta - p_{\theta'}\|_1 \gtrsim \delta\, \frac{h^{3/2}}{m_L}.$$

In particular, for the Varshamov packing where $h \geq c m_L$,

$$\mathrm{TV}(p_\theta, p_{\theta'}) \gtrsim \delta\sqrt{m_L}.$$

Proof. The L^2 identity is orthonormality. Put

$$f = p_\theta - p_{\theta'} = \delta \sum_{\lambda \in D} b_\lambda \Psi_\lambda, \qquad b_\lambda \in \{\pm 2\}, \quad |D| = h.$$

We deliberately avoid the endpoint L^1 square-function equivalence. Instead use the Littlewood–Paley inequality at the safe exponent $p = 4$:

$$\|f\|_4 \lesssim \|S(f)\|_4, \qquad S(f)^2 = \delta^2 \sum_{\lambda \in D} b_\lambda^2 \Psi_\lambda^2.$$

For a fixed resolution vector j with $|j|_+ = L$, bounded overlap gives

$$\left\| \sum_{k:(j,k)\in D} \Psi_{j,k}^2 \right\|_2 \lesssim 2^{L/2}\, |D_j|^{1/2}, \qquad D_j := \{k : (j,k) \in D\}.$$

Summing over the $O(L^{d-1})$ resolution vectors and using Cauchy–Schwarz,

$$\|S(f)^2\|_2 \lesssim \delta^2 2^{L/2} \sum_j |D_j|^{1/2} \lesssim \delta^2 2^{L/2} L^{(d-1)/2} h^{1/2} \lesssim \delta^2 m_L,$$

since $h \leq m_L$ and $m_L \asymp 2^L L^{d-1}$. Hence $\|f\|_4 \lesssim \delta m_L^{1/2}$. Interpolating between L^1, L^2, and L^4,

$$\|f\|_2 \leq \|f\|_1^{1/3} \|f\|_4^{2/3},$$

so

$$\|f\|_1 \geq \frac{\|f\|_2^3}{\|f\|_4^2} \gtrsim \delta \frac{h^{3/2}}{m_L}.$$

For $h \geq c m_L$ this gives $\|f\|_1 \gtrsim \delta\sqrt{m_L}$, and TV is one half of the L^1 distance. □

Lemma 5 (KL bound)**.** *If p, q are uniformly bounded above and below away from zero, then* $\mathrm{KL}(p\|q) \leq C\|p - q\|_2^2$.

Proof. Use $\mathrm{KL}(p\|q) \leq \int (p-q)^2/q$ and the lower bound on q. □

Lemma 6 (Varshamov–Gilbert packing)**.** *There exists $\{\theta^{(1)}, \ldots, \theta^{(M)}\} \subset \{\pm 1\}^{m_L}$ with pairwise Hamming distance at least $m_L/8$ and $\log M \gtrsim m_L$.*

Lemma 7 (Fano lower bound specialized to the packing)**.** *With the hyperbolic-shell family above and $\delta \asymp n^{-1/2}$, if $\delta \lesssim R 2^{-sL} m_L^{-1/2}$ then*

$$\inf_{\widehat{p}} \sup_{\theta} \mathbb{E}\,\mathrm{TV}(\widehat{p}, p_\theta) \gtrsim \delta\sqrt{m_L}.$$

Balancing gives the TV rate in Theorem 1.

Proof. The n-sample KL between packing elements is $O(n\delta^2 m_L) = O(m_L) = O(\log M)$, while Lemma 4 gives the metric radius because the packing has pairwise Hamming distance $h \asymp m_L$. Fano's inequality completes the proof. The balancing equation is $n^{-1/2} \asymp 2^{-sL} m_L^{-1/2}$, i.e. $2^{(2s+1)L} L^{d-1} \asymp n$. □

B.4 BALANCED-BLOCK W_1 LOWER BOUND

The TV packing uses the whole hyperbolic shell $|j|_+ = L$. For W_1, the relevant transport length of an anisotropic tensor wavelet is not $2^{-|j|_+}$ but the finest side length $2^{-\|j\|_\infty}$. We therefore use a balanced block rather than the full shell.

Lemma 8 (Anisotropic Lipschitz-dual scale). *Let $j \in \mathbb{N}_0^d$ be fixed, $|j|_+ = L$, $j_* = \|j\|_\infty$, and let $H \subset \mathcal{K}_j$ have cardinality h. For signs $a_k \in \{\pm 1\}$,*

$$\left\| \sum_{k \in H} a_k \Psi_{j,k} \right\|_{\mathsf{Lip}^*} \gtrsim 2^{-j_*}\, h\, 2^{-L/2}.$$

In particular, if $h \asymp 2^L$, the lower bound is $\gtrsim 2^{-j_*} 2^{L/2}$.

Proof sketch. For Haar wavelets, each tensor cell supports a local Lipschitz potential that vanishes at the cell boundary, has slope bounded by a universal constant, and pairs with $a_k \Psi_{j,k}$ at order $2^{-j_*}\|\Psi_{j,k}\|_1 \asymp 2^{-j_*} 2^{-L/2}$. Summing the boundary-vanishing local potentials over disjoint cells preserves the global Lipschitz constant. Smooth compactly supported wavelets are handled by the same local-duality argument with bounded overlap and a fixed loss in constants. The upper bound of this scale follows by integrating by parts in a coordinate attaining j_*. □

Proposition 1 (Balanced-block W_1 lower bound). *For fixed d and $s > 0$,*

$$\mathfrak{R}_n\big(W_1; \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})\big) \gtrsim n^{-(s+1/d)/(2s+1)}.$$

Proof. Choose a balanced multi-index j^{bal} with $|j^{\mathrm{bal}}|_+ = L$ and $\|j^{\mathrm{bal}}\|_\infty \le L/d + 1$. The block has $M_L = 2^L$ translations. For $\theta \in \{\pm 1\}^{M_L}$ set $g_\theta = \sum_{k \in \mathcal{K}_{j^{\mathrm{bal}}}} \theta_k \Psi_{j^{\mathrm{bal}},k}$ and $p_\theta = 1 + \delta g_\theta$. Besov membership requires $\delta \lesssim 2^{-sL} M_L^{-1/2}$, positivity follows from $\|g_\theta\|_\infty \lesssim 2^{L/2}$, and KL control requires $\delta \lesssim n^{-1/2}$. Balancing gives $2^{(2s+1)L} \asymp n$ and $\delta \asymp n^{-1/2}$. A Varshamov–Gilbert code has Hamming distance $h \asymp M_L$, and Lemma 8 gives

$$W_1(p_\theta, p_{\theta'}) \gtrsim \delta\, 2^{-\|j^{\mathrm{bal}}\|_\infty} M_L^{1/2} \gtrsim 2^{-sL} 2^{-L/d}.$$

Substituting $2^L \asymp n^{1/(2s+1)}$ yields the claim. □

The proposition improves the one-coordinate lower bound $n^{-(s+1)/(2s+1)}$ and explains why the original full-shell heuristic with a 2^{-L} transport gain was too strong: balanced mixed interactions have transport length about $2^{-L/d}$.

C IDEAL SPARSE-GRID DENSITY ESTIMATION IN W_1

This section is a benchmark, not a diffusion theorem. It shows that the balanced-block lower-bound geometry is not an artifact: a direct sparse-grid wavelet density estimator can match the candidate W_1 exponent for $d > 2$. The remaining open step for diffusion is to turn score error into the negative-order error used below.

For a zero-mean signed density $h = \sum_{j \neq -\mathbf{1}} \sum_k \theta_{j,k}(h) \Psi_{j,k}$ define

$$J(j) := \max_{1 \le m \le d} \max(j_m, 0), \qquad \|h\|_{\mathcal{N}}^2 := \sum_{j \neq -\mathbf{1}} 2^{-2J(j)} \sum_{k \in \mathcal{K}_j} \theta_{j,k}(h)^2.$$

The weight $2^{-J(j)}$ is the isotropic transport length of the finest side of the tensor cell. It is deliberately different from the mixed level $|j|_+$.

Lemma 9 (Lipschitz-dual domination). *For sufficiently regular tensor-product wavelets on $\mathbb{T}^d$, every zero-mean $h \in L^2(\mathbb{T}^d)$ satisfies*

$$\|h\|_{\mathsf{Lip}^*} := \sup_{\mathrm{Lip}(\varphi) \le 1} \int \varphi h \le C_d \|h\|_{\mathcal{N}}.$$

Proof sketch. The Lipschitz ball embeds into a bounded subset of the isotropic Besov/Sobolev class with one derivative in L^∞, and its wavelet coefficients obey

$$|\langle \varphi, \Psi_{j,k} \rangle| \lesssim 2^{-J(j)}.$$

More precisely, one integrates by parts in a coordinate attaining $J(j)$; the primitive of the one-dimensional wavelet contributes $2^{-J(j)}$, while the remaining tensor factors have bounded overlap after square-function summation. Cauchy–Schwarz in wavelet coefficients yields the displayed inequality. This is the standard wavelet characterization of a negative first-order isotropic norm, written in anisotropic tensor notation. □

Lemma 10 (Weighted sparse-grid counting)**.** *Let*

$$S_L := \sum_{|j|_+ \le L} 2^{|j|_+} 2^{-2J(j)}.$$

For fixed d,

$$S_L \asymp \begin{cases} 1, & d = 1, \\ L, & d = 2, \\ 2^{L(1-2/d)}, & d > 2. \end{cases}$$

Proof. Write $\ell = |j|_+$ and sum over active sets. If r coordinates are active, then $J(j) \ge \ell/r$. The exponential factor is at most $2^{\ell(1-2/r)}$, which is maximized by $r = d$ when $d > 2$. Contributions away from the balanced region $j_m \approx \ell/d$ lose a factor $2^{-2\Delta}$, where $\Delta = J(j) - \ell/d$, and the number of such compositions grows only polynomially in Δ; hence each shell contributes $\asymp 2^{\ell(1-2/d)}$. When $d = 2$ each shell contributes a constant, and summing shells gives L. When $d = 1$ the series $\sum_{\ell \le L} 2^{-\ell}$ is bounded. □

Theorem 4 (Ideal sparse-grid W_1 benchmark)**.** *Consider the direct orthogonal-series estimator*

$$\widetilde{p}_L = 1 + \sum_{\substack{j \ne -\mathbf{1} \\ |j|_+ \le L}} \sum_{k \in \mathcal{K}_j} \widehat{\theta}_{j,k} \Psi_{j,k}, \qquad \widehat{\theta}_{j,k} = \frac{1}{n} \sum_{i=1}^n \Psi_{j,k}(X_i),$$

followed, if necessary, by metric projection onto the set of probability densities in the Lip^* *norm. Uniformly over $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$,*

$$\mathbb{E} W_1(\widehat{p}_L, p) \lesssim 2^{-(s+1/d)L} + \frac{S_L^{1/2}}{\sqrt{n}}.$$

Consequently,

$$\inf_{\widehat{p}} \sup_{p \in \mathcal{B}^{s,\mathrm{mix}}} \mathbb{E} W_1(\widehat{p}, p) \lesssim \begin{cases} n^{-1/2}, & d = 1, \\ (\log n/n)^{1/2}, & d = 2, \\ n^{-(s+1/d)/(2s+1)}, & d > 2. \end{cases}$$

For $d > 2$, this matches the balanced-block lower exponent in Proposition 1.

Proof. Decompose

$$\widetilde{p}_L - p = (\widetilde{p}_L - \Pi_L p) + (\Pi_L p - p).$$

For the bias, since $J(j) \ge |j|_+/d$,

$$\|p - \Pi_L p\|_{\mathcal{N}}^2 = \sum_{|j|_+ > L} 2^{-2J(j)} \sum_k \theta_{j,k}^2 \le 2^{-2(s+1/d)L} \|p\|_{B^{s,\mathrm{mix}}_{2,2}}^2.$$

For the stochastic term, boundedness of p gives

$$\mathbb{E}(\widehat{\theta}_{j,k} - \theta_{j,k})^2 \lesssim n^{-1},$$

so

$$\mathbb{E}\|\widetilde{p}_L - \Pi_L p\|_{\mathcal{N}}^2 \lesssim \frac{1}{n} \sum_{|j|_+ \le L} 2^{-2J(j)} |\mathcal{K}_j| = \frac{S_L}{n}.$$

Lemma 9 transfers the bound to W_1. Projection onto densities changes the distance to p by at most a factor two, because p itself is feasible. The choices of L are: take $L \asymp \log n$ for $d = 1$; solve $2^{(2s+1)L} \asymp n/L$ for $d = 2$; and solve $2^{(2s+1)L} \asymp n$ for $d > 2$. □

Remark 1 (Why this is not yet a diffusion upper bound). Theorem 4 estimates density coefficients directly. The diffusion theorem in the main text controls an integrated $L^2(p_t)$ score error and converts it to TV through pathwise KL. The conditional result in Theorem 5 shows what is sufficient to recover the same W_1 exponent for diffusion: an adjoint smoothing envelope and an aggregate oracle for the forcing field. Proving those hypotheses for the sparse-grid score learner is the remaining bridge identified in Appendix I.

D NEGATIVE-ORDER SCORE-TO-W_1 STABILITY AND LINEARIZED DIFFUSION BENCHMARKS

This section refines the conditional W_1 route. The most natural object is not a per-block oracle for $Q_j(q_{T-t}e_t)$, because multiplication by the evolving density mixes wavelet blocks. Instead we use a time-dependent aggregate negative norm matched to the adjoint heat scale. This gives three nested statements: an aggregate forcing theorem, a linearized exact-density benchmark, and an L^2 perturbative corollary for the autonomous sampler.

For $j \in \mathbb{N}_{-1}^d$, let Q_j denote the wavelet block projection

$$Q_j f = \sum_{k \in \mathcal{K}_j} \langle f, \Psi_{j,k} \rangle \Psi_{j,k},$$

applied componentwise to vector fields. Recall $J(j) = \max_m \max(j_m, 0)$ and

$$S_L = \sum_{|j|_+ \le L} 2^{|j|_+} 2^{-2J(j)}.$$

For $M > 1$ and $t > 0$, define the time-dependent negative norm

$$\|f\|^2_{\mathcal{N}_{t,M}} := \sum_{j \neq -\mathbf{1}} (1 + 2^{J(j)}\sqrt{t})^{-2M} \|Q_j f\|_2^2. \tag{20}$$

The factor $(1 + 2^{J(j)}\sqrt{t})^{-M}$ suppresses frequencies finer than the heat radius $\sqrt{t}$.

Let $\widehat{\boldsymbol{s}}_t = \boldsymbol{s}_t + e_t$ be the learned score. Let q_τ be the density of the approximate reverse process, initialized at $q_0 = p_T$, and set

$$h_\tau = q_\tau - p_{T-\tau}, \qquad F_t(x) := q_{T-t}(x) e_t(x), \qquad 0 < t \le T.$$

The vector field F_t is the divergence forcing appearing in the density-error equation.

Lemma 11 (Adjoint error identity)**.** *Let φ be 1-Lipschitz and let v_t solve*

$$\partial_t v_t = \frac{1}{2}\Delta v_t + \boldsymbol{s}_t \cdot \nabla v_t, \qquad v_0 = \varphi.$$

If the reverse sampler is exactly initialized at p_T, then

$$\langle \varphi, h_T \rangle = \int_0^T \langle F_t, \nabla v_t \rangle \, dt.$$

Proof. In reverse time τ, the density error satisfies

$$\partial_\tau h_\tau = -\nabla \cdot (h_\tau \boldsymbol{s}_{T-\tau}) + \frac{1}{2}\Delta h_\tau - \nabla \cdot (q_\tau e_{T-\tau}).$$

Let $u_\tau = v_{T-\tau}$ solve the backward adjoint equation with terminal datum $u_T = \varphi$. Differentiating $\langle u_\tau, h_\tau \rangle$ cancels the linear forward-adjoint terms and leaves

$$\frac{d}{d\tau}\langle u_\tau, h_\tau \rangle = \int q_\tau e_{T-\tau} \cdot \nabla u_\tau.$$

Integrating over τ and using $h_0 = 0$, then changing variables $t = T - \tau$, gives the formula. □

Assumption 1 (Adjoint smoothing envelope)**.** There exist constants $M > 1$ and C_{adj} such that for every 1-Lipschitz terminal function φ, the solution v_t above satisfies, for all $t \in (0, T]$,

$$\sum_{j \neq -\mathbf{1}} (1 + 2^{J(j)}\sqrt{t})^{2M} \|Q_j \nabla v_t\|_2^2 \leq C_{\mathrm{adj}}.$$

Remark 2 (Why the envelope is the right target). If $\boldsymbol{s}_t \equiv 0$, then v_t is the heat-smoothed test function and the envelope follows from standard Littlewood–Paley heat estimates. For the true diffusion drift $\boldsymbol{s}_t = \nabla \log p_t$, the Doob-transform identity

$$v_t(y) = \frac{P_t(\varphi p_0)(y)}{P_t p_0(y)} = \mathbb{E}[\varphi(X_0) \mid X_t = y]$$

suggests the same scale: a Lipschitz test function is averaged over a posterior ball of radius $\sqrt{t}$, so frequencies with $2^{J(j)} \gg t^{-1/2}$ should be suppressed. A full proof follows from derivative bounds of the form $\|\nabla^r v_t\|_\infty \lesssim t^{-(r-1)/2}$ plus Littlewood–Paley estimates; we keep the envelope as an explicit hypothesis to isolate the remaining regularity input.

Assumption 2 (Aggregate forcing oracle)**.** For some resolution level L there are nonnegative numbers b_j with

$$\sum_{|j|_+ > L} b_j \leq R^2$$

such that, for all $t \in (0, T]$,

$$\mathbb{E}\|F_t\|_{\mathcal{N}_{t,M}}^2 \leq \frac{C_{\mathrm{forc}}}{t} \left[\frac{1}{n} \sum_{|j|_+ \leq L} (1 + 2^{J(j)}\sqrt{t})^{-2M} 2^{|j|_+} + \sum_{|j|_+ > L} (1 + 2^{J(j)}\sqrt{t})^{-2M} b_j 2^{-2s|j|_+} \right]. \tag{21}$$

The expectation includes training and sampling randomness.

Theorem 5 (Conditional aggregate negative-order diffusion W_1 bound)**.** *Assume exact initialization, Assumption 1, and Assumption 2. Then*

$$\mathbb{E}W_1(q_T, p_0) \lesssim 2^{-(s+1/d)L} + \sqrt{\frac{S_L}{n}} + \mathbf{1}_{\{d=2\}} \frac{L}{\sqrt{n}},$$

where the hidden constant depends on fixed problem parameters, T, M, C_{adj}, and C_{forc}, but not on n or L. Consequently,

$$\mathbb{E}W_1(q_T, p_0) \lesssim \begin{cases} n^{-1/2}, & d = 1, \\ n^{-1/2} \log n, & d = 2, \\ n^{-(s+1/d)/(2s+1)}, & d > 2. \end{cases}$$

Proof. Fix a 1-Lipschitz φ. Lemma 11, the definition of $\mathcal{N}_{t,M}$, and Cauchy–Schwarz give

$$|\langle \varphi, h_T \rangle| \leq C_{\mathrm{adj}}^{1/2} \int_0^T \|F_t\|_{\mathcal{N}_{t,M}} dt.$$

Taking the supremum over φ, then expectations and Jensen, gives

$$\mathbb{E}W_1(q_T, p_0) \lesssim \int_0^T \left(\mathbb{E}\|F_t\|_{\mathcal{N}_{t,M}}^2\right)^{1/2} dt.$$

Insert (21) and bound the square root of the sum by the sum of the square roots of the variance and bias parts. The resulting time integrals cannot be exchanged with the ℓ^2 sum over blocks, since Minkowski's inequality goes the other way; instead they are evaluated in Appendix J.4 by grouping the blocks by $J(j)$ and the time axis into dyadic scales. This gives a bias contribution $\lesssim 2^{-(s+1/d)L} R$ for every d, using $J(j) \geq |j|_+/d$, and a variance contribution $\lesssim \sqrt{S_L/n}$ for $d = 1$ and $d > 2$; for $d = 2$ the variance contribution is $\lesssim n^{-1/2} L$. The rate statements follow from Lemma 10, with $L \asymp \log n$ when $d = 2$. □

D.1 LINEARIZED EXACT-DENSITY FORCING BENCHMARK

The autonomous reverse sampler has forcing $q_{T-t}e_t$. A useful benchmark replaces this by the exact-density forcing $p_t e_t$. Define the signed linear response h_τ^{lin} by

$$\partial_\tau h_\tau^{\mathrm{lin}} = -\nabla\cdot(h_\tau^{\mathrm{lin}}\boldsymbol{s}_{T-\tau}) + \frac{1}{2}\Delta h_\tau^{\mathrm{lin}} - \nabla\cdot(p_{T-\tau}e_{T-\tau}), \qquad h_0^{\mathrm{lin}} = 0. \tag{22}$$

It has zero total mass, but $p_{T-\tau} + h_\tau^{\mathrm{lin}}$ need not be nonnegative; therefore we measure $\|h_T^{\mathrm{lin}}\|_{\mathsf{Lip}^*}$ rather than a literal distance between two probability laws.

Assumption 3 (Exact-density multiplier and score oracle)**.** For some $M' > M$, multiplication by p_t is bounded from $\mathcal{N}_{t,M'}$ to $\mathcal{N}_{t,M}$ uniformly in t:

$$\|p_t f\|_{\mathcal{N}_{t,M}} \le C_{\mathrm{mult}}\|f\|_{\mathcal{N}_{t,M'}}.$$

Moreover the score error satisfies, for nonnegative b_j with $\sum_{|j|_+>L} b_j \le R^2$,

$$\mathbb{E}\|e_t\|^2_{\mathcal{N}_{t,M'}} \le \frac{C_{\mathrm{score}}}{t}\left[\frac{1}{n}\sum_{|j|_+\le L}(1+2^{J(j)}\sqrt{t})^{-2M'}2^{|j|_+} + \sum_{|j|_+>L}(1+2^{J(j)}\sqrt{t})^{-2M'}b_j 2^{-2s|j|_+}\right].$$

Theorem 6 (Linearized exact-density forcing benchmark)**.** *Assume Assumption 1 and Assumption 3. Then the solution of* (22) *satisfies*

$$\mathbb{E}\|h_T^{\mathrm{lin}}\|_{\mathsf{Lip}^*} \lesssim 2^{-(s+1/d)L} + \sqrt{\frac{S_L}{n}} + \mathbf{1}_{\{d=2\}}\frac{L}{\sqrt{n}}.$$

Consequently the same dimension-dependent cases as in Theorem 5 hold; in particular, for $d > 2$ and $2^{(2s+1)L} \asymp n$,

$$\mathbb{E}\|h_T^{\mathrm{lin}}\|_{\mathsf{Lip}^*} \lesssim n^{-(s+1/d)/(2s+1)}.$$

Proof. The proof is Theorem 5 applied to $F_t^{\mathrm{lin}} = p_t e_t$. The multiplier bound transfers the aggregate oracle for e_t into the aggregate forcing oracle for $p_t e_t$, with M' replacing M and harmless constants. □

D.2 L^2 NONLINEAR PERTURBATION FOR THE AUTONOMOUS SAMPLER

The actual sampler differs from the linearized response by the nonlinear remainder

$$(q_{T-t} - p_t)e_t = h_{T-t}e_t.$$

The earlier TV–L^∞ product control is sufficient but unnecessarily strong. A sharper perturbative route uses only an L^2 score bound and the interpolation inequality

$$\|h\|_2^2 \le \|h\|_1\|h\|_\infty.$$

Thus, if the exact and approximate reverse marginals are uniformly bounded, then

$$\|h_{T-t}\|_2 \lesssim \mathrm{TV}(q_{T-t}, p_t)^{1/2}.$$

This improves the smoothness threshold for closing the autonomous-sampler W_1 rate.

Assumption 4 (Pathwise TV, density envelope, and integrated L^2 score control)**.** For the W_1-balanced resolution choice $2^{(2s+1)L} \asymp n$, assume

$$\mathbb{E}\left[\sup_{0\le\tau\le T}\mathrm{TV}(q_\tau, p_{T-\tau})\right] \le C_{\mathrm{tv}}\, n^{-s/(2s+1)}\,\mathrm{polylog}(n),$$

and assume the reverse marginals obey the uniform envelope

$$\sup_{0\le\tau\le T}\|q_\tau\|_\infty + \sup_{0\le t\le T}\|p_t\|_\infty \le C_{\mathrm{den}}.$$

Finally, assume

$$\mathbb{E}\left(\int_0^T \|e_t\|_2 dt\right)^2 \le C_2\, n^{-2s/(2s+1)}\,\mathrm{polylog}(n).$$

Corollary 2 (L^2 nonlinear perturbation)**.** *Assume Theorem 6 and Assumption 4 hold, and let $d > 2$. If*

$$s > \frac{2}{d},$$

then the autonomous reverse sampler satisfies

$$\mathbb{E}W_1(q_T, p_0) \lesssim n^{-(s+1/d)/(2s+1)} \operatorname{polylog}(n).$$

Proof. Let $\eta_T = h_T - h_T^{\mathrm{lin}}$. Its forcing is $h_{T-t}e_t$. For any 1-Lipschitz test function, the adjoint solution satisfies $\|\nabla v_t\|_\infty \lesssim 1$, hence

$$|\langle \varphi, \eta_T \rangle| \lesssim \int_0^T \|h_{T-t}\|_2 \|e_t\|_2 dt.$$

The density envelope and $\|h\|_2^2 \le \|h\|_1 \|h\|_\infty$ give

$$\|h_{T-t}\|_2 \lesssim \mathrm{TV}(q_{T-t}, p_t)^{1/2}.$$

Therefore

$$\|\eta_T\|_{\mathsf{Lip}^*} \lesssim \left(\sup_{\tau \le T} \mathrm{TV}(q_\tau, p_{T-\tau}) \right)^{1/2} \int_0^T \|e_t\|_2 dt.$$

Taking expectations and applying Cauchy's inequality together with Assumption 4 gives

$$\mathbb{E}\|\eta_T\|_{\mathsf{Lip}^*} \lesssim \left(\mathbb{E} \sup_{\tau \le T} \mathrm{TV}(q_\tau, p_{T-\tau}) \right)^{1/2} \left(\mathbb{E} \left(\int_0^T \|e_t\|_2 dt \right)^2 \right)^{1/2} \lesssim n^{-3s/[2(2s+1)]} \operatorname{polylog}(n).$$

This nonlinear remainder is lower order than the linearized benchmark whenever

$$\frac{3s}{2} > s + \frac{1}{d},$$

which is exactly $s > 2/d$. Since $q_T - p_0 = h_T^{\mathrm{lin}} + \eta_T$, the claim follows by the triangle inequality and Theorem 6. □

Remark 3 (What remains below $s = 2/d$). Corollary 2 improves the previous TV–L^∞ perturbative threshold $s > 1/2 + 1/d$ to $s > 2/d$ by using TV-to-L^2 interpolation and an integrated L^2 score bound. For $s \le 2/d$, generic Hölder/interpolation or paraproduct estimates based only on TV, bounded density envelopes, and score L^q control do not appear to close the sharp autonomous-sampler W_1 rate. Removing this remaining restriction likely requires a dynamics-specific cancellation for $h_{T-t}e_t$ rather than a black-box product estimate.

E SCORE-TO-DISTRIBUTION REDUCTION: FULL PROOFS

The proof follows the standard Girsanov and time-localization argument used in diffusion minimax analyses (Oko et al., 2023).

Lemma 12 (Pathwise KL control via Girsanov)**.** *Under the standing assumptions, conditional on the learned score family,*

$$\mathrm{KL}(\mathbb{Q} \| \widehat{\mathbb{Q}}) \le \frac{1}{2} \int_{\underline{t}}^{\bar{t}} \mathbb{E}_{X \sim p_t} \|\widehat{\boldsymbol{s}}_t(X) - \boldsymbol{s}_t(X)\|_2^2 dt.$$

Proof. Apply Cameron–Martin–Girsanov to the two reverse SDE path measures. The martingale term has zero mean, and the true reverse marginal at noise time t is p_t. □

Lemma 13 (Endpoint data processing)**.** *The endpoint marginal KL is bounded by the path KL.*

Proof. Taking the endpoint is a measurable map, and KL contracts under measurable transformations. □

Lemma 14 (Pinsker). $\mathrm{TV}(\mu,\nu) \le \sqrt{\mathrm{KL}(\mu\|\nu)/2}$.

Lemma 15 (Single-block W_1 perturbation bound). *Let $\underline{t} = t_0 < \cdots < t_K = \bar{t}$ with $t_k \le 2t_{k-1}$, and let $\mathsf{P}^{(k)}_{\underline{t}}$ be the endpoint law of the hybrid sampler, started from $p_{\bar{t}}$, that uses $\widehat{\boldsymbol{s}}$ on the k blocks nearest the output, $[t_0,t_k]$, and $\boldsymbol{s}$ on $[t_k,\bar{t}]$. Assume the following W_1-only stability condition: there are deterministic constants C_{lip} and C_{stab} such that, for every data set, $\|\nabla\widehat{\boldsymbol{s}}_t(x)\|_{\mathrm{op}} \le C_{\mathrm{lip}}/t$ for all $t \in [\underline{t},\bar{t}]$ and x, and the reverse dynamics driven by $\widehat{\boldsymbol{s}}$ from any level $t \in [\underline{t},\bar{t}]$ down to $\underline{t}$ is C_{stab}-Lipschitz in W_1 between initial laws. Any dependence of C_{lip} or C_{stab} on n or $\underline{t}$ is carried by $\widetilde{C} = 2^{C_{\mathrm{lip}}}C_{\mathrm{stab}}$. Then, for a block $[t_{k-1},t_k]$ in this time-localized hybrid sampler,*

$$\mathbb{E}W_1(\mathsf{P}^{(k)}_{\underline{t}},\mathsf{P}^{(k-1)}_{\underline{t}}) \lesssim \widetilde{C}\sqrt{t_{k-1}}\,\mathbb{E}\Big[\int_{t_{k-1}}^{t_k}\mathcal{E}(t)dt\Big]^{1/2}.$$

Proof. The laws $\mathsf{P}^{(k)}$ and $\mathsf{P}^{(k-1)}$ differ only on the block $[t_{k-1},t_k]$. Above it both processes follow the exact reverse dynamics from $p_{\bar{t}}$, so both have law p_{t_k} at level t_k; on the block the process X of $\mathsf{P}^{(k-1)}$ still follows the exact dynamics and has marginals p_t; below it both follow the learned dynamics. Couple the two processes synchronously on the block, from the same point at level t_k, and work with lifts to $\mathbb{R}^d$. Their difference Δ has no martingale part, and with $e_t = \widehat{\boldsymbol{s}}_t - \boldsymbol{s}_t$, in reverse time,

$$|\Delta|' \le |\widehat{\boldsymbol{s}}_t(X+\Delta) - \widehat{\boldsymbol{s}}_t(X)| + |e_t(X)| \le \frac{C_{\mathrm{lip}}}{t}|\Delta| + |e_t(X)|.$$

Grönwall's inequality gives $|\Delta_{t_{k-1}}| \le (t_k/t_{k-1})^{C_{\mathrm{lip}}}\int_{t_{k-1}}^{t_k}|e_t(X_t)|dt \le 2^{C_{\mathrm{lip}}}\int_{t_{k-1}}^{t_k}|e_t(X_t)|dt$. Since $X_t \sim p_t$, the Cauchy–Schwarz inequality in (t,ω) and $t_k - t_{k-1} \le t_{k-1}$ give, conditionally on the data, $\mathbb{E}|\Delta_{t_{k-1}}| \le 2^{C_{\mathrm{lip}}}\sqrt{t_{k-1}}\,(\int_{t_{k-1}}^{t_k}\mathcal{E}(t)dt)^{1/2}$. This bounds the W_1 distance of the laws at level t_{k-1}, and the learned dynamics below the block multiplies it by at most C_{stab}. Taking expectations over the data and applying Jensen's inequality gives the claim, because C_{lip} and C_{stab} are deterministic. On the torus no tail event is needed. □

Lemma 16 (Time-localized telescoping). *Replacing the true score by the learned score one block at a time and summing the hybrid endpoint distances gives the W_1 bound in Theorem 2.*

Proof. Order the hybrids as in Lemma 15: $\mathsf{P}^{(0)} = \mathsf{P}_{\underline{t}}$ uses the true score throughout, $\mathsf{P}^{(K)} = \mathsf{P}^{\widehat{\boldsymbol{s}}}_{\underline{t}}$ uses the learned score throughout, and $\mathsf{P}^{(k)}$ replaces the true score by the learned one on the k blocks nearest the output. In this order the block being replaced is preceded by the exact dynamics, so its reference marginal is p_t, and it is followed by the learned dynamics, whose stability constant is C_{stab}. The triangle inequality along $\mathsf{P}^{(0)},\dots,\mathsf{P}^{(K)}$ and Lemma 15 give (11) with $\widetilde{C} = 2^{C_{\mathrm{lip}}}C_{\mathrm{stab}}$. □

Theorem 2 follows by combining Lemmas 12–14 for TV and Lemmas 15–16 for W_1.

F Score regularity under mixed-smooth Besov assumptions

We use the following explicit multiplier condition. For $t \in [\underline{t},\bar{t}]$, assume multiplication by $1/p_t$ is bounded on $B^{s,\mathrm{mix}}_{2,2}$ with norm at most $M_s\underline{p}^{-1}$. For $s > 1/2$ this is not an additional assumption: the mixed Besov space is then an algebra, the heat marginals stay in the density class, and Proposition J.1 proves the bound uniformly in t and over the class. For $s \le 1/2$ we keep it as an explicit assumption. Without such a multiplier statement, the step $\partial_r p_t \mapsto \partial_r p_t/p_t$ is not justified by positivity alone.

Lemma 17 (Convolution stability). *If $f \in B^{s,\mathrm{mix}}_{2,2}$, then $\|f * K^{\mathbb{T}}_t\|_{B^{s,\mathrm{mix}}_{2,2}} \le \|f\|_{B^{s,\mathrm{mix}}_{2,2}}$ and*

$$\|\partial_r(f * K^{\mathbb{T}}_t)\|_{B^{s,\mathrm{mix}}_{2,2}} \lesssim t^{-1/2}\|f\|_{B^{s,\mathrm{mix}}_{2,2}}.$$

Proof. Heat convolution is a Fourier multiplier bounded by one; a derivative contributes a frequency factor controlled by $\sup_{u\ge 0} ue^{-ctu^2} \lesssim t^{-1/2}$. □

Lemma 18 (Score mixed-smooth bound)**.** *Under the multiplier condition above, for each coordinate r,*

$$\|\boldsymbol{s}_{t,r}\|_{B^{s,\mathrm{mix}}_{2,2}} = \|\partial_r p_t/p_t\|_{B^{s,\mathrm{mix}}_{2,2}} \lesssim M_s \underline{p}^{-1} t^{-1/2} R.$$

Lemma 19 (Fixed-t sparse-grid score risk)**.** *For the least-squares estimator in* (14) *and the event* $\mathcal{G}_t^{(1/4)} = \{\lambda_{\min}(\widehat{G}_t) \geq \underline{p}/4\}$,

$$\mathbb{E}\big[\mathcal{E}(t)\mathbf{1}_{\mathcal{G}_t^{(1/4)}}\big] \lesssim t^{-1}\left(2^{-2sL} + \frac{dm_L}{n}\right),$$

where the first term is the corrected $B^{s,\mathrm{mix}}_{2,2}$ sparse-grid bias and the second term is the linear-regression variance.

Proof. Apply Lemma 18 and the projection bound (23) componentwise. The excess-risk term is the finite-dimensional least-squares bound with response variance of order t^{-1}; Appendix J.2 gives the random-design argument in full (conditional variance, Gram matrix and leverage, projection residual), and Lemma J.8 gives one event contained in every $\mathcal{G}_t^{(1/4)}$, $t \in [\underline{t}, \overline{t}]$, on which the bound is integrated over t. □

G Heat-semigroup bias bounds

G.1 Wavelet tail control on a hyperbolic cross

For $f \in B^{s,\mathrm{mix}}_{2,2}$,

$$\|f - \Pi_L^{\mathrm{SG}} f\|_2^2 = \sum_{|j|_+ > L} \sum_k \theta_{j,k}^2 \leq 2^{-2sL} \|f\|^2_{B^{s,\mathrm{mix}}_{2,2}}. \tag{23}$$

This is the key corrected Hilbert-tail bound: the logarithmic factor appears through m_L, not through the sharp bias.

Lemma 20 (TV heat bias)**.** *Let $\alpha = \min(s, 2)$. If $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$, then*

$$\mathrm{TV}(p_t, p_0) \lesssim t^{\alpha/2}.$$

For $0 < s \leq 2$ this is the usual $t^{s/2}$ bias; for smoother classes the first-order heat semigroup difference saturates at order t unless higher-order bias correction is introduced.

Proof. On the unit-volume torus, $\|\cdot\|_1 \leq \|\cdot\|_2$. The heat multiplier satisfies

$$|e^{-t|\omega|^2/2} - 1| \lesssim t^{\alpha/2}(1 + |\omega|^2)^{\alpha/2}, \qquad 0 < \alpha \leq 2.$$

The mixed Besov norm controls the isotropic H^α norm for every $\alpha \leq s$, because on a tensor block $J(j) \leq |j|_+$ and hence $(1 + |\omega|^2)^\alpha \lesssim 2^{2\alpha J(j)} \leq 2^{2s|j|_+}$. Therefore

$$\|p_t - p_0\|_1 \leq \|p_t - p_0\|_2 \lesssim t^{\alpha/2} \|p_0\|_{B^{s,\mathrm{mix}}_{2,2}},$$

which gives the claim. □

Lemma 21 (Conservative W_1 heat bias)**.** *On the compact torus, $W_1(p_t, p_0) \leq \mathrm{diam}(\mathbb{T}^d)\,\mathrm{TV}(p_t, p_0)$, hence the $t^{\min(s,2)/2}$ order from Lemma 20 is always valid. A sharper extra-$\sqrt{t}$ transport bias requires an additional negative-order mixed-smooth argument and is not used in the revised main theorem.*

H Discretization and initialization effects

Theorem 3 is stated for a continuous-time reverse sampler with either oracle initialization at $p_{\overline{t}}$ or an initializer whose discrepancy from $p_{\overline{t}}$ is lower order than the statistical rate. A fixed $\overline{t} \asymp 1$ uniform initializer on the torus leaves a nonzero heat-mixing bias, so it is not automatically asymptotically negligible. Two standard closures are: take $\overline{t} \asymp \log n$ so that the heat flow is within $o(r_n)$ of uniform for the target rate r_n, or initialize from an independently noised empirical/reference sample and bound that initialization error separately. Appendix J.3 makes this explicit for the continuous-time sampler

with the certified output rule (Proposition J.3) and for an Euler scheme with the drift frozen at the points of a geometric time grid: comparing path laws along the true reverse process, the discretization error is controlled by the regularity of the true score alone, so no Lipschitz or tube condition on the learned score is needed, and a polynomial number of steps keeps it below the statistical TV rate. These terms are absorbed into the displayed $\mathrm{polylog}(n)$ factors in (18)–(19).

I Additional remarks and extensions

Nonperiodic domains. Boundary-corrected tensor wavelets should give analogous results on $[0,1]^d$, at the cost of boundary terms.

Other Besov parameters. Appendix K gives the immediate one-sided $p = 2$ summability extensions: the upper guarantee for $q \le 2$ under its score input, and the lower bounds for $q \ge 2$. Sharp logarithmic factors on the complementary sides and changes in p require further analysis.

Sharp W_1. Proposition 1 shows that balanced mixed interactions have transport scale $2^{-L/d}$. The ideal estimator benchmark, Theorem 4, matches this exponent for direct density estimation when $d > 2$. The conditional diffusion theorem, Theorem 5, shows that the same exponent follows from an adjoint smoothing envelope plus an aggregate forcing oracle. The linearized benchmark, Theorem 6, identifies the exact-density forcing mechanism; Corollary 2 shows that the autonomous sampler also attains this exponent under the perturbative condition $s > 2/d$. The unresolved task is to remove this remaining nonlinear-remainder restriction and to determine whether any logarithmic factor survives the anisotropic transport optimization.

J Supplementary derivations

This appendix gives complete versions of three arguments that support the upper bound of Theorem 3: the regularity of the score for $s > 1/2$ (Section J.1), the random-design analysis behind (15) (Section J.2), and explicit initialization, early-stopping and discretization controls (Section J.3). Section J.4 evaluates the time integral in the proof of Theorem 5. Constants may depend on $(s, d, R, \underline{p}, \overline{p})$ and on the wavelet system, never on n, L, t or p_0.

J.1 The multiplier condition holds for every $s > 1/2$

For $f \in L^2(\mathbb{T}^d)$ let $\|f\|_{s,d}^2 := \sum_{k\in\mathbb{Z}^d} \prod_{r=1}^d \langle k_r\rangle^{2s} |\widehat{f}(k)|^2$ with $\langle\kappa\rangle := (1+\kappa^2)^{1/2}$. For sufficiently regular wavelets this norm is equivalent to (3) (Vybiral, 2006; Dung et al., 2018); $c_{\mathrm{eq}} \ge 1$ denotes the equivalence constant. For a complex Hilbert space X and $a \ge 0$, let $H^a(\mathbb{T}; X)$ be the space of X-valued u with $\|u\|_{H^a(\mathbb{T};X)}^2 := \sum_{\kappa\in\mathbb{Z}} \langle\kappa\rangle^{2a} \|\widehat{u}(\kappa)\|_X^2 < \infty$. Singling out the first coordinate, $H^s_{\mathrm{mix}}(\mathbb{T}^d) = H^s(\mathbb{T}; X)$ isometrically with $X = H^s_{\mathrm{mix}}(\mathbb{T}^{d-1})$ (and $X = \mathbb{C}$ if $d = 1$).

Lemma J.1 (One-dimensional vector-valued estimates)**.** *Let X be a complex Hilbert space with a bilinear product satisfying $\|ab\|_X \le C_X \|a\|_X \|b\|_X$.*

(a) If $a > 1/2$, then $\sum_\kappa \|\widehat{u}(\kappa)\|_X \le c_a \|u\|_{H^a(\mathbb{T};X)}$ with $c_a^2 = \sum_\kappa \langle\kappa\rangle^{-2a}$; in particular u is continuous and $\sup_x \|u(x)\|_X \le c_a \|u\|_{H^a(\mathbb{T};X)}$.

(b) If $a, b \ge c \ge 0$ and $a + b - c > 1/2$, then $\|uv\|_{H^c(\mathbb{T};X)} \le C_{a,b,c} C_X \|u\|_{H^a(\mathbb{T};X)} \|v\|_{H^b(\mathbb{T};X)}$.

(c) If $0 < a < 1$, then $\|u\|_{H^a(\mathbb{T};X)}^2$ is equivalent, with constants depending only on a, to $\|u\|_{L^2(\mathbb{T};X)}^2 + \int_{\mathbb{T}}\int_{\mathbb{T}} \|u(x) - u(y)\|_X^2 \, |x-y|_{\mathbb{T}}^{-1-2a} \, dx\, dy$.

Proof. (a) is the Cauchy–Schwarz inequality. (c) follows from Parseval's identity for X-valued Fourier series, since $\int_{\mathbb{T}} |e^{2\pi i\kappa h} - 1|^2 |h|_{\mathbb{T}}^{-1-2a} dh \asymp |\kappa|^{2a}$ for $\kappa \ne 0$. For (b) it suffices to consider trigonometric polynomials. Then $\widehat{uv}(\kappa) = \sum_\lambda \widehat{u}(\lambda)\widehat{v}(\kappa - \lambda)$, so with $U = \|\widehat{u}(\cdot)\|_X$ and $V = \|\widehat{v}(\cdot)\|_X$ we get $\|\widehat{uv}(\kappa)\|_X \le C_X (U * V)(\kappa)$. Since $\langle\kappa\rangle^c \le 2^c(\langle\lambda\rangle^c + \langle\kappa-\lambda\rangle^c)$, it remains to bound $\|(\langle\cdot\rangle^c U) * V\|_{\ell^2}$ and $\|U * (\langle\cdot\rangle^c V)\|_{\ell^2}$. For the first, Young's inequality with $1/p + 1/q = 3/2$ and

Hölder's inequality with $\frac{1}{p} = \frac{1}{2} + \frac{1}{\alpha}$, $\frac{1}{q} = \frac{1}{2} + \frac{1}{\beta}$ and $\frac{1}{\alpha} + \frac{1}{\beta} = \frac{1}{2}$ give

$$\|(\langle\cdot\rangle^c U) * V\|_{\ell^2} \le \|\langle\cdot\rangle^a U\|_{\ell^2}\, \|\langle\cdot\rangle^{c-a}\|_{\ell^\alpha}\, \|\langle\cdot\rangle^b V\|_{\ell^2}\, \|\langle\cdot\rangle^{-b}\|_{\ell^\beta}.$$

The two weight norms are finite if $1/\alpha < a - c$ or $\alpha = \infty$, and if $1/\beta < b$ or $\beta = \infty$. A choice with $1/\alpha, 1/\beta \in [0, 1/2]$ exists because $(a - c) + b > 1/2$: take $1/\alpha = 0$ if $b > 1/2$, and otherwise $1/\beta$ slightly below b (or $1/\beta = 0$ if $b = 0$). The second convolution is symmetric. □

Lemma J.2 (Algebra property)**.** *If $s > 1/2$, then $\|fg\|_{s,d} \le C_{s,d}\|f\|_{s,d}\|g\|_{s,d}$.*

Proof. By induction on d, apply Lemma J.1(b) with $a = b = c = s$ in the first coordinate and $X = H^s_{\mathrm{mix}}(\mathbb{T}^{d-1})$, which is an algebra by the induction hypothesis ($X = \mathbb{C}$ for $d = 1$). □

Proposition J.1 (Inverse closure and uniform multiplier bound)**.** *Let $s > 1/2$ and $\underline{p} > 0$. There is a nondecreasing, polynomially bounded function $\Phi_{s,d}$ such that every real $f \in H^s_{\mathrm{mix}}(\mathbb{T}^d)$ with $f \ge \underline{p}$ satisfies $\|1/f\|_{s,d} \le \Phi_{s,d}(\|f\|_{s,d})$. Consequently, for every $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$, every $t \ge 0$ and every g,*

$$\|g/p_t\|_{s,d} \le C_{s,d}\, \Phi_{s,d}(c_{\mathrm{eq}} R)\, \|g\|_{s,d},$$

uniformly in t and p_0. In particular the multiplier condition of Appendix F holds for all $t > 0$, uniformly over the density class, and Lemma 18 gives $\|\boldsymbol{s}_{t,r}\|_{B^{s,\mathrm{mix}}_{2,2}} \lesssim t^{-1/2}$.

Proof. We argue by induction on d; for $d = 0$ set $\Phi_{s,0}(r) := 1/\underline{p}$. All functions $\Phi_{s,d}$ below also depend on $\underline{p}$, which is fixed and suppressed from the notation. Let $X = H^s_{\mathrm{mix}}(\mathbb{T}^{d-1})$, an algebra by Lemma J.2 with constant C_X, in which $\|1/g\|_X \le \Phi_{s,d-1}(\|g\|_X)$ whenever $g \ge \underline{p}$. Let $f \ge \underline{p}$ with $\|f\|_{H^s(\mathbb{T};X)} \le r$. By Lemma J.1(a), $x \mapsto f(x) \in X$ is continuous with $\sup_x \|f(x)\|_X \le c_s r$, and $f(x) \ge \underline{p}$ for every x. Hence $\sup_x \|1/f(x)\|_X \le \varphi := \Phi_{s,d-1}(c_s r)$.

Step 1 (fractional regularity). Let $a_0 := s$ if $s < 1$ and $a_0 := 3/4$ if $s \ge 1$, so $a_0 \in (1/2, 1)$. In X, $1/f(x) - 1/f(y) = (f(y) - f(x))\,(1/f(x))\,(1/f(y))$, hence $\|1/f(x) - 1/f(y)\|_X \le C_X^2 \varphi^2 \|f(x) - f(y)\|_X$. Lemma J.1(c) gives $\|1/f\|_{H^{a_0}(\mathbb{T};X)} \le A_0 := C(\varphi + C_X^2 \varphi^2 r)$. If $s < 1$ this is the claim.

Step 2 (bootstrap for $s \ge 1$). Suppose $\|1/f\|_{H^a(\mathbb{T};X)} \le A$ for some $a \in (1/2, s)$. Since $f \in H^1(\mathbb{T}; X)$ and $1/f$ is continuous with values in X, passing to the limit in the difference quotients of the identity above gives $\partial_1(1/f) = -(\partial_1 f)(1/f)^2$: with $D_h g := (g(\cdot + h) - g)/h$ we have $D_h(1/f) = -(D_h f)\,(1/f(\cdot + h))\,(1/f)$, where $D_h f \to \partial_1 f$ in $L^2(\mathbb{T}; X)$ and $1/f(\cdot + h) \to 1/f$ uniformly in X by uniform continuity, so $D_h(1/f)$ converges in $L^2(\mathbb{T}; X)$ to the right-hand side, which is therefore the weak derivative. Lemma J.1(b) with orders (a, a, a) gives $\|(1/f)^2\|_{H^a(\mathbb{T};X)} \le C C_X A^2$, and $\|\partial_1 f\|_{H^{s-1}(\mathbb{T};X)} \le 2\pi r$. With $c := \min(s - 1, a) \ge 0$, Lemma J.1(b) applies to orders $(s - 1, a, c)$: if $c = s - 1$ the condition reads $a > 1/2$, and if $c = a < s - 1$ it reads $s - 1 > 1/2$. Therefore $\|\partial_1(1/f)\|_{H^c(\mathbb{T};X)} \le C C_X^2 r A^2$ and $\|1/f\|_{H^{c+1}(\mathbb{T};X)} \le C(A + C_X^2 r A^2)$ with $c + 1 = \min(s, a + 1)$. Starting from a_0, at most $\lceil s \rceil$ such steps reach $a = s$. Each step maps a polynomial bound in r to a polynomial bound, which defines $\Phi_{s,d}$.

For the consequence, the heat semigroup multiplies Fourier coefficients by $e^{-2\pi^2 t|k|^2} \in (0, 1]$ and preserves the envelope, so $\|p_t\|_{s,d} \le \|p_0\|_{s,d} \le c_{\mathrm{eq}} R$ and $p_t \ge \underline{p}$ for all $t \ge 0$. Apply the first part to $f = p_t$ and Lemma J.2 to $g \cdot (1/p_t)$. □

The constant depends on $\underline{p}$ but not on $\overline{p}$, t or p_0. For $s \le 1/2$ the space contains unbounded functions and is not an algebra, so this argument does not apply; there the condition remains the explicit assumption stated in Appendix F.

J.2 Random-design analysis of the fixed-level least-squares estimator

Let $\Psi = (\Psi_\lambda)_{\lambda \in \Lambda_L} : \mathbb{T}^d \to \mathbb{R}^{m_L}$ be the sparse-grid basis of (13), built from a fixed compactly supported periodic wavelet system of class C^1. The training data are i.i.d. pairs (X_i, Z_i) with $X_i \sim p_0$ and $Z_i \sim \mathcal{N}(0, I_d)$ independent, independent of the Brownian motion of the sampler, and used at every level: $Y_{i,t} = (X_i + \sqrt{t}\, Z_i) \bmod 1$ and $U_{i,t} = \tau_t(X_i, Y_{i,t})$. For each fixed t the pairs $(Y_{i,t}, U_{i,t})$ are i.i.d. with $Y_{i,t} \sim p_t$; this is all the fixed-level analysis uses, and different levels are not assumed

independent. The estimator of Section 5 decouples over the score coordinates r into ordinary least squares of $U_{i,t,r}$ on $\Psi(Y_{i,t})$, and we take the minimum-norm solution. Let $G_t := \mathbb{E}\,\Psi(Y_t)\Psi(Y_t)^\top$ and $\widehat{G}_t := n^{-1}\sum_i \Psi(Y_{i,t})\Psi(Y_{i,t})^\top$. Because Ψ is L^2-orthonormal and $\underline{p} \le p_t \le \overline{p}$, $\underline{p} I \preceq G_t \preceq \overline{p} I$. Bounded overlap of the wavelets at each level gives $K_\Psi := \sup_x \|\Psi(x)\|_2^2 \le C_\Psi m_L$. Differentiating a basis function on block j in coordinate r costs a factor at most $C2^{j_r} \le C2^L$ in sup norm (and removes it if $j_r = -1$), so the same argument gives $D_\Psi := \sup_x \sum_\lambda \|\nabla \Psi_\lambda(x)\|_2^2 \le C_\Psi 2^{2L} m_L$, enlarging C_Ψ if necessary. For $\theta \in (0,1]$ let $\mathcal{G}_t^{(\theta)} := \{\lambda_{\min}(\widehat{G}_t) \ge \theta \underline{p}\}$, on which the least-squares solution is unique, and write $\mathcal{G}_t := \mathcal{G}_t^{(1/2)}$. The event in (15) is the larger event $\mathcal{G}_t^{(1/4)}$.

Lemma J.3 (Conditional second moment of the denoising target)**.** *For every y and r, $\mathbb{E}[\tau_{t,r}(X,Y_t)^2 \mid Y_t = y] \le (\overline{p}/\underline{p})\, t^{-1}$, and $|\boldsymbol{s}_{t,r}| \le (\overline{p}/\underline{p})(2/(\pi t))^{1/2}$.*

Proof. The conditional density of X given $Y_t = y$ is $K_t^{\mathbb{T}}(y-x)p_0(x)/p_t(y) \le (\overline{p}/\underline{p})K_t^{\mathbb{T}}(y-x)$. Hence the conditional second moment is at most $(\overline{p}/\underline{p})\int(\partial_r K_t^{\mathbb{T}})^2/K_t^{\mathbb{T}}$, which is $(\overline{p}/\underline{p})$ times the Fisher information of the wrapped normal in coordinate r. The wrapped normal is a sum of translates of the Gaussian density, and $(a,b) \mapsto a^2/b$ is jointly convex, so this Fisher information is at most that of the Gaussian, $1/t$. For the second claim, $|\partial_r p_t| \le \overline{p}\|\partial_r K_t^{\mathbb{T}}\|_{L^1(\mathbb{T}^d)} \le \overline{p}\,\|\partial_r \varphi_t\|_{L^1(\mathbb{R}^d)} = \overline{p}(2/(\pi t))^{1/2}$, where φ_t is the Gaussian density, and $p_t \ge \underline{p}$. □

Lemma J.4 (Gram event)**.** $\mathbb{P}(\mathcal{G}_t^c) \le m_L \exp(-c_0 n\underline{p}/K_\Psi)$ *with* $c_0 = (1-\ln 2)/2$.

Proof. Apply the matrix Chernoff lower-tail bound (Tropp, 2012, Theorem 1.1) to the independent positive semidefinite summands $n^{-1}\Psi(Y_{i,t})\Psi(Y_{i,t})^\top$, whose maximal eigenvalue is at most K_Ψ/n, with $\mu_{\min} = \lambda_{\min}(G_t) \ge \underline{p}$ and deviation parameter $1/2$. □

Lemma J.5 (Risk on the good event)**.** *Under the multiplier condition (in particular for every $s > 1/2$ by Proposition J.1), if $m_L \le n$ then the least-squares estimator $\widehat{\boldsymbol{s}}_{t,L}$ of Section 5 satisfies*

$$\mathbb{E}\big[\mathcal{E}(t)\,\mathbf{1}_{\mathcal{G}_t^{(1/4)}}\big] \lesssim \frac{1}{t}\Big(2^{-2sL} + \frac{d\,m_L}{n}\Big),$$

and hence the same bound on $\mathcal{G}_t \subseteq \mathcal{G}_t^{(1/4)}$.

Proof. Fix r and write $f = \boldsymbol{s}_{t,r}$, $f^* = \Psi^\top\beta^*$ for its $L^2(p_t)$-projection onto $\mathcal{V}_L^{\mathrm{SG}}$, $g = f - f^*$, and $\epsilon = U_{t,r} - f(Y_t)$, so that $\mathbb{E}[\epsilon \mid Y_t] = 0$ and $\mathbb{E}[\Psi g] = 0$. With $\xi_i = \Psi(Y_{i,t})(\epsilon_i + g(Y_{i,t}))$ and $a = n^{-1}\sum_i \xi_i$, the least-squares coefficients satisfy $\widehat{\beta} - \beta^* = \widehat{G}_t^{-1}a$ on $\mathcal{G}_t^{(1/4)}$. There $\widehat{G}_t \succeq (\underline{p}/4)I \succeq \kappa G_t$ with $\kappa = \underline{p}/(4\overline{p})$, hence $\widehat{G}_t^{-1}G_t\widehat{G}_t^{-1} \preceq \kappa^{-2}G_t^{-1}$ and

$$\|\widehat{f} - f^*\|_{L^2(p_t)}^2 = a^\top \widehat{G}_t^{-1}G_t\widehat{G}_t^{-1}a \le \kappa^{-2}a^\top G_t^{-1}a.$$

The ξ_i are i.i.d. and centered, so $\mathbb{E}\,a^\top G_t^{-1}a = n^{-1}\mathbb{E}\big[\Psi^\top G_t^{-1}\Psi\,(\mathbb{E}[\epsilon^2 \mid Y_t] + g^2)\big] \le n^{-1}\big(\sigma_t^2 m_L + (K_\Psi/\underline{p})\|g\|_{L^2(p_t)}^2\big)$, using $\mathbb{E}\,\Psi^\top G_t^{-1}\Psi = m_L$, $\Psi^\top G_t^{-1}\Psi \le K_\Psi/\underline{p}$ and $\sigma_t^2 \le (\overline{p}/\underline{p})t^{-1}$ from Lemma J.3. By Pythagoras, $\mathbb{E}[\|\widehat{f} - f\|_{L^2(p_t)}^2 \mathbf{1}_{\mathcal{G}_t^{(1/4)}}] \le (1 + \kappa^{-2}C_\Psi/\underline{p})\|g\|_{L^2(p_t)}^2 + \kappa^{-2}\sigma_t^2 m_L/n$. Finally $\|g\|_{L^2(p_t)}^2 \le \overline{p}\,\|f - \Pi_L f\|_{L^2}^2 \le \overline{p}\,2^{-2sL}\|f\|_{B^{s,\mathrm{mix}}_{2,2}}^2 \lesssim t^{-1}2^{-2sL}$ by (23) and Proposition J.1. Summing over the d coordinates proves the claim. □

Lemma J.5 is the bound (15); it controls one level at a time. The same observations and noise draws are used at every level, so the events at different levels are dependent. The discretized sampler of Proposition J.2 uses finitely many levels, and its events are combined by a union bound. The continuous-time sampler of Theorem 3 uses every level of $[\underline{t}, \overline{t}]$: Lemma J.8 constructs one computable event $\mathcal{C}_n \subseteq \bigcap_t \mathcal{G}_t^{(1/4)}$, with $\mathbb{P}(\mathcal{C}_n^c) = O(n^{-1})$ under the choices of Proposition J.3, and the certified output rule of Appendix J.3 specifies the sampler when $\mathcal{C}_n$ fails. At the balanced resolution $m_{L^\star} \asymp n^{1/(2s+1)}(\log n)^{2s(d-1)/(2s+1)}$ the failure probability $m_L \exp(-c_0 n\underline{p}/(C_\Psi m_L))$ of Lemma J.4 is smaller than any power of n.

J.3 Initialization, early stopping and an explicit discretization

Lemma J.6 (True-score regularity)**.** *For all $t > 0$, $|\boldsymbol{s}_t| \le C_0 t^{-1/2}$, $\|\nabla \boldsymbol{s}_t\|_{\mathrm{op}} \le C_1 t^{-1}$ and $|\partial_t \boldsymbol{s}_t| \le C_2 t^{-3/2}$, with C_i depending only on $(d, \underline{p}, \overline{p})$.*

Proof. Periodization does not increase L^1 norms, so $\|\partial^\alpha p_t\|_\infty \le \overline{p}\,\|\partial^\alpha \varphi_t\|_{L^1(\mathbb{R}^d)} \lesssim \overline{p}\, t^{-|\alpha|/2}$. The claims follow from $\boldsymbol{s}_t = \nabla p_t / p_t$, $\nabla \boldsymbol{s}_t = \nabla^2 p_t / p_t - \boldsymbol{s}_t \boldsymbol{s}_t^\top$, $\partial_t p_t = \frac{1}{2}\Delta p_t$ and $p_t \ge \underline{p}$. □

Lemma J.7 (Initialization and early stopping)**.** $\mathrm{TV}(\mathrm{Unif}, p_{\overline{t}}) \le \frac{1}{2} R e^{-2\pi^2 \overline{t}}$*, and a reverse dynamics applied to both initial laws cannot increase this distance. Moreover* $\mathrm{TV}(p_{\underline{t}}, p_0) \lesssim \underline{t}^{\min(s,2)/2}$ *(Lemma 20).*

Proof. Every nonconstant Fourier coefficient of $p_{\overline{t}}$ equals that of p_0 times at most $e^{-2\pi^2 \overline{t}}$, and $\|p_0 - 1\|_{L^2} \le R$; on the unit torus $\|\cdot\|_{L^1} \le \|\cdot\|_{L^2}$. Total variation contracts under Markov kernels. □

Fix $\varepsilon \in (0, 1]$ and the geometric grid $\tau_0 = \overline{t} > \tau_1 > \cdots > \tau_N = \underline{t}$ with $\tau_{j+1} = \max(\tau_j/(1+\varepsilon), \underline{t})$, so $h_j := \tau_j - \tau_{j+1} \le \varepsilon \tau_{j+1}$ and $N \le 1 + 2\varepsilon^{-1} \log(\overline{t}/\underline{t})$. Fit $\widehat{\boldsymbol{s}}_{\tau_j, L}$ at the grid levels and run

$$Y_{j+1} = Y_j + h_j\, \widehat{\boldsymbol{s}}_{\tau_j, L}(Y_j) + \sqrt{h_j}\, \xi_{j+1} \ (\mathrm{mod}\ 1), \qquad \xi_j \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_d), \qquad Y_0 \sim \mathrm{Unif}(\mathbb{T}^d).$$

Each step is the exact solution of the reverse SDE whose drift is frozen at the last grid point, in state and in time.

Proposition J.2 (Discretized sampler)**.** *Let $\mathbb{Q}$ be the law of the exact reverse process on $[\underline{t}, \overline{t}]$ started from $p_{\overline{t}}$, and let $\widehat{\mathbb{Q}}$ be the path law of the scheme above started from $p_{\overline{t}}$, that is, of its continuous-time interpolation, which on each step follows the reverse SDE with the drift frozen at the last grid point. Conditionally on the data,*

$$\mathrm{KL}(\mathbb{Q} \,\|\, \widehat{\mathbb{Q}}) \le \sum_{j<N} h_j\, \mathcal{E}(\tau_j) + C_3(d\varepsilon + \varepsilon^2)\big(1 + \log(\overline{t}/\underline{t})\big),$$

with C_3 depending only on $(d, \underline{p}, \overline{p})$. No regularity of the learned score is used beyond measurability and boundedness.

Proof. Both path laws have unit diffusion coefficient and bounded adapted drifts, so Girsanov's theorem gives $\mathrm{KL}(\mathbb{Q}\|\widehat{\mathbb{Q}}) = \frac{1}{2}\mathbb{E}_{\mathbb{Q}} \sum_j \int_{\tau_{j+1}}^{\tau_j} \|\boldsymbol{s}_t(X_t) - \widehat{\boldsymbol{s}}_{\tau_j}(X_{\tau_j})\|^2 dt$, the comparison being made along the *true* reverse path, as in Chen et al. (2023). Split the integrand as $2\|\boldsymbol{s}_t(X_t) - \boldsymbol{s}_{\tau_j}(X_{\tau_j})\|^2 + 2\|\boldsymbol{s}_{\tau_j}(X_{\tau_j}) - \widehat{\boldsymbol{s}}_{\tau_j}(X_{\tau_j})\|^2$. Under $\mathbb{Q}$, $X_{\tau_j} \sim p_{\tau_j}$, so the second part integrates to $h_j \mathcal{E}(\tau_j)$. The true reverse path is the time reversal of the forward heat flow, so $X_{\tau_j} = X_t + \sqrt{\tau_j - t}\, Z$ (mod 1) in law and $\mathbb{E}|X_t - X_{\tau_j}|_{\mathbb{T}}^2 \le d(\tau_j - t)$. By Lemma J.6, for $t \in [\tau_{j+1}, \tau_j]$,

$$\mathbb{E}\|\boldsymbol{s}_t(X_t) - \boldsymbol{s}_{\tau_j}(X_{\tau_j})\|^2 \le 2C_1^2 d\, \frac{h_j}{t^2} + 2C_2^2 \frac{h_j^2}{t^3} \le \frac{2C_1^2 d\varepsilon + 2C_2^2 \varepsilon^2}{\tau_{j+1}}.$$

Integrating over an interval of length $h_j \le \varepsilon \tau_{j+1}$ and summing over $N \le 1 + 2\varepsilon^{-1}\log(\overline{t}/\underline{t})$ intervals gives the second term. □

Corollary J.1 (End-to-end rate of the discretized sampler)**.** *Assume the multiplier condition (automatic for $s > 1/2$). Take $\overline{t} = \log n$, $\underline{t} = n^{-2s/((2s+1)\min(s,2))}$, $L = L^\star$ as in Section 5.3, and $\varepsilon = n^{-2s/(2s+1)}$, so that the scheme uses $N = O(n^{2s/(2s+1)} \log n)$ steps. Its output law $\widehat{\mathsf{P}}^{\mathrm{Eu}}$ satisfies*

$$\mathbb{E}\,\mathrm{TV}(\widehat{\mathsf{P}}^{\mathrm{Eu}}, \mathsf{P}_0) \lesssim n^{-s/(2s+1)} (\log n)^{\frac{s(d-1)}{2s+1}} (\log n)^{1/2},$$

which is the rate of Theorem 3, with $\mathrm{polylog}(n) = (\log n)^{1/2}$.

Proof. Decompose as in (8). The early-stopping term is $O(n^{-s/(2s+1)})$ and the initialization term is at most $\frac{1}{2}Rn^{-2\pi^2}$ by Lemma J.7. For the score term, let $\mathcal{G} = \bigcap_{j<N} \mathcal{G}_{\tau_j}$. Then $\mathbb{E}\,\mathrm{TV} \le \mathbb{E}[(\frac{1}{2}\,\mathrm{KL})^{1/2}\mathbf{1}_{\mathcal{G}}] + \mathbb{P}(\mathcal{G}^c)$, where $\mathbb{P}(\mathcal{G}^c) \le N m_L \exp(-c_0 n\underline{p}/(C_\Psi m_L))$ by Lemma J.4 is smaller than any power of n. By Jensen's inequality, Proposition J.2 and Lemma J.5 (note $\mathcal{G} \subseteq \mathcal{G}_{\tau_j}^{(1/4)}$ for every j),

$$\mathbb{E}[\mathrm{KL}\ \mathbf{1}_{\mathcal{G}}] \lesssim \sum_{j<N} \varepsilon\tau_j \cdot \tau_j^{-1}\Big(2^{-2sL} + \frac{dm_L}{n}\Big) + (d\varepsilon + \varepsilon^2)\log n \lesssim \log n\Big(2^{-2sL^\star} + \frac{dm_{L^\star}}{n} + \varepsilon\Big),$$

and the balance of Section 5.3 gives the stated rate. □

The continuous-time sampler. The sampler of Theorem 3 is driven by the least-squares scores at every level of $[\underline{t}, \bar{t}]$, so a single event must control all Gram matrices at once, and the output must be specified when this event fails.

Lemma J.8 (All-time Gram certificate)**.** *Fix $A \ge 1$ and $0 < \underline{t} < \bar{t}$. Let $B_n := (2d(A+3)\log(2nd))^{1/2}$, $\Lambda_n := 2(K_\Psi D_\Psi)^{1/2}B_n$ and $\delta_n := \underline{p}/(8\Lambda_n)$. Let $\mathcal{U}_n$ consist of the points $\sqrt{\underline{t}} + i\delta_n < \sqrt{\bar{t}}$, $i = 0, 1, \dots$, together with $\sqrt{\bar{t}}$, so that it has $N_n \le 2 + (\sqrt{\bar{t}} - \sqrt{\underline{t}})/\delta_n$ points, and define*

$$\mathcal{C}_n := \Big\{\max_{i\le n} \|Z_i\|_2 \le B_n\Big\} \cap \bigcap_{u\in\mathcal{U}_n} \mathcal{G}_{u^2}.$$

Then (i) $\mathcal{C}_n$ is determined by the training data and is checked by computing N_n Gram matrices; (ii) on $\mathcal{C}_n$, $\lambda_{\min}(\widehat{G}_t) \ge 3\underline{p}/8$ for every $t \in [\underline{t}, \bar{t}]$, so $\mathcal{C}_n \subseteq \mathcal{G}_t^{(1/4)}$; (iii) $\mathbb{P}(\mathcal{C}_n^c) \le n^{-A-2} + N_n m_L \exp(-c_0 n\underline{p}/K_\Psi)$; and (iv) under the conditions of Lemma J.5,

$$\mathbb{E}\Big[\mathbf{1}_{\mathcal{C}_n}\int_{\underline{t}}^{\bar{t}} \mathcal{E}(t)\,dt\Big] \lesssim \log\Big(\frac{\bar{t}}{\underline{t}}\Big)\Big(2^{-2sL} + \frac{d\,m_L}{n}\Big).$$

Proof. (i) is immediate. (ii) Write $u = \sqrt{t}$. Since Ψ is periodic and of class C^1, $\frac{d}{du}\Psi(Y_{i,u^2}) = \nabla\Psi(Y_{i,u^2})Z_i$, where the Jacobian $\nabla\Psi$ has Frobenius norm at most $D_\Psi^{1/2}$. Hence, on $\{\max_i \|Z_i\|_2 \le B_n\}$,

$$\Big\|\frac{d}{du}\widehat{G}_{u^2}\Big\|_{\mathrm{op}} \le \frac{2}{n}\sum_i \|\Psi(Y_{i,u^2})\|_2\,\|\nabla\Psi(Y_{i,u^2})Z_i\|_2 \le 2(K_\Psi D_\Psi)^{1/2}\max_i\|Z_i\|_2 \le \Lambda_n,$$

and by Weyl's inequality $u \mapsto \lambda_{\min}(\widehat{G}_{u^2})$ is Λ_n-Lipschitz. Every $u \in [\sqrt{\underline{t}}, \sqrt{\bar{t}}]$ lies within δ_n of $\mathcal{U}_n$, so on $\mathcal{C}_n$ we get $\lambda_{\min}(\widehat{G}_{u^2}) \ge \underline{p}/2 - \Lambda_n\delta_n = 3\underline{p}/8$. (iii) If $\|Z_i\|_2 > B_n$, then $|Z_{ir}| > B_n/\sqrt{d}$ for some r. The Gaussian tail $\mathbb{P}(|Z_{ir}| > x) \le 2e^{-x^2/2}$ and a union bound over the nd coordinates give $\mathbb{P}(\max_i\|Z_i\|_2 > B_n) \le 2nd\,e^{-B_n^2/(2d)} = (2nd)^{-A-2} \le n^{-A-2}$. At each of the N_n deterministic levels u^2 the pairs are i.i.d., so Lemma J.4 applies there, and a union bound over $\mathcal{U}_n$ gives the second term; no independence across levels is used. (iv) For fixed data, $\widehat{G}_t$ and $\widehat{b}_t := n^{-1}\sum_i \Psi(Y_{i,t})U_{i,t}^\top$ are continuous in t, and they are measurable in the data, hence jointly measurable; since the Moore–Penrose inverse is Borel, the minimum-norm coefficients and $\mathcal{E}(t) \ge 0$ are jointly measurable. By Tonelli's theorem and (ii), $\mathbb{E}[\mathbf{1}_{\mathcal{C}_n}\int\mathcal{E}(t)dt] = \int\mathbb{E}[\mathbf{1}_{\mathcal{C}_n}\mathcal{E}(t)]dt \le \int\mathbb{E}[\mathbf{1}_{\mathcal{G}_t^{(1/4)}}\mathcal{E}(t)]dt$, and Lemma J.5 bounds the integrand by $Ct^{-1}(2^{-2sL} + dm_L/n)$. □

Certified output rule. Given the training data, if $\mathcal{C}_n$ holds, the output $\widehat{\mathsf{P}}$ is the law at level $\underline{t}$ of the reverse diffusion

$$dY_\tau = \widehat{\boldsymbol{s}}_{\bar{t}-\tau,L}(Y_\tau)\,d\tau + dW_\tau \pmod 1, \qquad \tau \in [0, \bar{t}-\underline{t}], \qquad Y_0 \sim \mathrm{Unif}(\mathbb{T}^d),$$

driven by a Brownian motion W independent of the training data. If $\mathcal{C}_n$ fails, the output is $\widehat{\mathsf{P}} := \mathrm{Unif}(\mathbb{T}^d)$. On $\mathcal{C}_n$ the drift is defined at every level: $\|\widehat{G}_t^{-1}\|_{\mathrm{op}} \le 8/(3\underline{p})$, so the coefficients $\widehat{G}_t^{-1}\widehat{b}_t$

are continuous in t. Moreover $U_{i,t} = \nabla \log K_t^{\mathbb{T}}(\sqrt{t}\, Z_i)$, and $\log K_t^{\mathbb{T}}$ is smooth on $\mathbb{T}^d$ and depends continuously on $t > 0$, so

$$\sup_{t\in[\underline{t},\bar{t}],\, x\in\mathbb{T}^d} \|\widehat{\boldsymbol{s}}_{t,L}(x)\|_2 \le \frac{8\sqrt{d}\, K_\Psi}{3\underline{p}} \sup_{t\in[\underline{t},\bar{t}]} \sup_{z\in\mathbb{T}^d} \|\nabla \log K_t^{\mathbb{T}}(z)\|_2 < \infty.$$

A bounded, jointly measurable drift with unit diffusion determines a unique law, for instance through Girsanov's theorem, and this law depends measurably on the data. The rule leaves the least-squares scores unchanged: it only specifies the output when the certificate fails, and this costs $\mathbb{P}(\mathcal{C}_n^c)$ in the TV risk.

Proposition J.3 (Continuous-time TV bound)**.** *Assume the multiplier condition (automatic for $s > 1/2$), and take $\bar{t} = \log n$, $\underline{t} = n^{-2s/((2s+1)\min(s,2))}$, $L = L^\star$ and $A = 1$ in Lemma J.8. The certified output $\widehat{\mathsf{P}}$ satisfies*

$$\mathbb{E}\,\mathrm{TV}(\widehat{\mathsf{P}}, \mathsf{P}_0) \le C\underline{t}^{\frac{\min(s,2)}{2}} + \frac{R}{2} e^{-2\pi^2\bar{t}} + C\Big\{\log\Big(\frac{\bar{t}}{\underline{t}}\Big)\Big(2^{-2sL} + \frac{d\, m_L}{n}\Big)\Big\}^{\frac{1}{2}} + \mathbb{P}(\mathcal{C}_n^c),$$

and the right-hand side is $\lesssim n^{-s/(2s+1)}(\log n)^{s(d-1)/(2s+1)}(\log n)^{1/2}$, which is the rate of Theorem 3 with $\mathrm{polylog}(n) = (\log n)^{1/2}$.

Proof. If $\mathcal{C}_n$ fails, $\mathrm{TV}(\widehat{\mathsf{P}}, \mathsf{P}_0) \le 1$. On $\mathcal{C}_n$, let $\widehat{\mathsf{Q}}_\pi$ denote the law at level $\underline{t}$ of the reverse diffusion driven by $\widehat{\boldsymbol{s}}_{t,L}$ and started from π at level $\bar{t}$, so that $\widehat{\mathsf{P}} = \widehat{\mathsf{Q}}_{\mathrm{Unif}}$ and

$$\mathrm{TV}(\widehat{\mathsf{P}}, \mathsf{P}_0) \le \mathrm{TV}(\widehat{\mathsf{Q}}_{\mathrm{Unif}}, \widehat{\mathsf{Q}}_{p_{\bar{t}}}) + \mathrm{TV}(\widehat{\mathsf{Q}}_{p_{\bar{t}}}, p_{\underline{t}}) + \mathrm{TV}(p_{\underline{t}}, p_0).$$

The same Markov kernel acts in the first term, so it is at most $\mathrm{TV}(\mathrm{Unif}, p_{\bar{t}}) \le \frac{R}{2} e^{-2\pi^2\bar{t}}$ (Lemma J.7); the third term is at most $C\underline{t}^{\min(s,2)/2}$ (Lemma 20). For the second, the exact reverse process started from $p_{\bar{t}}$ has marginals p_t and drift $\boldsymbol{s}_t$, which is bounded on $[\underline{t}, \bar{t}]$ by Lemma J.6. Conditionally on training data in $\mathcal{C}_n$, both path laws have unit diffusion and bounded drifts, so Girsanov's theorem gives $\mathrm{KL} = \frac{1}{2}\int_{\underline{t}}^{\bar{t}} \mathcal{E}(t)dt$, and Pinsker's inequality with data processing gives $\mathrm{TV}(\widehat{\mathsf{Q}}_{p_{\bar{t}}}, p_{\underline{t}}) \le \frac{1}{2}(\int \mathcal{E}(t)dt)^{1/2}$. Taking expectations,

$$\mathbb{E}\,\mathrm{TV}(\widehat{\mathsf{P}}, \mathsf{P}_0) \le C\underline{t}^{\frac{\min(s,2)}{2}} + \frac{R}{2} e^{-2\pi^2\bar{t}} + \frac{1}{2}\mathbb{E}\Big[\mathbf{1}_{\mathcal{C}_n}\Big(\int_{\underline{t}}^{\bar{t}} \mathcal{E}(t)dt\Big)^{1/2}\Big] + \mathbb{P}(\mathcal{C}_n^c),$$

and $\mathbb{E}[\mathbf{1}_{\mathcal{C}_n} Z^{1/2}] \le (\mathbb{E}[\mathbf{1}_{\mathcal{C}_n} Z])^{1/2}$ for $Z \ge 0$, so Lemma J.8(iv) gives the first inequality. For the rate, $e^{-2\pi^2\bar{t}} = n^{-2\pi^2}$, $\underline{t}^{\min(s,2)/2} = n^{-s/(2s+1)}$ and $\log(\bar{t}/\underline{t}) \asymp \log n$, and the balance of Section 5.3 gives $2^{-2sL^\star} + dm_{L^\star}/n \asymp n^{-2s/(2s+1)}(\log n)^{2s(d-1)/(2s+1)}$, with $m_{L^\star} \le n$ for large n. Finally, $K_\Psi \le C_\Psi m_{L^\star}$ and $D_\Psi \le C_\Psi 2^{2L^\star} m_{L^\star}$ make Λ_n and N_n polynomial in n, while $n/m_{L^\star} \ge n^{2s/(2s+1)}/\mathrm{polylog}(n)$, so $\mathbb{P}(\mathcal{C}_n^c) \le n^{-3} + N_n m_{L^\star} \exp(-c_0 n\underline{p}/(C_\Psi m_{L^\star})) = O(n^{-1})$. □

J.4 The time integral in the proof of Theorem 5

The proof of Theorem 5 requires bounds on $\int_0^T t^{-1/2}\big(\sum_j a_j(1 + 2^{J(j)}\sqrt{t})^{-2M}\big)^{1/2}dt$ for non-negative weights a_j and $M > 1$. Group the blocks by $J = J(j)$ with $A_J := \sum_{J(j)=J} a_j$, and split the time axis into the dyadic scales $t \in (4^{-k-1}, 4^{-k}]$. On such a scale $t^{-1/2} \le 2^{k+1}$ and $(1 + 2^J\sqrt{t})^{-2M} \le \min\{1, 2^{-2M(J-k-1)}\}$, so the integral is at most

$$C\sum_{k\ge k_T} 2^{-k}\Big(\sum_{J\le k+1} A_J + \sum_{J>k+1} A_J 2^{-2M(J-k-1)}\Big)^{1/2}, \tag{24}$$

where k_T depends only on T. We evaluate (24) for the two terms of (21).

Bias. Here $a_j = b_j 2^{-2s|j|_+}\mathbf{1}_{|j|_+>L}$, and $J(j) \ge |j|_+/d > L/d$, so $A_J = 0$ for $J \le L/d$ and $\sum_{J\le k} A_J \le 2^{-2sL}R^2$. For $k < L/d$ the bracket is at most $2^{-2sL}R^2 2^{-2M(L/d-k-1)}$, and for $k \ge L/d$ it is at most $2^{-2sL}R^2$. Summing the two geometric series gives $\lesssim 2^{-(s+1/d)L}R$ for every d, which is the bias bound used in the proof.

Variance. Here $a_j = n^{-1}2^{|j|_+}\mathbf{1}_{|j|_+\le L}$. A multi-index with maximal coordinate J and $|j|_+ = \ell$ exists only for $\ell \le dJ$, and the number of them is $O((dJ-\ell+1)^{d-2})$. Hence $nA_J \lesssim 2^{dJ}$ for $J \le L/d$, and $nA_J \lesssim 2^L(J-L/d+1)^{d-2}$ for $J > L/d$. Put $J^* = L/d$. For $k \le J^*$ the bracket in (24) is at most a constant times $n^{-1}2^{dk}$ if $2M > d$, $n^{-1}(1+J^*-k)2^{dk}$ if $2M = d$, and $n^{-1}2^{2Mk+(d-2M)J^*}$ if $2M < d$: the sum over $k+1 < J \le J^*$ is dominated by its smallest, constant or largest J, respectively, and the terms with $J > J^*$ are of no larger order. With $h = J^* - k$, the corresponding terms of (24) are at most $n^{-1/2}2^{(d/2-1)J^*}$ times $2^{-(d/2-1)h}$, $2^{-(d/2-1)h}(1+h)^{1/2}$ or $2^{-(M-1)h}$. For $d > 2$ each is summable over $h \ge 0$ because $M > 1$, so the critical case $2M = d$ adds no logarithm. For $k > J^*$ the bracket grows only polynomially in $k - J^*$, while 2^{-k} decays. The sum is therefore of order $n^{-1/2}2^{L/2-J^*} = n^{-1/2}2^{L(1/2-1/d)} \asymp (S_L/n)^{1/2}$ by Lemma 10. For $d = 1$ the terms are $O(n^{-1/2}2^{-k/2})$ and sum to $O(n^{-1/2})$. For $d = 2$ each of the $\asymp L$ scales $k \le L/2$ contributes a term of order $n^{-1/2}$, so the variance part is $O(n^{-1/2}L)$, which exceeds $(S_L/n)^{1/2} \asymp (L/n)^{1/2}$ by a factor $L^{1/2}$.

With this evaluation, the proof of Theorem 5, and that of Theorem 6 which reuses it, yields the displayed bounds, including the case $d > 2$ used by Corollary 2; the term $\mathbf{1}_{\{d=2\}}L/\sqrt{n}$ records the critical case $d = 2$. This concerns the time integral of these two theorems only. The direct density-estimation benchmark, Theorem 4, aggregates squared coefficients instead, and its rate $(\log n/n)^{1/2}$ for $d = 2$ is unchanged.

K One-sided extensions in the Besov summability parameter

Throughout, the integrability parameter is fixed at $p = 2$ and $q \in (0,\infty]$ is the Besov summability parameter. With the tensor wavelet coefficients of Section 2, set, for $j \in \mathbb{N}_{-1}^d \setminus \{-\mathbf{1}\}$,

$$a_j(f) := 2^{s|j|_+}\Big(\sum_{k\in\mathcal{K}_j}\theta_{j,k}(f)^2\Big)^{1/2}, \qquad \|f\|_{B^{s,\mathrm{mix}}_{2,q}} := \big(|\bar f|^2 + \|a(f)\|_{\ell_q}^2\big)^{1/2},$$

where $\|a\|_{\ell_q} = (\sum_j |a_j|^q)^{1/q}$ for $q < \infty$ and $\|a\|_{\ell_\infty} = \sup_j |a_j|$. For $q = 2$ this is exactly the norm (3); for other q it is the usual mixed Besov sequence norm up to equivalence constants (Vybiral, 2006; Dung et al., 2018). Let $\mathcal{B}^{s,\mathrm{mix}}_q(R;\underline{p},\overline{p}) := \{p \in \mathcal{P}(\underline{p},\overline{p}) : \|p\|_{B^{s,\mathrm{mix}}_{2,q}} \le R\}$, so that $\mathcal{B}^{s,\mathrm{mix}}_2(R;\underline{p},\overline{p}) = \mathcal{B}^{s,\mathrm{mix}}(R;\underline{p},\overline{p})$, with the same radius and envelope as in the main text.

Lemma K.1 (Nesting of the density classes)**.** *If $0 < q \le 2 \le r \le \infty$, then $\|f\|_{B^{s,\mathrm{mix}}_{2,r}} \le \|f\|_{B^{s,\mathrm{mix}}_{2,2}} \le \|f\|_{B^{s,\mathrm{mix}}_{2,q}}$ for every f, and hence*

$$\mathcal{B}^{s,\mathrm{mix}}_q(R;\underline{p},\overline{p}) \subseteq \mathcal{B}^{s,\mathrm{mix}}(R;\underline{p},\overline{p}) \subseteq \mathcal{B}^{s,\mathrm{mix}}_r(R;\underline{p},\overline{p}).$$

Proof. It suffices to show $\|a\|_{\ell_{q_2}} \le \|a\|_{\ell_{q_1}}$ for $0 < q_1 \le q_2 \le \infty$, since the mean term is common to all three norms. If $\|a\|_{\ell_{q_1}} = 0$ there is nothing to prove; otherwise, by homogeneity, assume $\|a\|_{\ell_{q_1}} = 1$. Then $|a_j| \le 1$ for all j, so $\sup_j |a_j| \le 1$ and, for $q_2 < \infty$, $\sum_j |a_j|^{q_2} \le \sum_j |a_j|^{q_1} = 1$. The class inclusions follow because the envelope and the radius are the same. □

Corollary K.2 (Immediate consequences of the nesting)**.** *Let $s > 0$ and let d, R, $\underline{p}$, $\overline{p}$ be as in the main text.*

(i) Upper bounds for $0 < q \le 2$. *Every $p_0 \in \mathcal{B}^{s,\mathrm{mix}}_q(R;\underline{p},\overline{p})$ that satisfies the score-regularity condition of Appendix F (or the score-specific input $A_s(p_0) \le A$ of Proposition M.5) satisfies both bounds of Theorem 3, with the same estimator and the same constants. For $s > 1/2$ the condition holds for every member, uniformly, by Proposition J.1; hence the bounds* (18)–(19) *hold uniformly over $\mathcal{B}^{s,\mathrm{mix}}_q(R;\underline{p},\overline{p})$.*

(ii) Lower bounds for $2 \le q \le \infty$. *Both lower bounds of Theorem 1 hold with $\mathcal{B}^{s,\mathrm{mix}}(R;\underline{p},\overline{p})$ replaced by $\mathcal{B}^{s,\mathrm{mix}}_q(R;\underline{p},\overline{p})$, with the same orders.*

Proof. (i) By Lemma K.1, $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R;\underline{p},\overline{p})$. Theorem 3 and Proposition M.5 are statements about each such p_0 under its score input, with constants depending only on $(s,d,R,\underline{p},\overline{p})$, the wavelet

system and the constant of that input; they therefore apply unchanged. For $s > 1/2$, Proposition J.1 gives the multiplier bound with one constant for every density of $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$, in particular for every density of the smaller class. (ii) If $\Theta_1 \subseteq \Theta_2$, then $\sup_{p_0 \in \Theta_2} \mathbb{E}\, \mathsf{d}(\widehat{\mathsf{P}}, \mathsf{P}_0) \geq \sup_{p_0 \in \Theta_1} \mathbb{E}\, \mathsf{d}(\widehat{\mathsf{P}}, \mathsf{P}_0)$ for every estimator $\widehat{\mathsf{P}}$ and loss d, so the minimax risk over Θ_2 is at least that over Θ_1. Apply this with $\Theta_1 = \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$ and $\Theta_2 = \mathcal{B}^{s,\mathrm{mix}}_q(R; \underline{p}, \overline{p})$ (Lemma K.1) and to both bounds of Theorem 1. □

Remark K.3 (What the nesting does not give). The inclusions give no lower bound on the smaller classes $q < 2$ and no upper bound on the larger classes $q > 2$; whether the exponent $s/(2s+1)$ persists there, and which logarithmic powers are sharp for $q \neq 2$, requires separate packing and shell-wise ℓ_q–ℓ_2 arguments that we do not carry out. For $s \leq 1/2$ the score input is not removed by the embedding: part (i) still assumes it for each density. The integrability parameter $p = 2$ is kept throughout.

L Controlled study

Targets. All targets live on $\mathbb{T}^d$ and are built from the same ingredients: four random centers μ_j, wrapped Gaussians φ_{σ^2} with sixteen geometric scales $\sigma_\ell \in [0.004, 0.25]$, and a uniform floor $\lambda = 0.2$. The *mixed* target is a mixture of product densities, $p_0(x) = \lambda + \frac{1-\lambda}{4} \sum_j \prod_r g(x_r - \mu_{jr})$ with $g = \sum_\ell \pi_\ell \varphi_{\sigma_\ell^2}$ and $\pi_\ell \propto \sigma_\ell^{3/2}$. Its Fourier coefficients decay like $\prod_r \max(1, |k_r|)^{-3/2}$ over the resolved range, i.e. mixed smoothness $s = 1$ at the edge of the Hilbert ball (examples (i)–(ii) of Section 2). Two isotropic comparisons use one scale shared by all coordinates, $p_0(x) = \lambda + \frac{1-\lambda}{4} \sum_j \sum_\ell \pi_\ell \prod_r \varphi_{\sigma_\ell^2}(x_r - \mu_{jr})$. With $\pi_\ell \propto \sigma_\ell^{1+d/2}$ the coefficients decay like $|k|^{-1-d/2}$, i.e. isotropic smoothness $s = 1$ (*equal nominal smoothness*). With $\pi_\ell \propto \sigma_\ell^{3/2}$ and the same centers as the mixed target, every one-dimensional marginal coincides with that of the mixed target, while fine scales occur jointly in all coordinates (*matched marginals*). Joint fine scales make these targets far more concentrated: the ratio of the largest to the smallest value of p_0 on 2^{21} probe points (half drawn from p_0, half uniform) has median over seeds 28, 1.3×10^3, 9.5×10^4 and 6.8×10^8 for $d = 2, 3, 4, 6$, against 11, 44, 167 and 1.9×10^3 for the mixed targets and 11, 64, 489 and 1.5×10^4 for the equal-smoothness isotropic targets. Heat smoothing maps $\sigma_\ell^2 \mapsto \sigma_\ell^2 + t$, so p_t and $\boldsymbol{s}_t$ are in closed form; all targets are positive and are sampled exactly. We checked the closed-form scores against finite differences, the samplers against closed-form Fourier moments of the noised laws, the unbiasedness of the denoising target, and the equality of marginals for the matched-marginal targets.

Estimator and selection. At each of 33 noise levels, geometrically spaced in $[3 \times 10^{-5}, 0.3]$, we fit least squares on real Fourier features indexed by a hyperbolic cross, with the fitted score truncated at $(\overline{p}/\underline{p})(2/(\pi t))^{1/2}$, the bound of Lemma J.3; the truncation was never active in any run. Frequencies are ordered by $\prod_r \max(1, |k_r|)$ and then by the number of active coordinates; the candidate models are nested prefixes of this order, starting with the null model. The same n samples and one noise draw per sample are used at all levels, as in Section 5.2. The model size is selected by the denoising loss on a fixed held-out fifth of the sample, with no knowledge of s, and the selected model is refitted on all n points; oracle selection by the true error on an independent sample is reported for reference. The error at each level is the $L^2(p_t)$ error on a fresh sample, and the integrated error is the trapezoidal integral over levels. We use $n = 2^{11}, \ldots, 2^{20}$ (nested), $d \in \{2, 3, 4, 6\}$ and 16 seeds per target and dimension. In every run with a mixed or matched-marginal target the estimator beat the null model at every sample size, and the selected model size grew with n.

Law of the sampler ($d = 2$). The continuous-time reverse sampler of Theorem 3, without the certificate check of Appendix J.3, starts from the uniform law at $\overline{t} = 1$, stops at $\underline{t} = n^{-2/3}$, and uses the selected scores interpolated in $\log t$ between the fitted levels. We compute its law numerically by solving its Fokker–Planck equation (Strang splitting with an exact heat step and fourth-order Runge–Kutta advection on a 256^2 grid; TV evaluated on a 1024^2 grid). With the true score the exact solution is $p_{\underline{t}}$, and the solver reproduced it with TV error at most 6.1×10^{-5} in all oracle runs.

Results. Table 2 and Figure 1 summarize the study. For mixed targets the slopes are -0.67 in $d = 2$, the one-dimensional exponent, and flatten with d by about the size of the logarithmic factor in (17): over the same range of n, the local slope of $n^{-2/3}(\log n)^{2(d-1)/3}$ is -0.60, -0.54, -0.48

Table 2: Slopes of the integrated score error versus n (held-out selection; mean $\pm$ s.d. over 16 seeds) and paired per-seed slope differences, isotropic minus mixed, with the number of positive seeds. Asymptotic exponents: $-2/3$ for the mixed class and $-2/(2+d)$ for isotropic smoothness 1. †The isotropic cell for $d = 4$ failed the registered validity check (estimator no better than the null model at $n = 2^{11}$ in 4 of 16 seeds) and is not interpretable.

	mixed	isotropic, equal nominal smoothness		isotropic, matched marginals	
d	slope	slope	difference (> 0)	slope	difference (> 0)
2	-0.67 ± 0.03	-0.58 ± 0.03	$+0.09 \pm 0.05$ (16/16)	-0.51 ± 0.04	$+0.15 \pm 0.04$ (16/16)
3	-0.56 ± 0.02	-0.53 ± 0.03	$+0.02 \pm 0.04$ (12/16)	-0.42 ± 0.01	$+0.13 \pm 0.02$ (16/16)
4	-0.47 ± 0.01	-0.54 ± 0.02	-0.07 ± 0.03 (0/16)†	-0.38 ± 0.01	$+0.10 \pm 0.01$ (16/16)
6	-0.38 ± 0.007	-0.46 ± 0.03	-0.08 ± 0.02 (0/16)	-0.34 ± 0.01	$+0.04 \pm 0.01$ (16/16)

and -0.35 for $d = 2, 3, 4, 6$. This comparison was made after the run. The TV of the sampler in $d = 2$, computed from the numerical Fokker–Planck solution, decays with slope -0.32 ± 0.02 for the mixed target (from 0.089 at $n = 2^{11}$ to 0.011 at $n = 2^{20}$), against $-1/3$ predicted. For the matched-marginal target the slope is -0.24 ± 0.02, reaching 0.019 at $n = 2^{20}$. At equal feature budgets the hyperbolic-cross order has lower error than the full-grid order at $n = 2^{20}$ in 16 of 16 seeds for $d = 2, 3, 4$ (median error ratios 1.46, 1.54 and 1.22), and the gap grows with n; in $d = 6$ the two orders coincide on the first 861 features, which contain most selected models. The registered secondary outcome, the TV bound of the decomposition (8) evaluated with numerically estimated quantities (the early-stopping term by Monte Carlo, the score term through Pinsker's inequality from the integrated error on $[\underline{t}, 0.3]$ plus the score energy above 0.3, and the initialization term), decreases with n on average in every cell; it is an estimate, not a certified bound, and individual seed curves need not decrease at every n. For the mixed targets its mean slopes are -0.32, -0.26, -0.22 and -0.17 and its mean values at $n = 2^{20}$ are 0.032, 0.066, 0.115 and 0.240 for $d = 2, 3, 4, 6$.

The pre-registered primary comparison, with the equal-smoothness isotropic target, separated the two targets only in $d = 2$. At equal nominal smoothness the isotropic target decays like $|k|^{-1-d/2}$ along each axis, much faster than the mixed target, so at reachable n its error is dominated by low frequencies and its slope stays far from its asymptote $-2/(2+d)$. The matched-marginal control was designed after this outcome to remove that confound; its protocol was fixed before it was run. It separates the targets in every dimension, in all 16 seeds. Because the matched-marginal targets are also far more concentrated (see Targets), this control changes the envelope constants of Section 2 together with the interaction structure, and it does not isolate the latter.

Protocol history. The protocols, including the analysis scripts, were recorded with content hashes before the corresponding main runs. Pilot runs on separate seeds were used only to check that the estimator operates in the intended regime and to set dictionary caps. An earlier study with a different design was inconclusive: it had no null model, its targets were nearly uniform, and its noise window did not reach the levels where the problem is nonparametric. Its protocol and data will be included in the post-publication release.

M Scope of the assumptions, practical assessment and related analyses

This appendix collects results on the scope of the multiplier condition (Sections M.1 and M.2), on assessing the model in practice (Section M.3), on sparse-grid versus neural score estimators (Section M.4) and on closely related analyses (Section M.5). The theorems of the main text are unchanged.

M.1 Rough densities that satisfy the multiplier condition

Proposition J.1 proves the multiplier condition of Appendix F on the whole class for every $s > 1/2$; for $s \le 1/2$ it is assumed. This subsection shows that the condition holds for explicit families at $0 < s < 1/2$, including densities that are genuinely rough, and at the endpoint $s = 1/2$, and it records the exact scope. We use the norm $\|\cdot\|_{s,d}$, the spaces $H^a(\mathbb{T}; X)$ and the equivalence constant c_{eq} of

Appendix J.1. Throughout, X is a complex Hilbert space, $|x|_{\mathbb{T}}$ is the distance from x to the nearest integer, an *arc* is the image in $\mathbb{T}$ of an interval $[\alpha,\beta)$ with $0\le\beta-\alpha\le 1$, and $\mathrm{Var}(a):=\int_{\mathbb{T}}|a'|$ for $a\in C^1(\mathbb{T})$.

Lemma M.1 (Arcs and functions of bounded variation are multipliers for $s<1/2$)**.** *Let $0<s<1/2$. There is a constant C_s, depending only on s, such that for every complex Hilbert space X and every $u\in H^s(\mathbb{T};X)$:*

(a) $\int_{\mathbb{T}}\|u(x)\|_X^2\,|x-e|_{\mathbb{T}}^{-2s}\,dx\le C_s\|u\|^2_{H^s(\mathbb{T};X)}$ *for every $e\in\mathbb{T}$;*

(b) $\|\mathbf{1}_I u\|_{H^s(\mathbb{T};X)}\le C_s\|u\|_{H^s(\mathbb{T};X)}$ *for every arc I;*

(c) $\|au\|_{H^s(\mathbb{T};X)}\le\big(\|a\|_\infty+C_s\mathrm{Var}(a)\big)\|u\|_{H^s(\mathbb{T};X)}$ *for every real $a\in C^1(\mathbb{T})$.*

Proof. (a) Translations are isometries of $H^s(\mathbb{T};X)$, so take $e=0$. On $\mathbb{R}$, the proof of Lemma J.1(c), with the Fourier transform in place of Fourier series, gives $\|v\|^2_{H^s(\mathbb{R};X)}\asymp\|v\|^2_{L^2}+\iint\|v(x)-v(y)\|_X^2|x-y|^{-1-2s}dx\,dy$. For $w\in H^s(\mathbb{R};X)$, Lindemulder et al. (2018, Theorem 4.1), applied with $p=2$ and trivial weight (a Hilbert space is a UMD space), gives $\|\mathbf{1}_{(0,\infty)}w\|_{H^s(\mathbb{R};X)}\lesssim\|w\|_{H^s(\mathbb{R};X)}$ because $|s|<1/2$. In the double integral of $v=\mathbf{1}_{(0,\infty)}w$, the pairs $x>0>y$ contribute $\int_0^\infty\|w(x)\|_X^2\int_{-\infty}^0(x-y)^{-1-2s}dy\,dx=(2s)^{-1}\int_0^\infty\|w(x)\|_X^2x^{-2s}dx$. Applying this to w and to $w(-\cdot)$ gives the Hardy inequality $\int_{\mathbb{R}}\|w(x)\|_X^2|x|^{-2s}dx\lesssim\|w\|^2_{H^s(\mathbb{R};X)}$. Now view u as a 1-periodic function on $\mathbb{R}$ and fix $\chi\in C_c^\infty((-\frac12,\frac12))$ with $0\le\chi\le1$ and $\chi=1$ on $[-\frac14,\frac14]$. We claim $\|\chi u\|_{H^s(\mathbb{R};X)}\lesssim\|u\|_{H^s(\mathbb{T};X)}$. Write $\|\chi(x)u(x)-\chi(y)u(y)\|_X\le\chi(x)\|u(x)-u(y)\|_X+|\chi(x)-\chi(y)|\,\|u(y)\|_X$. In the first term, the pairs with $|x|<\frac12$ and $|x-y|\le\frac12$ map injectively into $\mathbb{T}^2$ and satisfy $|x-y|=|x-y|_{\mathbb{T}}$, so they contribute at most the double integral of u on $\mathbb{T}$. For the pairs with $|x-y|>\frac12$ we use $\|u(x)-u(y)\|_X^2\le2\|u(x)\|_X^2+2\|u(y)\|_X^2$; after integration in the other variable, these pairs and the second term have weights bounded by $C(1+|y|)^{-1-2s}$ times $\|u(y)\|_X^2$ (or a constant times $\|u(x)\|_X^2$ with $|x|<\frac12$), and summing over periods bounds them by $C\|u\|^2_{L^2(\mathbb{T};X)}$. This proves the claim, and hence

$$\int_{\mathbb{T}}\|u(x)\|_X^2|x|_{\mathbb{T}}^{-2s}dx\le\int_{\mathbb{R}}\|\chi(x)u(x)\|_X^2|x|^{-2s}dx+4^{2s}\|u\|^2_{L^2(\mathbb{T};X)}\lesssim\|u\|^2_{H^s(\mathbb{T};X)}.$$

(b) Let $v=\mathbf{1}_I u$ and use Lemma J.1(c). Clearly $\|v\|_{L^2}\le\|u\|_{L^2}$. In the double integral, pairs in $I\times I$ contribute at most the double integral of u, and pairs outside I contribute nothing. If $x\in I$ and $y\notin I$, then $|x-y|_{\mathbb{T}}\ge\delta(x):=\min_i|x-e_i|_{\mathbb{T}}$, where e_1,e_2 are the endpoints of I, and $\int_{\{|h|_{\mathbb{T}}\ge\delta\}}|h|_{\mathbb{T}}^{-1-2s}dh\le s^{-1}\delta^{-2s}$. Since $\delta(x)^{-2s}\le\sum_i|x-e_i|_{\mathbb{T}}^{-2s}$, the mixed pairs contribute at most $2s^{-1}\sum_{i=1,2}\int_{\mathbb{T}}\|u(x)\|_X^2|x-e_i|_{\mathbb{T}}^{-2s}dx$, which (a) bounds. The constant depends neither on the position nor on the length of I; the argument uses the metric of $\mathbb{T}$ directly, so short and wrapping arcs need no separate treatment.

(c) Identify $\mathbb{T}$ with $[0,1)$. Then $a(x)=a(0)+\int_0^1a'(y)\mathbf{1}_{[y,1)}(x)\,dy$, and each $[y,1)$ is an arc. For an X-valued trigonometric polynomial g and the $L^2(\mathbb{T};X)$ pairing $\langle\cdot,\cdot\rangle$, Fubini's theorem gives $\langle au,g\rangle=a(0)\langle u,g\rangle+\int_0^1a'(y)\langle\mathbf{1}_{[y,1)}u,g\rangle\,dy$, so by (b) $|\langle au,g\rangle|\le(|a(0)|+C_s\mathrm{Var}(a))\|u\|_{H^s(\mathbb{T};X)}\|g\|_{H^{-s}(\mathbb{T};X)}$, where $\|g\|^2_{H^{-s}(\mathbb{T};X)}:=\sum_\kappa\langle\kappa\rangle^{-2s}\|\widehat g(\kappa)\|_X^2$. Taking the supremum over such g with $\|g\|_{H^{-s}(\mathbb{T};X)}\le1$ gives (c). □

Proposition M.2 (A genuinely rough family for $0<s<1/2$)**.** *Let $0<s<1/2$, let $h:=\mathbf{1}_{[0,1/2)}-\mathbf{1}_{[1/2,1)}$ on $\mathbb{T}$, fix $0<a_*<1$, and for $a\in[-a_*,a_*]^d$ let*

$$p_a(x):=\prod_{r=1}^d\big(1+a_r\,h(x_r)\big),\qquad x\in\mathbb{T}^d.$$

(i) p_a is a probability density with $(1-a_*)^d\le p_a\le(1+a_*)^d$, and $\|p_a\|^2_{B^{s,\mathrm{mix}}_{2,2}}=\prod_{r=1}^d(1+a_r^2\kappa_s^2)$ with $\kappa_s:=\|h\|_{B^s_{2,2}(\mathbb{T})}<\infty$ (the case $d=1$ of (3)). Hence every p_a lies in $\mathcal{B}^{s,\mathrm{mix}}(R;\underline p,\overline p)$ as soon as $(1-a_*)^d\ge\underline p$, $(1+a_*)^d\le\overline p$ and $(1+a_*^2\kappa_s^2)^d\le R^2$. If $a_r\neq0$ for some r, then $p_a\notin H^{1/2}_{\mathrm{mix}}(\mathbb{T}^d)$.

(ii) For every $t > 0$ and every g,

$$\|g/p_{a,t}\|_{s,d} \le \mu_s(a_*)^d \, \|g\|_{s,d}, \qquad \mu_s(a_*) := (1-a_*)^{-1} + 4C_s a_*(1-a_*)^{-2},$$

with C_s from Lemma M.1. Hence the multiplier condition of Appendix F holds at every $t > 0$, uniformly over the family, with $M_s = c_{\mathrm{eq}}^2 \, \underline{p} \, \mu_s(a_*)^d$.

Proof. (i) The function h has mean zero and $|h| = 1$, so each factor $q_{a_r} := 1 + a_r h$ is a density with values in $[1-a_*, 1+a_*]$. Tensor wavelet coefficients of a product factorize, so (3) of p_a is the product of the one-dimensional norms (Section 2, example (i)), and $\|q_a\|^2_{B^s_{2,2}(\mathbb{T})} = 1 + a^2\kappa_s^2$ because the mean of q_a is 1 and its other coefficients are those of ah. Since $\widehat{h}(k) = 2/(\pi i k)$ for odd k and $\widehat{h}(k) = 0$ otherwise, $\sum_k \langle k\rangle^{2\sigma} |\widehat{h}(k)|^2 < \infty$ exactly when $\sigma < 1/2$. By norm equivalence $\kappa_s < \infty$, while $\|p_a\|_{1/2,d} = \prod_r \|q_{a_r}\|_{H^{1/2}(\mathbb{T})} = \infty$ if some $a_r \neq 0$, since every factor has norm at least $|\widehat{q}_{a_r}(0)| = 1$.

(ii) The heat kernel of $\mathbb{T}^d$ is a product of one-dimensional heat kernels, so $p_{a,t}(x) = \prod_r q_{a_r,t}(x_r)$ with $q_{a,t} = 1 + a h_t$ and $h_t := h * K_t^{\mathbb{T}} \in C^\infty(\mathbb{T})$. Convolution with a probability density gives $|h_t| \le 1$ and $\mathrm{Var}(h_t) \le \mathrm{Var}(h) = 4$. Hence $1 - a_* \le q_{a,t} \le 1 + a_*$, and $b_t := 1/q_{a,t}$ satisfies $\|b_t\|_\infty \le (1-a_*)^{-1}$ and $\mathrm{Var}(b_t) = \int_{\mathbb{T}} |a\, h_t'| \, q_{a,t}^{-2} \le 4a_*(1-a_*)^{-2}$. By Lemma M.1(c), multiplication by b_t has norm at most $\mu_s(a_*)$ on $H^s(\mathbb{T}; X)$ for every Hilbert space X. Singling out the coordinate r, $H^s_{\mathrm{mix}}(\mathbb{T}^d) = H^s(\mathbb{T}; X_r)$ isometrically with $X_r = H^s_{\mathrm{mix}}(\mathbb{T}^{d-1})$ (Appendix J.1), and multiplication by $b_{r,t}(x_r) := 1/q_{a_r,t}(x_r)$ acts on the variable x_r only. Writing $1/p_{a,t} = \prod_r b_{r,t}(x_r)$ and applying the d factors in turn gives the bound, which depends neither on t nor on a. The constant c_{eq} converts it to the norm (3). □

Remark M.3 (Scope of the rough family). (1) Part (ii) uses only the lower bound and the variation of each factor. It therefore applies verbatim to every product $\prod_r q_r(x_r)$ of periodic densities of bounded variation with $q_r \ge c > 0$, with $1 - a_*$ and $4a_*$ replaced by c and $\mathrm{Var}(q_r)$; histogram densities with finitely many bins in each coordinate are examples. Each such factor lies in $H^s(\mathbb{T})$ for every $s < 1/2$, since $\|q(\cdot + h) - q\|_{L^2}^2 \le 2\|q\|_\infty |h|_{\mathbb{T}} \mathrm{Var}(q)$, so membership in $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$ reduces to the radius and envelope checks of part (i). (2) The family is fixed: a_* does not depend on n. For fixed $(R, \underline{p}, \overline{p})$, a_* must decrease with d (the envelope requires $a_* \lesssim d^{-1}$), and the constants depend on d; as everywhere in the paper, d is fixed. (3) Theorem 3 applies to every member, with constants uniform over the family. The family has only d parameters, so it is not a hard subclass: neither the lower bound of Theorem 1 nor any statement about the whole ball $\mathcal{B}^{s,\mathrm{mix}}$ at $s < 1/2$ follows from it.

Proposition M.4 (Endpoint $s = 1/2$)**.** *Let $0 < b_* < 1$ and $p_b(x) := \prod_{r=1}^d (1 + b_r \cos 2\pi x_r)$ with $|b_r| \le b_*$. Then $(1-b_*)^d \le p_b \le (1+b_*)^d$ and $\|p_b\|^2_{B^{1/2,\mathrm{mix}}_{2,2}} = \prod_r (1 + b_r^2 \kappa'^2)$ with $\kappa' := \|\cos 2\pi \cdot\|_{B^{1/2}_{2,2}(\mathbb{T})}$. There is an absolute constant C' such that for every $t \ge 0$ and every g,*

$$\|g/p_{b,t}\|_{1/2,d} \le \nu(b_*)^d \|g\|_{1/2,d}, \qquad \nu(b_*) := C'\big((1-b_*)^{-1} + 2\pi b_*(1-b_*)^{-2}\big).$$

Proof. The first two claims follow as in Proposition M.2(i). Since $\cos 2\pi x$ is an eigenfunction of the heat semigroup, $p_{b,t} = p_{b'}$ with $b_r' = b_r e^{-2\pi^2 t}$, so it suffices to take $t = 0$. The factor $a := 1/(1 + \beta \cos 2\pi x)$ with $|\beta| \le b_*$ satisfies $\|a\|_\infty \le (1-b_*)^{-1}$ and $|a(x) - a(y)| \le \Lambda |x-y|_{\mathbb{T}}$ with $\Lambda := 2\pi b_*(1-b_*)^{-2}$. In $\|a(x)u(x) - a(y)u(y)\|_X \le |a(x)| \, \|u(x) - u(y)\|_X + \Lambda |x-y|_{\mathbb{T}} \|u(y)\|_X$, the second term contributes at most $2\Lambda^2 \|u\|^2_{L^2(\mathbb{T};X)}$ to the double integral of Lemma J.1(c) with $a = 1/2$, whose kernel is $|x-y|_{\mathbb{T}}^{-2}$. Hence $\|au\|_{H^{1/2}(\mathbb{T};X)} \le C'(\|a\|_\infty + \Lambda)\|u\|_{H^{1/2}(\mathbb{T};X)}$, and the d factors are applied in turn as in Proposition M.2(ii). □

The cosine family is smooth: it shows that the endpoint condition is satisfied by nonconstant densities with any prescribed radius and envelope (for small b_*), not that it holds on the endpoint ball. The step family does not reach $s = 1/2$, because $h \notin H^{1/2}(\mathbb{T})$ and indicators of arcs are not multipliers of $H^{1/2}(\mathbb{T})$. For $s > 1/2$, Proposition J.1 covers the whole class, including the products, mixtures, low-order interactions and Gibbs densities of Section 2.

M.2 A SCORE-SPECIFIC SUFFICIENT INPUT

The multiplier condition requires division by p_t to be bounded for *every* $g \in B^{s,\mathrm{mix}}_{2,2}$. The proof of Theorem 3 divides only $g = \partial_r p_t$, and uses the quotient $\boldsymbol{s}_{t,r} = \partial_r p_t / p_t$ only through the error of its best sparse-grid approximation. For $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$ define

$$A_s(p_0)^2 := \sup_{t>0,\ L\geq 0} t\, 2^{2sL} \sum_{r=1}^{d} \inf_{v \in \mathcal{V}_L^{\mathrm{SG}}} \|\boldsymbol{s}_{t,r} - v\|^2_{L^2(p_t)}. \tag{25}$$

Proposition M.5 (A score-specific sufficient input). *(i)* $A_s(p_0)^2 \leq \overline{p}\, \sup_{t>0} t \sum_{r=1}^{d} \|\boldsymbol{s}_{t,r}\|^2_{B^{s,\mathrm{mix}}_{2,2}}$. *If the multiplier condition holds for all $t > 0$ with constant M_s (as under Propositions J.1, M.2 and M.4), then $A_s(p_0) \lesssim \sqrt{d}\, M_s \underline{p}^{-1} R$.*

(ii) Fix $A < \infty$. Lemma J.5, Lemma J.8(iv), Corollary J.1 and Proposition J.3, and hence both bounds of Theorem 3, hold with the multiplier condition replaced by $A_s(p_0) \leq A$, uniformly over $\{p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p}) : A_s(p_0) \leq A\}$ and with constants depending on A in place of M_s. In particular, Corollary 1 holds whenever $A_s \leq A$ uniformly over the model class.

Proof. (i) The best approximation in $L^2(p_t)$ is at most the $L^2(p_t)$ error of the L^2 projection, so $p_t \leq \overline{p}$ and (23) give $\inf_v \|\boldsymbol{s}_{t,r} - v\|^2_{L^2(p_t)} \leq \overline{p}\, 2^{-2sL} \|\boldsymbol{s}_{t,r}\|^2_{B^{s,\mathrm{mix}}_{2,2}}$; Lemma 18 gives the second claim.

(ii) In the proof of Lemma J.5 the multiplier condition enters only in the last step, through the bound $\|g\|^2_{L^2(p_t)} \lesssim t^{-1} 2^{-2sL}$ for the residual g of the $L^2(p_t)$ projection of $\boldsymbol{s}_{t,r}$ onto $\mathcal{V}_L^{\mathrm{SG}}$. Summed over r, these residuals are exactly the quantity in (25), so $A_s(p_0) \leq A$ gives the same bound. The other ingredients, namely Lemmas J.3, J.4, J.6, J.7 and 20 and parts (i)–(iii) of Lemma J.8, use only the envelope and $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$, and the remaining listed results use the multiplier condition only through Lemma J.5. □

Remark M.6 (What the weaker input does and does not say). By part (i), every density covered by the multiplier condition, including all examples above, satisfies $A_s < \infty$. For a product density $p_0 = \prod_r q_r(x_r)$, the score component $\boldsymbol{s}_{t,r} = q'_{r,t}/q_{r,t}$ depends on x_r only, and its mixed norm equals its one-dimensional norm, so $A_s(p_0)^2 \leq \overline{p} \sum_r \sup_{t>0} t\, \|q'_{r,t}/q_{r,t}\|^2_{B^s_{2,2}(\mathbb{T})}$ involves only one-dimensional scores. Appendix N establishes A_s uniformly on the one-dimensional and product classes and proves strict separation from the multiplier condition for $0 < s \leq 1/2$. Uniform control on general joint low-smoothness balls, and necessity for the diffusion endpoint rate, remain unresolved. The hypotheses of Theorem 3 and Corollary 1 are unchanged.

M.3 WHERE THE STRUCTURE MAY HOLD, AND ASSESSING IT AT FINITE RESOLUTION

Settings. Dominating mixed smoothness is a statement about a fixed, meaningful coordinate system: it allows sharp variation along single coordinates and penalizes fine scales that occur jointly in many coordinates. Settings where this is plausible are those in which coordinates are interpretable variables whose dependence is dominated by low-order interactions: product-form models and their mixtures, additive or ANOVA-type log-densities with pairwise or few-variable terms, and Gibbs densities with local potentials (Section 2). The torus is the natural domain for angular variables such as phases, directions and molecular torsion angles; diffusion models on the hypertorus of torsion angles are used for molecular conformer generation (Jing et al., 2022). These are settings in which the assumption *may* hold; we make no claim about how often it holds in practice.

A finite-resolution diagnostic. Membership in $\mathcal{B}^{s,\mathrm{mix}}$ concerns all scales and cannot be established from finitely many samples. What a sample can test is whether the wavelet energy of p_0 on the resolved scales behaves as the model predicts. For $l \geq 0$ let $\mathcal{S}_l := \{(j,k) : j \neq -\mathbf{1},\ |j|_+ = l,\ k \in \mathcal{K}_j\}$, $N_l := |\mathcal{S}_l| \asymp 2^l l^{d-1}$ and $E_l := \sum_{\lambda \in \mathcal{S}_l} \theta_\lambda^2$ with $\theta_\lambda = \mathbb{E}\, \Psi_\lambda(X)$, so that $\|p_0\|^2_{B^{s,\mathrm{mix}}_{2,2}} = 1 + \sum_{l \geq 0} 2^{2sl} E_l$ by (3).

1. *Estimate each shell energy without bias.* The plug-in value $\sum_{\lambda \in \mathcal{S}_l} \widehat{\theta}_\lambda^2$ is biased upward by $n^{-1} \sum_{\lambda \in \mathcal{S}_l} \mathrm{Var}\, \Psi_\lambda(X)$, which is of order N_l/n, grows with l, and by itself makes fine shells

look rough. The U-statistic

$$\widehat{E}_l := \frac{1}{n(n-1)} \sum_{\lambda \in \mathcal{S}_l} \Big[\Big(\sum_{i=1}^{n} \Psi_\lambda(X_i) \Big)^2 - \sum_{i=1}^{n} \Psi_\lambda(X_i)^2 \Big]$$

is unbiased for E_l; so is $\sum_\lambda \widehat{\theta}_{A,\lambda} \widehat{\theta}_{B,\lambda}$ computed from two independent halves A, B of the sample.

2. *Report the noise.* $\widehat{E}_l$ can be negative and should be reported without clipping, with error bars from resampling whole observations or from independent split replicates. The Hoeffding decomposition of $\widehat{E}_l$ gives $\mathrm{sd}(\widehat{E}_l) \le \big(4\overline{p} E_l/n + 2\overline{p}^2 N_l/(n(n-1))\big)^{1/2}$, so shells with E_l below the noise floor $\overline{p}\sqrt{2N_l}/n$ are not resolved and should not enter a slope fit. If $E_l \asymp 2^{-2sl}$, the resolved shells are those with $2^{(2s+1/2)l} l^{(d-1)/2} \lesssim n$, somewhat finer than the resolution $L^\star$ of the estimator, because a quadratic functional is easier to estimate than the density itself.
3. *Compare with the model.* Plot $\widehat{E}_l$ against l together with the envelope 2^{-2sl} for candidate values of s, and the weighted cumulative sums $\sum_{l \le L} 2^{2sl} \widehat{E}_l$ over the resolved range. A cumulative sum that stabilizes is consistent with $B^{s,\mathrm{mix}}_{2,2}$ at that resolution. A matching slope alone is not, since $\sum_l 2^{2sl} E_l$ diverges when $E_l \asymp 2^{-2sl}$.
4. *Locate the interactions.* Split each shell by the active set $\{m : j_m \ge 0\}$. Energy concentrated on small active sets supports low-order interaction structure; energy that spreads over large active sets at fixed l reveals higher-order interactions but does not by itself identify an isotropic smoothness class. The energy captured by hyperbolic and by isotropic index sets with the same number of coefficients, or the held-out denoising score-matching loss of the two truncations, gives a matched-budget comparison.
5. *State the coordinates.* The diagnostic depends on the coordinate system and on the scaling of each variable; a rotation can move a density out of the mixed regime. The axes, their units and the resolved range should be reported with the result.

This is a check of finite-resolution plausibility. It does not prove membership at all scales, it does not verify the radius, the envelope or the multiplier condition, and it is not needed to run the estimator. Appendix O implements a truncated Fourier analogue on observed wind directions; dependence-aware resampling there is a sensitivity analysis, not the independent-sample guarantee above.

M.4 Sparse-grid and neural score estimators

Representation versus learning. Sparse-grid structure also admits a constructive neural representation. Using the spline construction of Suzuki (2019), fix d and a spline degree $r > s+1$, and consider an expansion in periodized cardinal tensor B-splines $f_L = \sum_{\lambda=1}^{M} a_\lambda B_\lambda$ with $M = O_{d,r}(m_L)$ terms on a hyperbolic cross and $A = \sum_\lambda |a_\lambda|$. The hat function $(1-|u|)_+ = \sigma(u+1) - 2\sigma(u) + \sigma(u-1)$, with $\sigma(u) = \max\{u, 0\}$, is represented exactly by ReLU units. Higher-degree splines and their tensor products can be approximated by ReLU multiplication modules; periodicization uses only a fixed number of shifted copies for fixed d, r. Approximating each B_λ to accuracy $\varepsilon/(1+A)$ and summing gives $\|f_L - \Phi\|_\infty \le \varepsilon$ with $O_{d,r}(M[1 + \log((1+A)/\varepsilon)])$ nonzero parameters, for $0 < \varepsilon \le 1$. Thus, at $L = L^\star = O(\log n)$, if A and ε^{-1} are polynomially bounded in n, a componentwise neural representation has $\widetilde{O}(m_{L^\star}) = n^{1/(2s+1)} \,\mathrm{polylog}(n)$ nonzero parameters for fixed d. This recovers the sparse-grid effective size at the representation level. The construction uses a spline dictionary with the same hyperbolic-cross geometry, not an exact neural realization of the wavelet basis used in our estimator; the coefficient budget is explicit, rather than uniform over an unbounded linear space. It is a spatial construction at each noise level, not a guarantee for a single time-conditioned network. The separate learning requirement is described next.

Statistical rates. Theorem 1 bounds the risk of every estimator, so no score parameterization, neural or otherwise, improves on $n^{-s/(2s+1)} (\log n)^{s(d-1)/(2s+1)}$ in TV uniformly over the class, and Theorem 2 converts the integrated score error of any estimator into TV. Suzuki (2019) shows that deep ReLU networks attain mixed-smooth rates in L^2 *regression*; a diffusion analogue additionally needs a generalization bound for time-indexed network score matching, whose response variance grows like t^{-1}, and we do not provide one. We therefore claim neither that sparse grids outperform

neural scores on this class nor the converse. What is proved is that a convex estimator with explicit complexity attains the TV minimax scale in the coordinate-aligned regime.

Computation. Table 3 compares the costs at one noise level for n training pairs, $m = m_L$ basis functions and d score components. The sparse-grid counts are derived operation counts for dense linear algebra, not measurements; for the neural score they depend on the architecture, the number U of optimizer steps, the batch size B and the per-example costs $C_{\mathrm{train}}(P, d)$ and $C_{\mathrm{eval}}(P, d)$ of a network with P parameters. We report no speed or memory comparison. Compactly supported wavelets make most basis values zero at any point, which sparse implementations can exploit; we have not measured this. The polynomially many Gram matrices of Lemma J.8 belong to the certified output rule used in the proof, not to the per-level fit.

Table 3: Sparse-grid least squares and a neural score model. Derived operation counts, not measurements.

Aspect	Sparse-grid least squares (Section 5)	Time-conditioned neural score
Statistical guarantee	TV minimax scale under the multiplier (or score) condition, Theorem 3	lower bound of Theorem 1 applies; no diffusion rate proved here
Training, one level	Gram $O(nm^2)$, right-hand sides $O(nmd)$, factorization $O(m^3)$, solves $O(m^2 d)$	$U \cdot B \cdot C_{\mathrm{train}}(P, d)$, architecture dependent
Across K levels	K separate fits (no parameter sharing)	one network shared through time conditioning
Memory	$O(m^2 + md)$ while fitting; $O(Kdm)$ stored coefficients	$O(P)$ parameters plus optimizer state
Score evaluation	$O(dm)$ after the basis values	$C_{\mathrm{eval}}(P, d)$
Interpretation	coefficient indexed by output, active coordinates, scale and location	learned features, not tied to coordinates
Suited structure	meaningful coordinates, low-order interactions, fixed moderate d	unknown, rotated or curved structure

Interpretability. The fitted score is $\widehat{\boldsymbol{s}}_{t,L,r} = \sum_\lambda a_{r,\lambda} \Psi_\lambda$ with $\lambda = (j, k)$, so every coefficient belongs to an output coordinate r, a set of active input coordinates $\{m : j_m \geq 0\}$, a resolution j and a location k. Aggregating $a^2_{r,\lambda}$ by active set and by $|j|_+$ shows which interactions and which scales the fitted score uses, and the same aggregation of the data coefficients is the diagnostic of Section M.3. This is interpretation at the level of the representation, not a causal explanation, and it depends on the chosen coordinates.

Structural regimes. The sparse-grid estimator is designed for coordinate-aligned mixed smoothness: meaningful coordinates, dependence dominated by low-order interactions and a fixed, moderate dimension. Its guarantees need not transfer to coordinates in which the mixed regularity is lost. Neural score features need not be tied to a prescribed coordinate-aligned sparse-grid basis; their performance on other structures depends on the architecture and training, and is not compared here.

M.5 Relation to closely related analyses

Oko et al. Theorem 2 restates the score-to-distribution reduction of Oko et al. (2023), and we use the same pipeline of forward noising, denoising score matching and reverse sampling. The following steps are specific to the mixed-smooth class and are proved here: the hyperbolic-shell lower bound with the factor $(\log n)^{s(d-1)/(2s+1)}$ (Theorem 1, Appendix B); score regularity in the mixed class, namely the inverse-closure argument for $s > 1/2$ (Proposition J.1) and, at $s \leq 1/2$, the examples and the score-specific input of this appendix; the random-design analysis of the linear sparse-grid score estimator with one Gram certificate for all noise levels and the certified output rule (Appendix J.2, Lemma J.8, Proposition J.3); and the balanced-block W_1 lower bound, which shows that balanced mixed interactions have transport scale $2^{-L/d}$ (Proposition 1).

Dimension-free rates under a score-side assumption. Cole & Lu (2024) prove that score-based generative models attain a dimension-independent TV rate for sub-Gaussian targets whose log-relative density with respect to the standard Gaussian is locally approximable by a neural network

with suitably bounded parameters. Their assumption is placed on the log-relative density rather than on a smoothness class, and no lower bound is involved, so the two results are not directly comparable; ours concerns dominating mixed smoothness on the torus with a matching lower bound.

Flow matching. Zhou & Liu (2025) give an end-to-end analysis of flow matching: the velocity field of a continuous normalizing flow with a Gaussian source is learned by least squares over Lipschitz-constrained ReLU networks, and the Euler-discretized flow is stopped early. For targets supported on $[0,1]^d$ they prove consistency in probability in W_2; under an additional Lipschitz condition on the target score (a potential with bounded Hessian), they obtain $W_2 = \widetilde{O}(n^{-1/(3(d+5))})$ with probability at least $1 - 1/n$ for their prescribed n-dependent step size and stopping time (Zhou & Liu, 2025, Theorems 1.6 and 1.7). Their analysis concerns a deterministic flow, the W_2 metric and a class defined by support and score regularity; ours concerns the stochastic reverse diffusion, TV, and densities with dominating mixed smoothness, for which we also prove the matching lower bound. The dimension enters the exponent of their rate and only logarithmic factors and constants of ours, but the metrics and classes differ, so we draw no ranking between the results.

Discrete-state diffusion. Wakasugi & Suzuki (2025) study diffusion models on a *finite state space* of size M; "discrete" refers to the states, not to time discretization. They learn density-ratio scores with a score-entropy (Bregman) loss over ReLU networks and bound $\mathrm{KL}(p_\delta \| \widehat{q}_{T-\delta})$ between the forward law at an early-stopping time δ and the output. A naive bound grows linearly in M. Using a polynomial decay of the coefficients of p_0 in the eigenbasis of the rate matrix and anisotropic-Besov approximations of the eigenvectors on a continuous embedding, they remove the polynomial dependence on M; on the hypercube $\{0,1\}^D$ the sample exponent does not involve D and $M = 2^D$ enters only logarithmically when their decay exponent is at least one (Wakasugi & Suzuki, 2025, Corollary 1). Pinsker's inequality turns this into TV to the same law p_δ; a TV statement about p_0 also needs the early-stopping term. Their work, like ours, uses structure to avoid a cost exponential in the dimension; the mechanisms (spectral decay on a finite state space versus dominating mixed smoothness of a continuous density), the losses and the classes differ, and no ranking between the results follows.

N Low smoothness: the score input beyond the multiplier condition

This appendix sharpens Appendices M.1 and M.2. It shows that the score-specific input A_s of Proposition M.5 holds on the whole one-dimensional class and on all product densities for $0 < s < 1$, that the multiplier condition fails on some densities of the class for every $0 < s \le 1/2$ although A_s is finite for them, and it separates three necessity questions. In particular it proves the strict separation stated in Remark M.6 and shows that A_s is necessary for the per-level risk bound. The results of the main text are unchanged. We use the notation of Appendices J.1 and M.1; P_t denotes the heat semigroup on $\mathbb{T}$ or $\mathbb{T}^d$ (Fourier multiplier $e^{-2\pi^2 t|k|^2}$), and for $0 < s < 1$ we use the equivalent norm of Lemma J.1(c).

N.1 The score input on one-dimensional and product classes

Proposition N.1 (One-dimensional score bound)**.** *Let $0 < s < 1$, $0 < \rho \le \bar{\rho}$, and let $q_0 \in H^s(\mathbb{T})$ satisfy $\rho \le q_0 \le \bar{\rho}$. Put $q_t = P_t q_0$ and $b_t = q_t'/q_t$. There is C_s, depending only on s, such that*

$$\sup_{t>0} \sqrt{t}\, \|b_t\|_{H^s(\mathbb{T})} \le C_s\big(\rho^{-1} + \bar{\rho}\rho^{-2}\big)\|q_0\|_{H^s(\mathbb{T})}.$$

Proof. Heat convolution preserves the bounds, so $\rho \le q_t \le \bar{\rho}$. For $h \in \mathbb{T}$ write $D_h u = u(\cdot + h) - u$. Then

$$D_h b_t = \frac{D_h q_t'}{q_t(\cdot + h)} - \frac{q_t'\, D_h q_t}{q_t(\cdot + h)\, q_t}, \qquad \text{so} \qquad |D_h b_t|^2 \le 2\rho^{-2}|D_h q_t'|^2 + 2\rho^{-4}\|q_t'\|_\infty^2 |D_h q_t|^2.$$

Integrating against $|h|_{\mathbb{T}}^{-1-2s}$ and using $\|b_t\|_{L^2} \le \rho^{-1}\|q_t'\|_{L^2}$, Lemma J.1(c) gives $\|b_t\|_{H^s} \le C_s(\rho^{-1}\|q_t'\|_{H^s} + \rho^{-2}\|q_t'\|_\infty \|q_t\|_{H^s})$. Finally $\|q_t\|_{H^s} \le \|q_0\|_{H^s}$ and $\|q_t'\|_{H^s} \le$

$\sup_k 2\pi|k|e^{-2\pi^2 tk^2}\,\|q_0\|_{H^s} \le t^{-1/2}\|q_0\|_{H^s}$, while $\|q_t'\|_\infty \le \|q_0\|_\infty \|\partial K_t^{\mathbb{T}}\|_{L^1(\mathbb{T})} \le \bar\rho(2/(\pi t))^{1/2}$ as in Lemma J.3. □

Corollary N.2 (Whole one-dimensional class and product classes)**.** *(i) Let $0 < s < 1$. For every d, the input A_s of (25) is bounded by a constant A, depending only on $(s, d, R, \underline{p}, \overline{p})$ and the wavelet system, on the product densities $p_0 = \prod_{r=1}^d q_r(x_r)$ of $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$. Hence, by Proposition M.5(ii), both bounds of Theorem 3 hold uniformly on this product subclass, without the multiplier condition.*

(ii) For $d = 1$ every density of the class $\mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$ is a product. Hence the bounds of Theorem 3 hold uniformly on it for $0 < s < 1$ by (i), and for $s > 1/2$ by Proposition J.1; so Corollary 1 holds for $d = 1$ and every $s > 0$.

Proof. Since $p_0 \ge \underline{p} > 0$, each factor has an a.e. constant sign, so we may take $q_r > 0$ with $\prod_r \int q_r = 1$ and normalize the factors to be densities; then each q_r is the r-th marginal of p_0, so $\underline{p} \le q_r \le \overline{p}$. Tensor wavelet coefficients factorize and every factor has norm at least its mean 1, so $\|q_r\|_{B^s_{2,2}(\mathbb{T})} \le \|p_0\|_{B^{s,\mathrm{mix}}_{2,2}} \le R$. The heat flow preserves products, and $\boldsymbol{s}_{t,r} = b_{r,t}(x_r)$ with $b_{r,t} = q_{r,t}'/q_{r,t}$ depends on x_r only; its mixed norm is the one-dimensional norm of $b_{r,t}$, because the constant factor 1 has norm 1. Proposition M.5(i), norm equivalence and Proposition N.1 give $A_s(p_0)^2 \le \overline{p}\, c_{\mathrm{eq}}^2 \sum_r \sup_t t\|b_{r,t}\|_{H^s}^2 \le C\, d\, c_{\mathrm{eq}}^4 R^2$ with $C = C_s^2(\underline{p}^{-1} + \overline{p}\,\underline{p}^{-2})^2\overline{p}$. □

N.2 For $0 < s \le 1/2$ the multiplier condition is strictly stronger than A_s

Proposition N.3 (Strictness for $0 < s < 1/2$)**.** *Let $0 < s < 1/2$, $d \ge 1$, $R > 1$ and $0 < \underline{p} < 1 < \overline{p}$. There is $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$, depending on x_1 only, such that $\sup_{0<t\le 1} \|g \mapsto g/p_t\|_{H^s_{\mathrm{mix}} \to H^s_{\mathrm{mix}}} = \infty$ while $A_s(p_0) < \infty$. In particular, for every constant M_s the multiplier condition of Appendix F fails for p_0 once $\underline{t}$ is small enough, the condition does not hold uniformly over the class, and Theorem 3 nevertheless applies to p_0 through Proposition M.5.*

Proof. It suffices to take $d = 1$: for a function of x_1 only and a test function $g(x_1) \otimes 1$, all norms involved reduce to one-dimensional ones.

Stripes. For an interval I of length $\ell \le 1/4$ and $N \ge 1$, split I into $2N$ cells of width $h = \ell/(2N)$ and let $f = \pm 1$ alternately on the cells and $f = 0$ off I. Then $\int f = 0$, $|f| \le 1$, and $\|D_z f\|_{L^2}^2 \le C\min(\ell, N|z|)$: for $|z| < h$ the difference is nonzero only within $|z|$ of the $2N + 1$ jumps, so $\|D_z f\|_{L^2}^2 \le 4(2N + 1)|z|$, and for $|z| \ge h$ we have $\|D_z f\|_{L^2}^2 \le 4\ell \le 8N|z|$. Integrating against $|z|^{-1-2s}$ and splitting at $|z| = \ell/N$ gives $\|f\|_{H^s} \le C_s \ell^{1/2-s} N^s$; only the piece $|z| < \ell/N$ needs $s < 1/2$. The same computation with $N = 1$ and without sign changes gives $\|\mathbf{1}_I\|_{H^s} \le C_s \ell^{1/2-s}$.

Lower bound. For $0 < a \le 1/2$, the function $v = \mathbf{1}_I/(1 + af)$ equals $1/(1 + a)$ and $1/(1 - a)$ on alternate cells, a difference $\delta = 2a/(1 - a^2) \ge 2a$. The N disjoint pairs of adjacent cells contribute to the double integral at least $N \cdot 2\delta^2 h^2 (2h)^{-1-2s} \ge c_s a^2 \ell^{1-2s} N^{2s}$, hence $\|v\|_{H^s} \ge c_s a\, \ell^{1/2-s} N^s$ and

$$\|\mathbf{1}_I/(1 + af)\|_{H^s} \ge c_s\, a\, N^s\, \|\mathbf{1}_I\|_{H^s}. \tag{26}$$

One density. Let $N_m = 2^{m^2}$ (any $N_m \to \infty$ would do) and $\ell_m = c\,(2^{-m} N_m^{-s})^{1/(1/2-s)}$ for $m \ge 2$, with $c > 0$ so small that $\sum_m \ell_m < 1/4$, so that $\ell_m^{1/2-s} N_m^s = c^{1/2-s} 2^{-m}$. Place the stripes f_m on disjoint intervals $I_m \subset (0, 1/4)$ and set $p_0 = 1 + a\sum_m f_m$. By disjointness $|p_0 - 1| \le a$ and $\int p_0 = 1$, and $\|p_0 - 1\|_{H^s} \le C_s a\, c^{1/2-s} \sum_m 2^{-m}$, the series converging in H^s to the pointwise sum. Choosing a small makes p_0 a member of the class, by norm equivalence.

Multiplier failure. Since $f_{m'} = 0$ on I_m for $m' \ne m$, $\mathbf{1}_{I_m}/p_0 = \mathbf{1}_{I_m}/(1 + af_m)$, and (26) shows that the multiplier norm of $1/p_0$ is at least $c_s a N_m^s \to \infty$. If $\|g \mapsto g/p_t\| \le K$ for all $0 < t \le 1$, then for $g = \mathbf{1}_{I_m}$ we have $\|g/p_t - g/p_0\|_{L^2} \le \underline{p}^{-2}\|p_t - p_0\|_{L^2} \to 0$, so the Fourier coefficients converge and Fatou's lemma for the weighted sum gives $\|g/p_0\|_{H^s} \le K\|g\|_{H^s}$ for every m, a contradiction. (The argument shows that the multiplier norm of $1/p_t$ tends to infinity as $t \downarrow 0$.)

Finite A_s. p_0 is a positive density in $H^s(\mathbb{T})$ between $1-a$ and $1+a$, so $A_s(p_0) < \infty$ by Corollary N.2. □

Proposition N.4 (The endpoint $s = 1/2$)**.** *The conclusion of Proposition N.3 also holds for $s = 1/2$.*

Proof. Again take $d = 1$. Fix an odd $\varphi \in C_c^\infty((-\frac{1}{2}, \frac{1}{2}))$ with $|\varphi| \le 1$, not identically zero. For $m \ge 2$ let $\xi_m = 2^{-m}$, $\eta_m = 2^{-m-4}$, $\varepsilon_m = \eta_m \exp(-2^{4m})$, $a_m = a2^{-m}$, $\varphi_m(x) = \varphi((x - \xi_m)/\varepsilon_m)$ and $p_0 = 1 + \sum_m a_m \varphi_m$.

Membership. The supports $|x - \xi_m| < \varepsilon_m/2$ are disjoint, so $|p_0 - 1| \le a/4$ and $\int p_0 = 1$. On $\mathbb{R}$ the homogeneous $H^{1/2}$ seminorm is dilation invariant, and $\|\varphi_m\|_{L^2} \le \varepsilon_m^{1/2}$. For a function supported in an arc of length less than $1/2$, the double integral on $\mathbb{T}$ is at most the double integral on $\mathbb{R}$ of its zero extension (take $y \in (x - \frac{1}{2}, x + \frac{1}{2}]$), so $\|\varphi_m\|_{H^{1/2}(\mathbb{T})} \le C_\varphi$, so the series converges in $H^{1/2}$ with norm at most $C_\varphi a$. For small a, p_0 lies in the class.

Trace bound. For U on the half-cylinder $\mathbb{T} \times [0, \infty)$ with compact support and finite Dirichlet energy, and $g = U(\cdot, 0)$, write $\widehat{U}_k(y)$ for the Fourier coefficients in x. For $k \ne 0$, $|\widehat{g}_k|^2 = -\int_0^\infty (|\widehat{U}_k|^2)' dy \le \int_0^\infty \big((2\pi|k|)^{-1}|\widehat{U}_k'|^2 + 2\pi|k||\widehat{U}_k|^2\big) dy$, so by Parseval $\sum_{k\ne 0} 2\pi|k||\widehat{g}_k|^2 \le \int |\nabla U|^2$ and

$$\|g\|^2_{H^{1/2}(\mathbb{T})} \le C\Big(\|g\|^2_{L^2} + \int_{\mathbb{T}\times(0,\infty)} |\nabla U|^2\Big).$$

Test functions. Let $\varrho = ((x-\xi_m)^2 + y^2)^{1/2}$ and $U_m = 1$ for $\varrho \le \varepsilon_m$, $U_m = \log(\eta_m/\varrho)/\log(\eta_m/\varepsilon_m)$ for $\varepsilon_m < \varrho < \eta_m$, and $U_m = 0$ otherwise; it is supported in one period. Its Dirichlet energy on the half-cylinder is $\pi/\log(\eta_m/\varepsilon_m) = \pi 2^{-4m}$, and $g_m = U_m(\cdot, 0)$ satisfies $\|g_m\|^2_{L^2} \le 2\varepsilon_m + 4\eta_m 2^{-8m}$. Hence $\|g_m\|_{H^{1/2}} \le C2^{-2m}$.

Lower bound. On the core $|x - \xi_m| < \varepsilon_m/2$ we have $g_m = 1$ and $g_m/p_0 = 1/(1 + a_m\varphi_m)$; off the core but inside $|x - \xi_m| < \eta_m$, $p_0 = 1$. Restricting the double integral of Lemma J.1(c) (kernel $|x - y|^{-2}$) to the core and rescaling by ε_m, which leaves it invariant,

$$\|g_m/p_0\|^2_{H^{1/2}} \ge c \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} \frac{|w(u) - w(v)|^2}{|u - v|^2}\, du\, dv, \qquad w = \frac{1}{1 + a_m\varphi},$$

and $|w(u) - w(v)| \ge a_m|\varphi(u) - \varphi(v)|/(1 + a_m)^2$, so the right-hand side is at least $c_\varphi a_m^2$. The multiplier ratio is therefore at least $c\, a_m 2^{2m} = c\, a\, 2^m \to \infty$. The Fatou argument of Proposition N.3 excludes a bound uniform in t, and $A_{1/2}(p_0) < \infty$ by Corollary N.2. □

N.3 Three necessity questions

Proposition N.5 (Per-level necessity and a direct-density benchmark)**.** *Let $s > 0$.*

(i) Necessary for the per-level risk bound. *Let $p_0 \in \mathcal{B}^{s,\mathrm{mix}}(R; \underline{p}, \overline{p})$. If the bound* (15) *holds for p_0 with a constant C for all $t > 0$, $L \ge 0$ and $n \ge m_L$, then $A_s(p_0)^2 \le C$.*

(ii) A minimax benchmark without an additional score assumption. *For every $s > 0$, the hyperbolic-cross projection estimator $\widehat{p} = \sum_{\lambda\in\Lambda_{L^\star}} \widehat{\theta}_\lambda \Psi_\lambda$, $\widehat{\theta}_\lambda = n^{-1}\sum_i \Psi_\lambda(X_i)$ (with Λ_L as in* (13)*; the constant term has $\widehat{\theta} = 1$), renormalized as $\widetilde{p} = \widehat{p}_+/\int \widehat{p}_+$, satisfies $\sup_{p_0} \mathbb{E}\,\mathrm{TV}(\widetilde{p}, p_0) \lesssim n^{-s/(2s+1)}(\log n)^{s(d-1)/(2s+1)}$ over the class. With Theorem 1, the TV minimax rate of the class is $n^{-s/(2s+1)}(\log n)^{s(d-1)/(2s+1)}$ for every $s > 0$, with no score input.*

Proof. (i) The least-squares score lies in $(\mathcal{V}_L^{\mathrm{SG}})^d$, so $\mathcal{E}(t) \ge B_L(t) := \sum_r \inf_{v\in\mathcal{V}_L^{\mathrm{SG}}} \|\boldsymbol{s}_{t,r} - v\|^2_{L^2(p_t)}$ deterministically, and $\mathbb{E}[\mathcal{E}(t)\mathbf{1}_{\mathcal{G}_t^{(1/4)}}] \ge B_L(t)\mathbb{P}(\mathcal{G}_t^{(1/4)})$. By Lemma J.4, $\mathbb{P}(\mathcal{G}_t^{(1/4)}) \to 1$ as $n \to \infty$ with t, L fixed; letting $n \to \infty$ in (15) gives $B_L(t) \le Ct^{-1}2^{-2sL}$ for all t, L, which is $A_s(p_0)^2 \le C$. (ii) For the at most m_L nonconstant terms, $\mathbb{E}(\widehat{\theta}_\lambda - \theta_\lambda)^2 \le n^{-1}\int \Psi_\lambda^2 p_0 \le \overline{p}/n$, and by (23) the discarded coefficients contribute at most $2^{-2sL}R^2$; hence $\mathbb{E}\|\widehat{p} - p_0\|^2_{L^2} \le R^2 2^{-2sL} + \overline{p}\, m_L/n$, which at $L = L^\star$ is $\lesssim n^{-2s/(2s+1)}(\log n)^{2s(d-1)/(2s+1)}$ (Section 5.3). Since $\int \widehat{p} = 1$, $Z := \int \widehat{p}_+ \ge 1 > 0$,

so $\widetilde{p}$ is well defined. Since $p_0 \geq 0$, $\|\widehat{p}_+ - p_0\|_{L^1} \leq \|\widehat{p} - p_0\|_{L^1}$ and $|\int \widehat{p}_+ - 1| \leq \|\widehat{p}_+ - p_0\|_{L^1}$, so $\|\widetilde{p} - p_0\|_{L^1} \leq 2\|\widehat{p} - p_0\|_{L^1} \leq 2\|\widehat{p} - p_0\|_{L^2}$ on the unit torus. Take expectations and use Jensen. □

The direct-density result establishes the class-wide rate without imposing an additional score condition. It does not construct a density with $A_s = \infty$ and therefore does not by itself give a pointwise counterexample to necessity of A_s.

Remark N.6 (What remains open, and where the attempts stop). (1) *Necessity for the diffusion estimator.* Whether $A_s < \infty$ is necessary for the TV rate of the diffusion sampler is not settled: Girsanov's theorem and Pinsker's inequality give only an upper bound on TV by the integrated score error, and a small endpoint TV does not force a small score error, so the bound cannot be reversed. (2) *Joint classes,* $d \geq 2$, $0 < s \leq 1/2$. Whether A_s is bounded on the whole ball is open. For $d = 2$ and $b = \partial_1 q/q$, $q = p_t$, the mixed difference is $D^{(1)}D^{(2)}b = (D^{(1)}D^{(2)}\partial_1 q)\,\tau w + (D^{(2)}\partial_1 q)\,\tau D^{(1)}w + (D^{(1)}\partial_1 q)\,\tau D^{(2)}w + \partial_1 q\, D^{(1)}D^{(2)}w$ with $w = 1/q$ and shifts τ. The first term and the second-order part of the last are bounded by t^{-1} times the mixed norm, as in Proposition N.1; the remaining terms require a bilinear square-function bound of the form $\int_{\mathbb{T}^2} S_1(\partial_1 p_t)^2\, S_2(p_t)^2\, dx \lesssim t^{-1}$, where S_i is the one-directional square function of the Gagliardo seminorm, and its analogues. Bounding one factor in L^∞ (using $\|S_2 p_t\|_\infty^2 \lesssim t^{-s}$) gives $\|\boldsymbol{s}_{t,1}\|^2_{B^{s,\mathrm{mix}}_{2,2}} \lesssim t^{-1-s}$, and bounding both factors through partial L^∞ norms (a Sobolev embedding in the other variable) heuristically gives a loss of order $t^{-(1-2s)}$ instead; either loss would change the rate after integration over t. We could not prove the bilinear bound, nor construct a density of the ball with $A_s = \infty$; any such counterexample must be non-product by Corollary N.2, and the densities of Propositions N.3 and N.4 are not counterexamples to A_s.

O THE FINITE-RESOLUTION DIAGNOSTIC ON OBSERVED WIND DIRECTIONS

This appendix applies the diagnostic of Section M.3 to one real angular system, with the protocol, stations and cutoffs fixed before the data were downloaded. It is descriptive; it neither certifies nor refutes membership in a Besov class.

Data and estimand. We use hourly wind directions for 2024 at four nearby stations in eastern Massachusetts (Boston Logan, Hanscom, Norwood, Beverly) from the NOAA Integrated Surface Database, ISD-Lite files (NOAA National Centers for Environmental Information, 2001). A timestamp is retained only if all four stations report a valid, non-calm direction; this leaves $n = 5131$ of 8784 hours, spread over all 53 weeks. Directions are mapped to $\mathbb{T}$ by dividing by $360°$, so $d = 4$; the first two stations give a $d = 2$ projection on the same hours. The estimand is the joint law of *reported* directions at retained hours. Every station reports directions in steps of $10°$, so the recorded law is discrete; we use Fourier modes $k \in \{-4, \ldots, 4\}^d$, far below the aliasing limit, as the finite-resolution analogue of the wavelet shells, grouped by the total level $\ell = \sum_{k_r \neq 0}(1 + \lfloor \log_2 |k_r| \rfloor) \leq 6$ and by the number of active coordinates. These are truncated frequency bins, not full dyadic shells: coordinate level 3 contains only $|k_r| = 4$, not $4, \ldots, 7$, and higher total levels inherit this truncation. Their values cannot be treated as a fitted asymptotic smoothness law.

Estimates and dependence. Bin energies are the diagonal-corrected sums $2\sum_k (|\sum_i e^{2\pi i \langle k, X_i \rangle}|^2 - n)/(n(n-1))$ over one mode from each $\pm k$ pair. The observations are strongly dependent (lag-one-hour autocorrelation of the direction cosines about 0.9, lag-one-day about 0.3) and seasonal, so the independent-sample variance bound of Section M.3 is not a confidence statement here. We report, as sensitivity summaries only, intervals from resampling whole 7-day blocks (199 replicates; 1- and 14-day blocks give the same picture), a cross-fit estimate from even versus odd weeks, and quarterly curves.

What is resolved. For the prespecified truncated bins, the main-effect estimates are 0.41, 0.17 and 0.07 at levels 1–3 for $d = 4$. Interaction estimates dominate several bins and fluctuate substantially across levels: for $d = 4$ the truncated-bin energies at levels $1, \ldots, 6$ are 0.41, 9.2, 3.3, 27.9, 25.0 and 31.3, and for $d = 2$ they are 0.23, 1.61, 0.38, 1.82, 0.15 and 0.30, with the mass on the modes $(a, \pm a)$ (Figure 2); the cross-fit estimates agree within 5% for the dominant interaction bins, while discrepancies in some smaller bins reach approximately 18%. Nearby stations report nearly equal

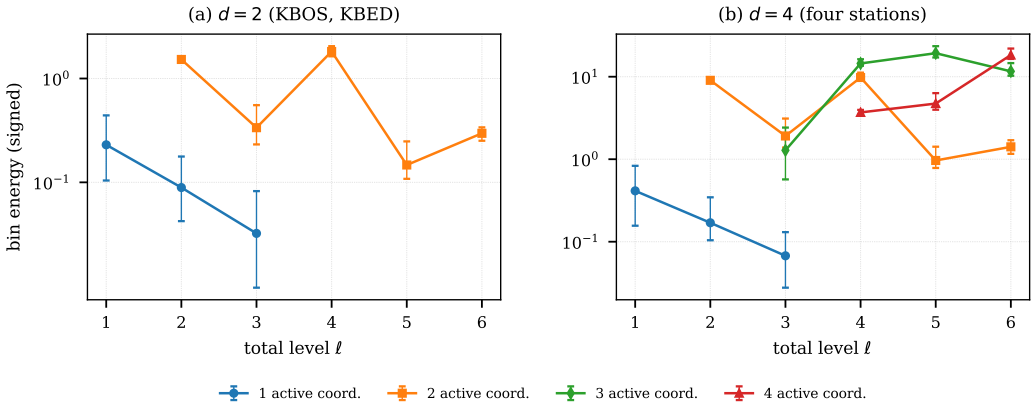

Figure 2: Signed energies in the prespecified truncated Fourier bins of hourly wind directions (NOAA ISD-Lite, 2024), by total level and number of active coordinates, with 7-day block-bootstrap sensitivity intervals. (a) Two stations; (b) four stations.

directions, so the joint law concentrates near the diagonal of $\mathbb{T}^d$. Interaction energy is larger in winter than in summer. The weighted cumulative sums of Section M.3 increase over the whole resolved range for $s = 0.25, 0.5$ and 1.

Interpretation. In the reported station coordinates the selected frequencies exhibit substantial cross-station dependence rather than only marginal variation. This illustrates why angular coordinates and small marginal energies alone do not justify the mixed model. Because the bins are truncated, the observations are quantized and dependent, and no fixed (s, R) envelope was tested, these plots establish neither membership nor exclusion of a particular mixed Besov ball. No sparse-grid fit or matched-budget comparison was run here, so they also establish no computational or statistical performance disadvantage. We did not analyze other coordinates, such as differences between stations, to avoid choosing the representation after seeing the results.