SPECTRAL BIAS AND FINITE SAMPLE EFFECTS IN DENOISING SCORE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based diffusion models achieve remarkable performance on high-dimensional tasks, yet the inductive biases that shape their learned scores remain poorly understood. Their training can be viewed as a sequence of denoising regression problems. In the infinite-width limit, a network is described by its NNGP kernel, whose spectrum induces a well-understood spectral bias in supervised regression. In this work, we show that this spectral bias is largely accurate in high-dimensional denoising score matching, and wide networks essentially act as supervised learning on the score. At small diffusion time, we derive two correction terms, one involving a heteroscedastic ridge and another involving a pointwise bias to the ensemble mean score. These corrections become significant when data have low intrinsic dimension but are embedded in a high-dimensional space with many weak-variance directions. We validate these predictions using finite-width networks on seven synthetic distributions, MNIST, and CIFAR-10, finding close agreement.

1 INTRODUCTION

Diffusion models learn scores through denoising regression (Vincent, 2011; Ho et al., 2020; Song et al., 2021). With finite training data, the learned score is shaped by both the sampled examples and the network's inductive bias. We ask how these factors determine which features of the population score are learned and how the prediction varies across inputs.

For supervised regression, kernel theory offers a useful answer. Infinite-width networks admit NNGP or NTK descriptions (Lee et al., 2018; Jacot et al., 2018), whose spectra determine which target components are learned: modes with large eigenvalues are retained, while those with small eigenvalues are suppressed. The threshold is set by an effective ridge combining explicit regularization with the predictor's remaining uncertainty (Bordelon et al., 2020; Canatar et al., 2021; Cohen et al., 2021). Denoising score matching (DSM), however, couples the input and target: the target is the noise used to perturb the input. Repeated noise draws constrain the network over a neighbourhood of each clean sample, making the loss sensitive to local derivatives. Whether the supervised spectral picture carries over is therefore not obvious.

We show that this picture largely survives, with additional local effects specific to denoising. For wide networks trained by weight-decay Langevin dynamics to equilibrium, we derive pointwise predictions at small diffusion times, near the end of the reverse process. In our global approximation, DSM smooths the population score with a similar operator as supervised regression but changes the overall scale and adds a pointwise bias driven by the spatial variation.

The data's local geometry is important. When uncertainty varies substantially across the data, our refined approximation replaces the constant effective ridge with a heteroscedastic one. This is relevant when variance concentrates in a few dominant directions with a long tail of weak ones, as can occur when the data has low intrinsic dimension relative to ambient dimension.

1.1 MAIN CONTRIBUTIONS

- **Pointwise score predictions for DSM.** We derive the score predictor ensemble mean and variance to leading order in small diffusion time (Eqs. (7) and (8)), connecting supervised spectral bias to denoising and identifying an additional uncertainty-dependent score bias.

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.

- **Pointwise score bias from data geometry.** The ensemble-mean score is not simply the population score smoothed by the covariance kernel. At leading order, the kernel acts on the population score together with a geometry-induced correction $D(x)$. Here $D(x) = \frac{1}{2}\nabla \log(\kappa^2 + R_x)$ is a pointwise bias generated by the spatial variation of the local posterior variance R_x. In the same regime, the covariance itself is geometry-dependent: the smoothing operator in Eq. (5) is no longer diagonal in the kernel eigenbasis, so it mixes spectral modes. Both effects become important for geometrically concentrated data with weak-variance directions, where R_x changes across the data manifold ; when R_x is constant, the bias vanishes, and the usual diagonal spectral filter is recovered.
- **Numerical validation with finite-width networks.** Ensembles of finite-width networks trained by Langevin dynamics on seven synthetic distributions in two families and on MNIST and CIFAR-10 agree closely with the predictions.

2 Related work

Score estimation and kernel regression. Cui & Zdeborová (2024) derive denoising errors for fixed-width nonlinear autoencoders, while George et al. (2025); George & Macris (2026) obtain learning curves for random-feature score models with Gaussian and manifold-supported data. We characterize the mean predictor and its variance at each input, rather than only aggregate error. Latourelle-Vigeant et al. (2026) analyze DSM with inner-product kernels, deriving risk trajectories and the generated law through high-dimensional kernel linearization. Wang & Pehlevan (2025) study spectral learning dynamics for linear and convolutional denoisers, averaging over both data and noise; we retain a finite random clean dataset. The resulting predictor also connects to the score-smoothing interpretation of Gabriel et al. (2025), who analyze kernel-smoothed empirical scores and conjecture an equivalent-kernel connection. Our theory retains nonlinear score structure and derives how local uncertainty shapes both the smoothing operator, determined self-consistently by the ensemble covariance h, and an additional score correction D.

Biases in diffusion models. For shallow autoencoders trained on Gaussian-mixture data, Cui et al. (2024) quantify how limited training data shift the generated mixture's mean away from its target. Cui et al. (2025) study how learning shapes the generated distribution and identify failures such as mode collapse. These works characterize distortions in the generated data. Training duration provides another source of bias: Bonnaire et al. (2026) show that models can generalize before longer training leads to memorization. Another proposed mechanism is randomness in the training target. Vastola (2025) relate score-estimator variability to reverse dynamics for a fixed empirical training distribution, while Bertrand et al. (2025) find that replacing stochastic flow-matching targets with closed-form targets preserves sample quality in experiments. Our focus is how a finite training set systematically distorts the learned score relative to the population score. Averaging over independently drawn clean datasets, we predict this distortion at each input. It persists at equilibrium with corruption noise averaged out of the loss, so it does not rely on early stopping or stochastic target sampling.

Score smoothing. Farghly et al. (2025) show how log-domain smoothing preserves manifold structure, with a score/log-density correspondence requiring location-independent kernels. Chen (2026) connect smoothing to neural-network regularization and analyze interpolation for uniformly spaced data in a one-dimensional subspace. Our analysis jointly determines the smoothing operator and ensemble variance from DSM training. In the geometry-aware specialization, $R_x = h(x, x)$ sets a local ridge $\kappa^2 + R_x$ that can mix kernel eigenmodes, and generates $D = \frac{1}{2}\nabla \log(\kappa^2 + R_x)$. Thus finite-sample uncertainty changes both the operator and the score being smoothed, which corresponds to the reweighted density $p_{\text{eff}} \propto p\sqrt{\kappa^2 + R_x}$.

3 Setup

Data and forward process. Let p be a smooth, strictly positive density on $\mathbb{R}^d$ and let $\{x_0^{(\mu)}\}_{\mu=1}^P$ be P clean samples drawn i.i.d. from p. We write $\mathrm{d}\mu(x) = p(x)\,\mathrm{d}x$ and $s(x) = \nabla \log p(x)$ for the score. At noise level t the forward process maps a clean point to

$$x_t \;=\; x_0 + \sqrt{\beta_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_d), \tag{1}$$

with variance schedule β_t.[1] We fix one level t and write $\beta \equiv \beta_t$; everything below holds at each t separately. We work in the small-noise regime: β is small compared with the squared distances between data points and with the length scale of the kernel K defined below.

Each clean sample is presented with M independent noise draws $\epsilon^{(\mu,m)} \sim \mathcal{N}(0, I_d)$, $m = 1, \dots, M$, giving $N = MP$ noisy inputs $x_\beta^{(\mu,m)} = x_0^{(\mu)} + \sqrt{\beta}\,\epsilon^{(\mu,m)}$. We train a network $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$ on the denoising loss

$$\mathcal{L}_M(\boldsymbol{\theta}) \;=\; \frac{1}{2\kappa^2} \sum_{\mu=1}^{P} \frac{1}{M} \sum_{m=1}^{M} \big\| \epsilon^{(\mu,m)} - f_{\boldsymbol{\theta}}\big(x_\beta^{(\mu,m)}\big) \big\|^2, \tag{2}$$

where $\kappa^2 > 0$ is a hyperparameter whose role is fixed below. As $M \to \infty$, $\mathcal{L}_M \to \mathcal{L}(\boldsymbol{\theta}) \equiv \frac{1}{2\kappa^2} \sum_{\mu=1}^{P} \mathbb{E}_\epsilon \|\epsilon - f_{\boldsymbol{\theta}}(x_0^{(\mu)} + \sqrt{\beta}\,\epsilon)\|^2$. Following analytic treatments of diffusion training (Vincent, 2011; Wang & Pehlevan, 2025), the derivation works with this limit at fixed κ^2, so that the clean dataset $\{x_0^{(\mu)}\}$ is the only quenched disorder (at finite M the noise draws would have to be replicated too).

Score model. The noise-averaged loss matches the network output to the noise ϵ (Ho et al., 2020; Nichol & Dhariwal, 2021; Kingma & Gao, 2023), which is the default convention in practice (Rombach et al., 2022; von Platen et al., 2022). The score, is recovered as $s_{\boldsymbol{\theta}}(x) = -f_{\boldsymbol{\theta}}(x)/\sqrt{\beta}$.

We work in the kernel regime. The prior over functions induced by a wide network $f_{\boldsymbol{\theta}}$ is a Gaussian process with the NNGP kernel (Lee et al., 2018), and weight-decay Langevin dynamics run to equilibrium samples the corresponding Bayesian posterior (Naveh et al., 2021; Li & Sompolinsky, 2021; Seroussi et al., 2023).[2] Since $f_{\boldsymbol{\theta}}$ is vector-valued and the readout weights of different output indices are independent, the prior covariance is diagonal in the output index, $\langle f_i(x) f_j(x') \rangle = K(x, x')\,\delta_{ij}$. We assume K is trace class on $L^2(\mu)$, and write its eigendecomposition in $L^2(\mu)$ as $K(x, y) = \sum_\rho \lambda_\rho \phi_\rho(x) \phi_\rho(y)$ with $\int \mathrm{d}\mu\, \phi_\rho \phi_\sigma = \delta_{\rho\sigma}$.

Under the loss (2) the posterior is again Gaussian, with mean

$$\bar{f}_i(x) \;=\; \sum_{a,b=1}^{N} K\big(x, x_\beta^{(a)}\big) \Big[\big(\mathbf{K} + M\kappa^2\, I_N\big)^{-1} \Big]_{ab} \epsilon_i^{(b)}, \qquad \mathbf{K}_{ab} = K\big(x_\beta^{(a)}, x_\beta^{(b)}\big), \tag{3}$$

where $a = (\mu, m)$ runs over the $N = MP$ noisy inputs: kernel ridge regression from the noisy inputs to the noise, with κ^2 the observation-noise variance per clean sample in the Bayesian reading and the ridge strength in the regression reading. The object of study is this predictor at an input y, averaged over draws of the clean dataset $\{x_0^{(\mu)}\}_{\mu=1}^{P}$; the average turns the $N \times N$ inverse in Eq. (3) into a closed-form expression in β, κ^2, and the statistics of p.

The supervised baseline. For ordinary supervised regression with a fixed target $g : \mathbb{R}^d \to \mathbb{R}^d$ on the same P inputs, kernel, and ridge (Canatar et al., 2021; Cohen et al., 2021), the same machinery gives an ensemble of trained predictors that is Gaussian to leading order, with covariance $\delta_{ij} h_0(x, y)$ and mean $\bar{g}(y) = \frac{P}{\kappa^2 + R} \int \mathrm{d}\mu(x)\, h_0(y, x)\, g(x)$, where

$$h_0(x, y) = \sum_\rho \frac{\lambda_\rho}{1 + \lambda_\rho P/(\kappa^2 + R)}\, \phi_\rho(x) \phi_\rho(y), \qquad R = \int \mathrm{d}\mu(x)\, h_0(x, x). \tag{4}$$

Thus $\bar{g}_\rho = \lambda_\rho\, g_\rho / \big(\lambda_\rho + (\kappa^2 + R)/P\big)$: modes with $\lambda_\rho \gg (\kappa^2 + R)/P$ are learned, modes with $\lambda_\rho \ll (\kappa^2 + R)/P$ are suppressed, and the threshold is set by the *effective ridge* $\kappa_{\text{eff}}^2 = \kappa^2 + R$, fixed self-consistently by the total predictor variance R.

[1] This is the variance-exploding form (Song et al., 2021). The variance-preserving form of DDPM (Ho et al., 2020), $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, equals $\sqrt{\bar{\alpha}_t}$ times Eq. (1) with $\beta_t = (1 - \bar{\alpha}_t)/\bar{\alpha}_t$; the constant rescaling of the network input can be absorbed into the kernel, so the analysis does not depend on this choice.

[2] For networks trained by gradient descent to convergence, the same formulas hold with the NTK in place of the NNGP kernel (Jacot et al., 2018; Lee et al., 2019).

4 MAIN RESULTS

This section presents the derived formulas for the ensembled mean and variance of the trained denoiser at an input y, averaged over the dataset, to leading order in $\sqrt{\beta}$. Both are functionals of the ensemble covariance h, Eq. (5), a supervised propagator whose effective ridge varies along the dominant directions of the data. The mean, Eq. (7), is the population score smoothed by h plus a bias driven by the spatial variation of the ensemble variance; the variance, Eq. (8), is $h(y, y)$ itself. Full derivations are in Appendix A.

4.1 HETEROSCEDASTIC RIDGE, MEAN, AND VARIANCE

The self-consistent equations. Let $E \subset \mathbb{R}^d$ be the subspace of the few dominant directions of the data and $E^\perp$ the directions along which the population has small spread $\sigma_\perp$. Write $x = (s, u)$ with $s = \Pi_E x \in E$ and $u = (1 - \Pi_E)x \in E^\perp$. To leading order, the trained ensemble is a Gaussian field, diagonal in the output index, $\langle f_i(x) f_j(y) \rangle = \delta_{ij}\, h(x, y) + O(\beta)$, whose covariance solves

$$h \;=\; \Big(K^{-1} + \frac{P}{\kappa^2}\,\hat{\gamma}\Big)^{-1}, \qquad \gamma_s \;=\; \frac{\kappa^2}{\kappa^2 + R_s} \tag{5}$$

where K^{-1} is the inverse of the kernel as an operator on $L^2(\mu)$ and $\hat{\gamma}$ is the multiplication operator $(\hat{\gamma}\varphi)(x) = \gamma_{s(x)}\varphi(x)$. With $R_s = \mathbb{E}[\, R_x \mid \Pi_E x = s \,]$ with the self-consistent pointwise variance $R_x = h(x, x)$, so the ridge factor $\gamma_{s(x)}$ depends on x only through its dominant coordinates. In the kernel eigenbasis,

$$\big[h^{-1}\big]_{\rho\sigma} \;=\; \frac{\delta_{\rho\sigma}}{\lambda_\rho} \;+\; \frac{P}{\kappa^2}\int \mathrm{d}\mu(x)\; \gamma_{s(x)}\, \phi_\rho(x)\phi_\sigma(x), \tag{6}$$

so the local factor $\gamma_{s(x)}$ mixes the eigenfunctions: h is diagonal in the modes only if γ_s is constant over the population. We write $R = \int \mathrm{d}\mu(x)\, R_x = \mathrm{Tr}\, h$ for the total variance.

The position dependence comes from applying the Gaussian-equivalence step of the supervised theory conditionally on the dominant directions of the data rather than to the population as a whole (Sec. 5); In particular, this local description matters: when R_x derivatives are large, i.e., the pointwise posterior variance across the data manifold. This variation may stem from data concentration near a low-dimensional subspace, which in turn increases score bias.

Mean predictor and Score bias (*uncertainty-reweighted* density). With h from Eq. (5),

$$\begin{aligned}
\langle f_i(y) \rangle \;&=\; -\frac{P\sqrt{\beta}}{\kappa^2}\int \mathrm{d}\mu(x)\; h(y, x)\Big[\, s_i(x) \;+\; \frac{\gamma_x}{2\kappa^2}\,\partial_i R_x \Big] \;+\; O(\beta^{3/2}) \\
&=\; -\frac{P\sqrt{\beta}}{\kappa^2}\int \mathrm{d}\mu(x)\; h(y, x)\Big[\, \underbrace{s_i(x)}_{\text{score}} \;+\; \underbrace{\tfrac{1}{2}\,\partial_i \log\big(\kappa^2 + R_x\big)}_{D_i(x)} \Big] \;+\; O(\beta^{3/2}),
\end{aligned} \tag{7}$$

$$\mathrm{Var}[f_i(y)] \;=\; h(y, y) \;=\; R_y \;+\; O(\beta), \tag{8}$$

the two lines of Eq. (7) being equal by the definition of γ_x. $s = -f\sqrt{\beta}$, the ensemble-mean score is,

$$\langle s_i(y) \rangle \;=\; \frac{P}{\kappa^2}\int \mathrm{d}\mu(x)\, h(y, x)\, \big[s_i(x) + D_i(x)\big] \;+\; O(\beta), \tag{9}$$

and the bracket is a single gradient,

$$s_i(x) + D_i(x) \;=\; \partial_i \log p_{\text{eff}}(x), \qquad p_{\text{eff}}(x) \;:=\; \frac{p(x)\,\sqrt{\kappa^2 + R_x}}{Z}, \quad Z = \int \mathrm{d}\mu(x)\, \sqrt{\kappa^2 + R_x}. \tag{10}$$

The population score thus enters smoothed by h: the spectral filter of supervised regression, with its threshold set locally by $\kappa^2 + R_x$. The source, however, is $s + D$ rather than s. The score bias D is generated purely by the gradient of the pointwise posterior variance, has no analog in supervised regression on a fixed target, and vanishes wherever R_x is flat; by Eq. (10), the ensemble learns the smoothed score of the *uncertainty-reweighted* density p_{eff}. The variance is independent of β and of the data score at this order.

4.2 CONSTANT RIDGE AND THE SUPERVISED SPECTRAL BIAS

Replacing γ_x in Eq. (5) by its population average $\gamma = \kappa^2/(\kappa^2 + R)$ collapses the integral in Eq. (6) to $\delta_{\rho\sigma}$ by orthonormality and, since $P\gamma/\kappa^2 = P/(\kappa^2 + R)$, returns the supervised propagator $h = h_0$ of Eq. (4): diagonal in the modes, with the constant effective ridge $\kappa^2_{\text{eff}} = \kappa^2 + R$. This is what the Gaussian-equivalence step gives when applied to the population as a whole, as in the supervised theory (Sec. 5); it is exact when $R_x \equiv R$ on the support of p and an approximation otherwise. Equations (7), (10) hold with $h = h_0$ and $R_x = h_0(x, x)$, and Eq. (8) becomes precisely the self-consistent effective-ridge variance of Canatar et al. (2021). We refer to this reduction as the *homoscedastic ridge* and to Eq. (5) as the *heteroscedastic ridge*; Sec. 6 compares the two.

5 DERIVATION SKETCH

In this section, we describe the main steps of the derivation of the mean ensemble score Eq. (7) and the ensemble variance Eq. (8). We draw on tools from Gaussian process theory, field theory, and statistical physics methods. See Ringel et al. (2025) for the review of the approach in the context of deep neural networks. As in the supervised calculation, we replicate the partition function. Two steps are new: since the target is coupled to the input through the noise, we expand the action in the noise standard deviation; and we apply Gaussian equivalence locally, yielding a heteroscedastic ridge.

Partition function and replica action. Weight-decay Langevin training equilibrates to $P_{\text{eq}}(\boldsymbol{\theta}) \propto e^{-\|\boldsymbol{\theta}\|^2/2-\mathcal{L}(\boldsymbol{\theta})}$, and marginalizing $\boldsymbol{\theta}$ against the infinite-width Gaussian prior gives the function-space partition function $Z = \int Df\, \exp[-\frac{1}{2}\langle f, K^{-1} f\rangle - \frac{1}{2\kappa^2}\sum_\mu \mathbb{E}_\epsilon \|f(x_0^{(\mu)} + \sqrt{\beta}\,\epsilon) - \epsilon\|^2]$, Averaging Z^n over the i.i.d. dataset factorizes over data points, and exponentiating the resulting factor (exact as $n \to 0$) gives a replica action with a prior term and a data term $P - P\int \mathrm{d}\mu(x)\, e^{-\frac{1}{2\kappa^2}\sum_a L_a(x)}$. Integrating the cross term of L_a by parts in ϵ turns the noise target into a divergence, $\mathbb{E}_\epsilon f\cdot\epsilon = \sqrt{\beta}\,\mathbb{E}_\epsilon \nabla\cdot f$, and expanding in $\sqrt{\beta}$ gives $L_a = |f^a|^2 + d - 2\sqrt{\beta}\,\nabla\cdot f^a + O(\beta)$. The action splits as $S = S_{\text{ref}} + S_{\text{v}}$: the $\beta = 0$ action S_{ref}, even under $f \mapsto -f$, and the odd vertex $S_{\text{v}} = -\frac{P\sqrt{\beta}}{\kappa^2}\int \mathrm{d}\mu(x)\, e^{-\tilde{U}(x)} \sum_a \nabla\cdot f^a(x)$ with $\tilde{U}(x) = \frac{1}{2\kappa^2}\sum_b |f^b(x)|^2$.

Reference measure: Gaussian equivalence on fibers. The data term of S_{ref} is non-Gaussian in f through $e^{-\tilde{U}}$. Following the supervised theory, we treat the field values $Y = f(x)$ at a random data point $x \sim p$ as Gaussian for a fixed field configuration, conditionally on the coordinates $s = \Pi_E x$ along a subspace E, i.e. separately on each fiber N_s, with fiber mean $\bar{Y}_s$ and covariance C_s. For $E = \{0\}$ there is a single fiber and the step is the global Gaussian equivalence of the supervised calculation, with a single overlap matrix Q_{ab}. A shifted-Gaussian identity (Lemma 1) evaluates $\mathbb{E}[e^{-|Y|^2/2\kappa^2}]$; the fibre mean drops out at quadratic order (Eq. (41)), and linearizing around a replica-symmetric mean field $Q_s^{\text{MF}} = R_s I_n + T_s \Gamma$ makes the data term quadratic with a position-dependent coefficient, $S_{\text{ref}} \approx \frac{1}{2}\sum_a \langle f^a, K^{-1} f^a\rangle + \frac{P}{2\kappa^2}\int \mathrm{d}\mu(x)\, \gamma_{s(x)} \sum_a |f^a(x)|^2$, whose propagator is Eq. (5) (Eq. (4) when $E = \{0\}$). Moment matching closes the self-consistency, $R_s = \int \mathrm{d}\mu_s\, h(x', x')$, and forces $T_s = 0$. We write $\langle\cdot\rangle_{\text{ref}}$ for the average under $e^{-S_{\text{ref}}}$.

Mean and variance. The exact identity $\langle F\rangle = \langle F e^{-S_{\text{v}}}\rangle_{\text{ref}} / \langle e^{-S_{\text{v}}}\rangle_{\text{ref}}$ expands as a linked-cluster series $\sum_k \frac{(-1)^k}{k!}\langle F S_{\text{v}}^k\rangle_{\text{ref},c}$, where the subscript c denotes the connected part, $\langle F S_{\text{v}}\rangle_{\text{ref},c} = \langle F S_{\text{v}}\rangle_{\text{ref}} - \langle F\rangle_{\text{ref}}\langle S_{\text{v}}\rangle_{\text{ref}}$. For the odd observable $F = f_i(y)$, parity kills the $k = 0$ and $k = 2$ terms, so $\langle f_i(y)\rangle = -\langle F S_{\text{v}}\rangle_{\text{ref}} + O(\beta^{3/2})$. The damping factor $e^{-\tilde{U}}$ inside S_{v} is handled exactly by a Gaussian damping lemma (Lemma 2): it contributes $\langle e^{-\tilde{U}}\rangle \to 1$ and replaces the covariances by their damped versions. The direct contraction of F with the derivative field gives $\partial_{2,i} h(y, x)$, which integrates by parts against p into $-h(y, x) s_i(x)$ with no damping factor; the contraction routed through the damped block gives $-h(y, x)\,\partial_i R_x/2(\kappa^2 + R_x)$. Their sum is Eq. (7) (Appendix A.9.1 recovers both terms by a leave-one-out argument). For the even observable $F = f_i(y) f_j(y)$, all odd-k terms vanish and the $k = 0$ term is $\delta_{ij} h(y, y)$, giving Eq. (8).

6 EMPIRICAL VALIDATION

This section tests the two predictions of Sec. 4, the ensemble mean of Eq. (7) and the ensemble variance of Eq. (8), both built on the heteroscedastic ridge of Eq. (5), against ensembles of finite-width denoising networks. We first describe the protocol: how the ensembles are generated, how κ^2 and β are set, and which theory curves are compared with them, including the constant ridge of Sec. 4.2 and a baseline without the score bias. We then present the experiments: seven synthetic distributions in an isotropic and an anisotropic embedding, where the population score is known, and the theory can be evaluated exactly (Fig. 4); MNIST and CIFAR-10 along a geodesic through the data (Fig. 2); and, last, the variance, which tests the reference theory on its own. Throughout, there is one network per draw of the training set, so that the ensemble average is over the dataset randomness the theory averages over, and the ensemble variance is the quantity R_x predicts.

Protocol. The network is a one-hidden-layer erf network in the NNGP parameterization, of width $N_{\mathrm{w}} = 5000$ on the synthetic distributions and $N_{\mathrm{w}} = 60\,000$ on image data (Appendix D). In every ensemble, each member has its own draw of P clean points, each presented with $M = 64$ noise draws, so that the member is trained on the loss of Eq. (2) exactly as written. The number of noise draws is checked against the $M \to \infty$ limit in which the theory is derived, for details see Appendix D. The synthetic ensembles have 256 members and the image ensembles 32. Circles in the figures are ensemble means with ± 1 s.e. error bars. We set κ^2 through the ratio $R/\kappa^2 : R/\kappa^2 = 1$ in the isotropic embedding and on image data, $R/\kappa^2 = 10$ in the anisotropic embedding, in both cases far from the equivalent-kernel limit $R/\kappa^2 \to 0$, so that the corrections have a chance to be visible; $\sqrt{\beta} = \kappa^2/100$, and κ^2 and $\sqrt{\beta}$ are printed in each panel (Figs. 4., 2, 1). Theory curves solve the self-consistency for h numerically on a quadrature sample of the population (Appendix B) and insert h into Eqs. (7), (8). Where the population score is known, we also plot the score term of Eq. (7) alone, evaluated with the constant ridge, $\frac{P}{\kappa^2} \int \mathrm{d}\mu(x)\, h_0(y, x)\, s(x)$: labeled as spectral bias in the plot.

Synthetic distributions. Each synthetic density is Gaussian along $d - 1$ coordinates, with the specified distribution along the coordinate x_1. The network output is projected onto e_1. In the *isotropic* embedding the remaining coordinates are standard Gaussian, so every direction carries comparable variance and R_x varies little over the data (Fig. 7). In the *anisotropic* embedding they have variance $\sigma_\perp^2 \ll 1$, so the data have intrinsic dimension one inside a d-dimensional space with many weak-variance directions, and R_x varies by $O(1)$ along x_1. The seven profiles are, in the isotropic embedding ($d = 64$, $P = 256$), a standard Gaussian, a Student-t with $\nu = 3$, the two-component mixture $\frac{1}{2}\mathcal{N}(\pm 2, 1)$ and a heavy-tailed scale mixture of three Gaussians of widths 1, 3, 10, probed along $y = y_1 e_1$; and, in the anisotropic embedding ($d = 64$, $P = 512$, $\sigma_\perp = 0.15$), a Gaussian, a Student-t with $\nu = 3$ and a three-component Gaussian mixture, each scaled to variance $\sigma_1^2 = 16$ along x_1 and probed along $y = y_1 e_1 + \sigma_\perp \sqrt{d-1}\, e_2$, the typical transverse radius of the data, for $|y_1| \le 1.5\sigma_1$.

Figure 1 tests Eq. (7). In the isotropic embedding (top row), the constant and the heteroscedastic ridge nearly coincide, the only visible difference being in the tails of the Student-t, and both track the ensemble mean: this is the regime in which DSM behaves as supervised regression on the score, and the score bias is a small correction, visible only as the gap between the dashed curve, which omits the score bias term, and the ensemble. In the anisotropic embedding (bottom row), the corrections are large. The score bias is of the same order as the score itself, so the score term alone (dashed) misses the ensemble throughout; the constant ridge misplaces the ensemble mean where R_x departs most from its population average, at the center of the Gaussian, where R_x dips, in the tails of the Student-t, and at the peaks of the three-component mixture, while the heteroscedastic ridge tracks the ensemble in every panel. The mixture is the least accurate panel; its residual is larger in the $d = 16$ experiment of Fig. 4, and we attribute it to the mean-field propagator, whose accuracy degrades as P approaches the number of learned modes, rather than to the score bias.

MNIST and CIFAR-10. Figure 2 applies the same predictor to image data, with the ensembles of the protocol: MNIST at 14×14 ($d = 196$) and grey-scale CIFAR-10 at 16×16 ($d = 256$), each dataset draw being $P = 256$ images of the training set (Appendix D). The probes are 17 points evenly spaced in arc length ℓ along a geodesic through the data manifold, a shortest path through a density-restricted k-nearest-neighbour graph in a whitened top-PCA subspace (Appendix E.3), and the d-dimensional mean and variance are projected onto the local tangent of the path. Image data

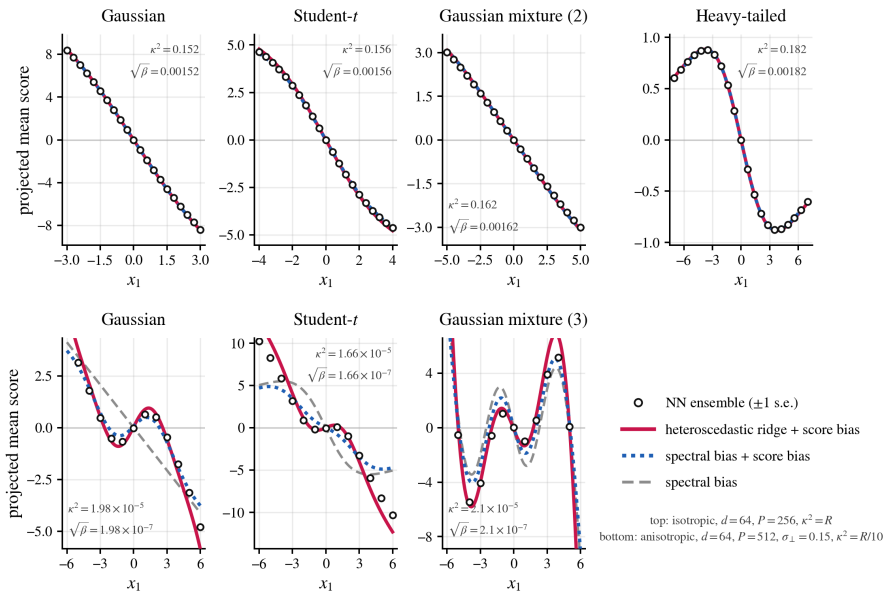

Figure 1: **Synthetic distributions: the theory predicts the ensemble mean.** Projected mean score of a denoising ensemble against the non-Gaussian coordinate x_1. *Top row (isotropic embedding, $d = 64$, $P = 256$, $\kappa^2 = R$):* Gaussian, Student-t, two-component Gaussian mixture, and heavy-tailed profiles, with the remaining coordinates standard Gaussian. *Bottom row (anisotropic embedding, $d = 64$, $P = 512$, $\sigma_\perp = 0.15$, $\kappa^2 = R/10$):* Gaussian, Student-t, and three-component Gaussian mixture profiles, with the remaining coordinates Gaussian of variance $\sigma_\perp^2$. *Circles*: ensemble mean over 4096 dataset realizations, one width-5000 network per realization with its readout posterior evaluated in closed form in the $M \to \infty$ limit (Appendix C), ± 1 s.e. *Solid*: Eq. (7) with the heteroscedastic ridge. *Dotted*: Eq. (7) with the constant and the score bias term. *Dashed*: the score term of Eq. (7) with the constant ridge, $\frac{P}{\kappa^2} \int \mathrm{d}\mu \, h_0(y, x) \, s(x)$, i.e. kernel ridge regression on the true score with the same kernel, ridge and prefactor; it differs from the dotted curve only by the score bias D. Variance panels are in (Fig. 3); the same experiment at $d = 16$ is in Appendix E Fig. 6.

have no canonical dominant subspace E, so we use the pointwise form of Eq. (5); the population score is not known, so no supervised baseline is available, and the score term of Eq. (7) is evaluated by integrating it by parts onto the kernel (Appendix B). The theory reproduces the overall shape and much of the fine non-monotonic structure of both the mean and the variance along the geodesic on both datasets. At this operating point ($R/\kappa^2 = 1$) the score bias is small on image data, so the image experiments test the smoothing operator and the variance rather than D; the position dependence of the ridge is what improves on the constant ridge, which lies farther from the ensemble mean on both datasets (not shown). The per-class breakdown (Appendix E.2) gives the same picture on each of the twenty class populations.

Validation of variance prediction. The variance panels of Fig. 3 and of Fig. 2 test Eq. (8). Being β-independent and insensitive to the score at this order, the variance is a test of the reference theory alone. The ensemble variance follows $R_x = h(x, x)$ of Eq. (5), including its variation along x_1 in the anisotropic embedding, where it dips at the center of the Gaussian and the Student-t and oscillates with the components of the mixture, and along ℓ on image data. On the synthetic distributions, the width-5000 ensembles lie a small, nearly uniform factor below the infinite-width prediction: the ratio of ensemble variance to R_x is constant along the probe, so the shape of R_x is reproduced, and only its level is shifted, an $O(P/N_\mathrm{w})$ finite-width correction that grows with P/N_w (Fig. 5). In the isotropic embedding the constant and the heteroscedastic ridge give nearly the same variance profile; in the anisotropic embedding the heteroscedastic ridge is the closer, the difference being concentrated where R_x departs most from its population average.

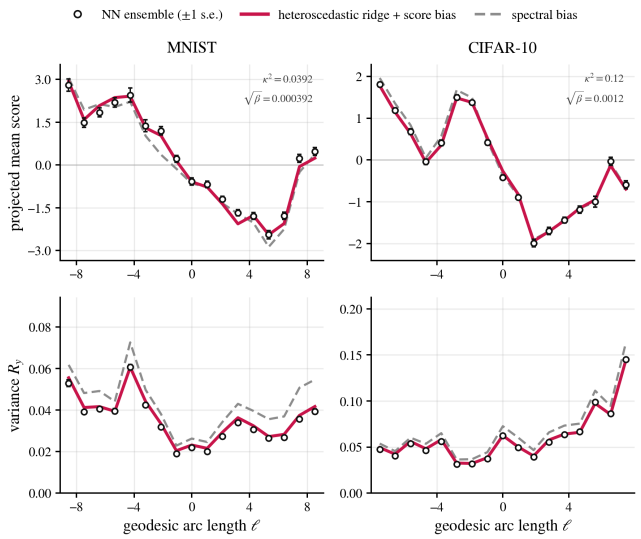

Figure 2: **Image data.** Mean (top) and variance (bottom) of the denoising ensemble, projected onto the local tangent of the path, against geodesic arc length ℓ for MNIST (left) and CIFAR-10 (right), pooled over all classes; κ^2 and $\sqrt{\beta}$ as printed. *Circles*: ensemble of 32 width-60 000 networks, one per draw of $P = 256$ training images, each with its readout posterior evaluated in closed form in the $M \to \infty$ limit (Appendix C); the top panels show the ensemble mean ± 1 s.e., the bottom panels the posterior variance of the network averaged over the ensemble. *Solid*: Eqs. (7) and (8) with the heteroscedastic ridge. The coordinate ℓ parameterizes a shortest path through neighbouring training images in a density-restricted graph, which migrates between classes without leaving the density-supported region; the probes are evenly spaced in arc length along it, each lying between two neighbouring training images (Appendix E.3, Fig. 10).

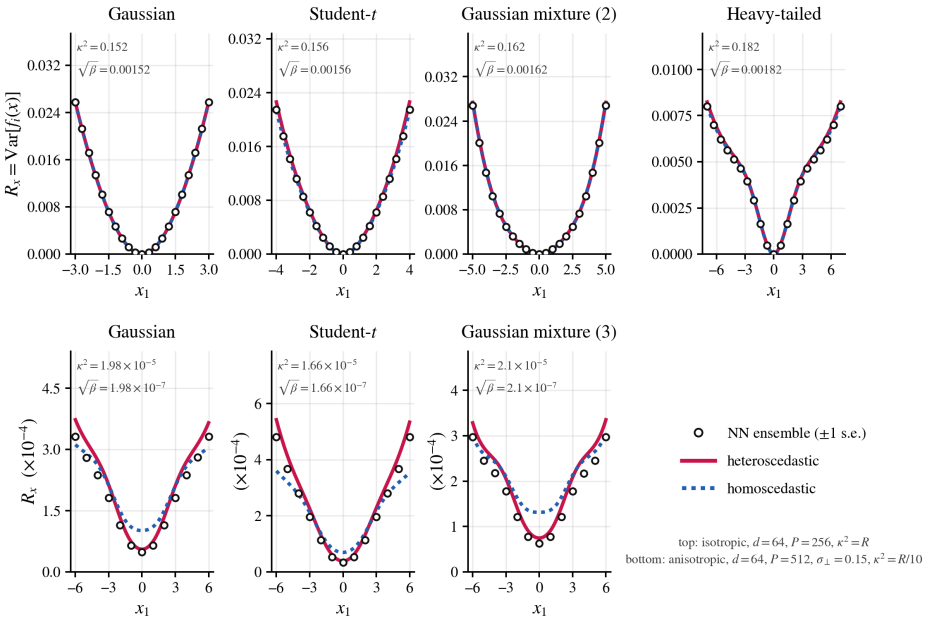

Figure 3: **Ensemble variance on the synthetic distributions.** Variance of the projected ensemble output along the probe line, $R_x = \mathrm{Var}[f_i(x)]$, for the seven distributions and settings of Fig. 4. *Circles*: posterior variance of the network averaged over the same 4096 dataset realizations as in Fig. 4, which to leading order equals the across-realization variance, ± 1 s.e.; *solid*: $R_x = h(x, x)$ with h from the heteroscedastic ridge of Eq. (5) in its pointwise form; *dotted*: $h_0(x, x)$ from the constant ridge of Eq. (4). In the isotropic embedding, the two coincide; in the anisotropic embedding, the heteroscedastic ridge reproduces the depth of the central dip of all three profiles, which the constant ridge underestimates.

7 DISCUSSION

Our theory links supervised spectral bias to the score learned by denoising. Local uncertainty affects how strongly the score is smoothed, while its variation across inputs produces an additional score bias. The global approximation retains the familiar spectral filter, and the geometry-aware refinement allows the effective regularization to vary across inputs. Experiments on synthetic data, MNIST, and CIFAR-10 support these predictions and motivate extensions beyond the present equilibrium, small-noise setting.

Limitations and Future directions. Our theory describes the learned score at equilibrium in the infinite-width limit. A natural next step is to study how the score and its bias evolve during training, clarifying which score components are learned first and how early stopping affects the prediction. A second direction is to incorporate feature learning in finite-width networks (Seroussi et al., 2023; Rubin et al., 2024), examining how evolving representations change local uncertainty and the learned score. Comparisons across network widths and data geometries would also help assess the Gaussian-equivalence and pointwise-variance approximations.

Our predictions concern a single, small noise level. Extending the analysis beyond the small-β expansion would test whether the same smoothing and bias mechanisms persist earlier in the reverse process. A joint treatment of different noise levels would account for their coupling through shared network parameters. Combining these extensions with reverse dynamics would connect the predicted score bias to the generated distribution, including its mode coverage and memorization.

AI USE STATEMENT

We used large language models (LLMs) at several stages of this work. In writing, they polished phrasing and checked notation consistency; the authors wrote all sections. In the literature search, LLMs helped identify candidate related work; the authors read and verified every cited reference. In research, LLMs were used for brainstorming and helped implement and debug the numerical experiments. In the derivations in Appendix A, LLMs were used to check intermediate algebraic steps and test candidate identities; the authors independently verified every step and take full responsibility for the correctness of all claims. We did not use AI tools to generate data.

REFERENCES

Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0289.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *Advances in Neural Information Processing Systems*, 38:141266–141286, 2026.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.

George Casella and Christian P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83 (1):81–94, 1996.

Zhengdao Chen. On the interpolation effect of score smoothing in diffusion models. In *International Conference on Learning Representations*, 2026.

Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2):023034, 2021.

Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104018, 2024.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity. In *International Conference on Learning Representations*, volume 2024, pp. 51929–51955, 2024.

Hugo Cui, Cengiz Pehlevan, and Yue M. Lu. A solvable model of learning generative diffusion: Theory and insights. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0187.

Tyler Farghly, Peter Potaptchik, Samuel Howard, George Deligiannidis, and Jakiw Pidstrigach. Diffusion models and the manifold hypothesis: Log-domain smoothing is geometry adaptive. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4095.

Franck Gabriel, François Ged, Maria Han Veiga, and Emmanuel Schertzer. Kernel-smoothed scores for denoising diffusion: A bias-variance study. *arXiv preprint arXiv:2505.22841*, 2025.

Anand Jerry George and Nicolas Macris. Asymptotic learning curves for diffusion models with random features score and manifold data. *arXiv preprint arXiv:2603.22962*, 2026.

Anand Jerry George, Rodrigo Veiga, and Nicolas Macris. Denoising score matching with random features: Insights on diffusion models from precise learning curves. *arXiv preprint arXiv:2502.00336*, 2025.

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hugo Latourelle-Vigeant, Sinho Chewi, Aram-Alexandre Pooladian, John Sous, and Theodor Misiakiewicz. Generalization, memorization, and overfitting for diffusion models trained in the lazy high-dimensional regime. *arXiv preprint arXiv:2608.23938*, 2026.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Physical Review E*, 104(6):064301, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, and Inbar Seroussi. Applications of statistical field theory in deep learning. *arXiv preprint arXiv:2502.18553*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *International Conference on Learning Representations*, volume 2024, pp. 23881–23904, 2024.

Inbar Seroussi, Gadi Naveh, and Zohar Ringel. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, 2023.

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.

John Vastola. Generalization through variance: how noise shapes inductive biases in diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 34922–34960, 2025.

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

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3207.

A Full derivation of the mean and variance predictors

A.1 Notation and standing assumptions

- **Data.** p is a smooth, strictly positive density on $\mathbb{R}^d$; $\mathrm{d}\mu(x) = p(x)\,\mathrm{d}x$; the score is $s(x) = \nabla \log p(x)$, so $\nabla p = p\,s$. The dataset $\{x^{(\mu)}\}_{\mu=1}^{P}$ is i.i.d. from p, and $\mathbb{E}[\cdot]$ denotes the (quenched) average over it.
- **Diffusion.** $x_t = x_0 + \sqrt{\beta}\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I_d)$, at a single fixed time t; the t argument is suppressed throughout. $\sqrt{\beta}$ is the perturbative parameter: $O(\beta^{k/2})$ counts vertex insertions.
- **Kernel.** The main text writes $K(x, y)$ for the scalar NNGP kernel. In this appendix we write $k(x, y)$ for the scalar kernel and $K_{ij}(x, y) = k(x, y)\delta_{ij}$ for the output-index-diagonal matrix kernel; $k(x, y) = \sum_\rho \lambda_\rho \phi_\rho(x)\phi_\rho(y)$ with $\int \mathrm{d}\mu\, \phi_\rho \phi_\sigma = \delta_{\rho\sigma}$.
- **Indices.** Output indices $i, j, l, m \in \{1, \dots, d\}$; kernel modes ρ, σ; data points μ; replicas $a, b, c \in \{1, \dots, n\}$, with a_0 a fixed external replica. All sums are written explicitly.
- **Replica matrices.** δ^{ab} is the Kronecker delta. Γ is the $n \times n$ all-ones matrix, $\Gamma_{ab} = 1$; it satisfies $\Gamma^2 = n\Gamma$.
- **Measures and averages.** $\langle\cdot\rangle_{\mathrm{prior}}$: GP prior. $\langle\cdot\rangle_{\mathrm{ref}}$: reference measure (Sec. A.6). $\langle\cdot\rangle'$: locally damped Gaussian of Lemma 2. Subscript c: connected average, Eq. (62). Inside the replica computation, an average with no subscript means the reference measure.
- **Boundary terms.** Every integration by parts over x assumes vanishing surface terms: $p(x)\,q(x) \to 0$ as $|x| \to \infty$ for every integrand factor q that appears. Here q is always a product of h, its first derivatives, R_x, and γ_x, all bounded in our setting, so sufficient decay of p and ∇p suffices. Integration by parts in ϵ uses the Gaussian decay of the noise density.
- **Shorthands.** $\gamma_x = (1 + R_x/\kappa^2)^{-1} = \kappa^2/(\kappa^2 + R_x)$; in the constant-ridge reduction, $\gamma = (1 + R/\kappa^2)^{-1} = \kappa^2/(\kappa^2 + R)$, $\kappa_{\mathrm{eff}}^2 = \kappa^2 + R$, $\tilde{\lambda}_\rho = (1/\lambda_\rho + P\gamma/\kappa^2)^{-1}$.

A.2 From Langevin training to a function-space partition function

We consider a network with parameters $\boldsymbol{\theta}$, denoted $f_{\boldsymbol{\theta}}(x_t, t)$, whose output has the dimension of its input. In the variance-preserving forward process, $x_t^{(\mu)} = \sqrt{\bar{\alpha}_t}\, x_0^{(\mu)} + \sqrt{1 - \bar{\alpha}_t}\, \epsilon^{(\mu)}$ with $\bar{\alpha}_t =$

$\prod_{u\le t}(1-\beta_u)$. At small times, where the individual variances $\beta_u \ll 1$, write $\beta_{\mathrm{total},t} \equiv \sum_{u\le t}\beta_u$; then $\bar\alpha_t \approx 1-\beta_{\mathrm{total},t}$, $\sqrt{\bar\alpha_t} \approx 1-\frac{1}{2}\beta_{\mathrm{total},t}$, and to leading order

$$x_t^{(\mu)} \approx x_0^{(\mu)} + \sqrt{\beta_{\mathrm{total},t}}\,\epsilon^{(\mu)}. \tag{11}$$

From here on we abbreviate $\beta \equiv \beta_{\mathrm{total},t}$, which is Eq. (1) of the main text (and is exact for the variance-exploding process).

The loss. For a single fixed time t, with the noise averaged over,

$$L(\boldsymbol{\theta}) = \frac{1}{2\kappa^2}\sum_{\mu=1}^{P}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\Big[\big\|\epsilon - f_{\boldsymbol{\theta}}(x_t^{(\mu)},t)\big\|^2\Big]. \tag{12}$$

Langevin dynamics. The parameters follow

$$\mathrm{d}\boldsymbol{\theta}(\tau) = -\nabla_{\boldsymbol{\theta}}\left[\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \eta\,L(\boldsymbol{\theta}(\tau))\right]\mathrm{d}\tau + \sqrt{2}\,\mathrm{d}W(\tau), \tag{13}$$

with τ the training time, W a standard Wiener process, and the factor $\sqrt{2}$ setting the temperature so that the equilibrium density is $P_{\mathrm{eq}}(\boldsymbol{\theta}) \propto \exp[-\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 - \eta\,L(\boldsymbol{\theta})]$. We set $\eta=\lambda=1$ (any other choice rescales κ^2 and the prior variance).

Partition function. Marginalizing over $\boldsymbol{\theta}$ and working in function space, insert $1 = \int Df\,\delta(f - f_{\boldsymbol{\theta}}) = \int Df\int D\tilde f\; e^{i\langle\tilde f,\,(f-f_{\boldsymbol{\theta}})\rangle}$, with $\langle g,h\rangle = \int \mathrm{d}\mu_x \sum_j g_j(x)h_j(x)$. Then, up to constants,

$$Z = \int \mathrm{d}\boldsymbol{\theta}\;\exp\!\big(-L(f_{\boldsymbol{\theta}}) - \tfrac{1}{2}\|\boldsymbol{\theta}\|^2\big) = \int Df\; e^{-L(f)}\int D\tilde f\; e^{-i\langle\tilde f,f\rangle}\;\mathbb{E}_{\boldsymbol{\theta}}\Big[e^{\,i\langle\tilde f,f_{\boldsymbol{\theta}}\rangle}\Big].$$

The last factor is the characteristic function of $f_{\boldsymbol{\theta}}$ under the Gaussian weight-decay prior $\boldsymbol{\theta}\sim\mathcal{N}(0,I)$. At infinite width, $f_{\boldsymbol{\theta}}$ is a Gaussian process on its inputs with the NNGP kernel $K(x,y) = \mathbb{E}_{\boldsymbol{\theta}}[f_{\boldsymbol{\theta}}(x)\,f_{\boldsymbol{\theta}}(y)^\top]$, so $\mathbb{E}_{\boldsymbol{\theta}}[e^{\,i\langle\tilde f,f_{\boldsymbol{\theta}}\rangle}] = e^{-\frac{1}{2}\langle\tilde f,K\tilde f\rangle}$, and the Gaussian integral over $\tilde f$ gives

$$Z = \int Df\;\exp\Big[-\frac{1}{2}\langle f,K^{-1}f\rangle\;-\;\frac{1}{2\kappa^2}\sum_{\mu=1}^{P}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\Big\|f(x_0^{(\mu)}+\sqrt{\beta}\epsilon)-\epsilon\Big\|^2\Big], \tag{14}$$

where $\langle f,K^{-1}f\rangle := \int \mathrm{d}\mu(x)\int\mathrm{d}\mu(y)\sum_{i,j} f_i(x)\,(K^{-1})_{ij}(x,y)\,f_j(y)$. The first term is the prior from the NNGP kernel; the second is the data-dependent loss.

A.3 Replica action

For compactness write the single-replica denoising loss as

$$L_a(x_0) := \mathbb{E}_\epsilon\sum_{i=1}^{d}\big(f_i^a(x_0+\sqrt{\beta}\,\epsilon)-\epsilon_i\big)^2. \tag{15}$$

We replicate and average over the data. Since the $x^{(\mu)}$ are i.i.d. from p, the data-dependent factor of Z^n factorizes across data points:

$$\mathbb{E}[Z^n] = \int\prod_{a=1}^{n}Df^a\;\exp\left[-\frac{1}{2}\sum_{a=1}^{n}\langle f^a,K^{-1}f^a\rangle\right]\left(\int\mathrm{d}\mu(x_0)\exp\left[-\frac{1}{2\kappa^2}\sum_{a=1}^{n}L_a(x_0)\right]\right)^P. \tag{16}$$

Exponentiation (the "Poisson trick"). Write $X_n := \int\mathrm{d}\mu(x_0)\exp[-\frac{1}{2\kappa^2}\sum_a L_a(x_0)]$, so the data factor is $X_n^P = e^{P\log X_n}$. At $n=0$ the exponent is an empty sum and $X_0=1$; for small n, $X_n = 1+O(n)$, hence

$$P\log X_n = P\,(X_n-1) + O\big(P\,n^2\big), \tag{17}$$

and the $O(n^2)$ piece does not survive the replica limit. (An equivalent exact reading: Poissonize the dataset size, $P\sim\mathrm{Poisson}(\bar P)$; the exponential is then exact and the $n\to0$ physics is identical.) Therefore $\mathbb{E}[Z^n] = \int\prod_a Df^a\;e^{-S[\{f^a\}]}$ with

$$S[\{f^a\}] = \frac{1}{2}\sum_{a=1}^{n}\langle f^a,K^{-1}f^a\rangle + P\;-\;P\int\mathrm{d}\mu(x_0)\exp\left[-\frac{1}{2\kappa^2}\sum_{a=1}^{n}L_a(x_0)\right]. \tag{18}$$

A.4 NOISE INTEGRATION BY PARTS

For one replica, expanding the square in (15),

$$L_a(x_0) = \mathbb{E}_\epsilon |f^a(x_0 + \sqrt{\beta}\,\epsilon)|^2 \;-\; 2\,\mathbb{E}_\epsilon \sum_{i=1}^{d} f_i^a(x_0 + \sqrt{\beta}\,\epsilon)\,\epsilon_i \;+\; d. \tag{19}$$

Write the cross term as a Gaussian integral and integrate by parts in ϵ (Stein's identity, $\epsilon_i \mathcal{P}(\epsilon) = -\partial_{\epsilon_i}\mathcal{P}(\epsilon)$ for the standard Gaussian density $\mathcal{P}$; the surface term vanishes by Gaussian decay):

$$\mathbb{E}_\epsilon \sum_{i=1}^{d} f_i^a(x_0 + \sqrt{\beta}\,\epsilon)\,\epsilon_i = \int \mathrm{d}^d\epsilon\, \mathcal{P}(\epsilon) \sum_i \frac{\partial}{\partial \epsilon_i} f_i^a(x_0 + \sqrt{\beta}\,\epsilon) = \sqrt{\beta}\,\mathbb{E}_\epsilon\, \nabla\cdot f^a(x_0 + \sqrt{\beta}\,\epsilon). \tag{20}$$

Thus

$$S[\{f^a\}] = S_{\text{prior}} + P - P\int \mathrm{d}\mu(x_0) \exp\left[-\frac{1}{2\kappa^2}\sum_{a=1}^{n} E_a(x_0)\right], \tag{21}$$
$$E_a(x_0) := \mathbb{E}_\epsilon |f^a(x_0 + \sqrt{\beta}\,\epsilon)|^2 - 2\sqrt{\beta}\,\mathbb{E}_\epsilon \nabla\cdot f^a(x_0 + \sqrt{\beta}\,\epsilon) + d,$$

where S_{prior} denotes the first (kernel) term of (18).

A.5 SMALL-β EXPANSION

Taylor expand each replica field, $f^a(x_0 + \sqrt{\beta}\,\epsilon) = f^a + \sqrt{\beta}\,(\epsilon\cdot\nabla)f^a + \frac{\beta}{2}(\epsilon\cdot\nabla)^2 f^a + O(\beta^{3/2})$, all fields on the right evaluated at x_0. Squaring, averaging over ϵ (odd terms vanish; $\mathbb{E}[\epsilon_l\epsilon_m] = \delta_{lm}$ gives $\mathbb{E}_\epsilon[(\epsilon\cdot\nabla)f_i]^2 = |\nabla f_i|^2$ and $\mathbb{E}_\epsilon[(\epsilon\cdot\nabla)^2 f_i] = \nabla^2 f_i$),

$$\mathbb{E}_\epsilon |f^a(x_0 + \sqrt{\beta}\,\epsilon)|^2 = |f^a|^2 + \beta\sum_i \left(f_i^a \nabla^2 f_i^a + |\nabla f_i^a|^2\right) + O(\beta^2) = |f^a|^2 + \frac{\beta}{2}\nabla^2|f^a|^2 + O(\beta^2), \tag{22}$$

the last equality from $\partial_l\partial_l(f_i f_i) = 2(\partial_l f_i)(\partial_l f_i) + 2f_i\,\partial_l\partial_l f_i$ (summed over i, l). Similarly $\mathbb{E}_\epsilon \nabla\cdot f^a(x_0 + \sqrt{\beta}\,\epsilon) = \nabla\cdot f^a + \frac{\beta}{2}\nabla^2(\nabla\cdot f^a) + O(\beta^2)$. Hence

$$E_a = |f^a|^2 + d - 2\sqrt{\beta}\,\nabla\cdot f^a + \frac{\beta}{2}\nabla^2|f^a|^2 + O(\beta^{3/2}), \tag{23}$$

where the omitted $O(\beta^{3/2})$ piece, $-\beta^{3/2}\nabla^2(\nabla\cdot f^a)$, is odd under $f \mapsto -f$ and enters the mean only at $O(\beta^{3/2})$. Exponentiating and expanding the β-dependent part,

$$\exp\left[-\frac{1}{2\kappa^2}\sum_a E_a\right] = e^{-nd/2\kappa^2}\; e^{-\frac{1}{2\kappa^2}\sum_a |f^a|^2}\left[1 + \frac{\sqrt{\beta}}{\kappa^2}\sum_a \nabla\cdot f^a + O(\beta)\right], \tag{24}$$

where the $O(\beta)$ collects terms *even* under $f \mapsto -f$ (the $\nabla^2|f^a|^2$ insertion and the square of the divergence term); their role in the error budget is settled by parity in Sec. A.9.[3] Keeping the terms needed for the $O(\sqrt{\beta})$ mean,

$$S[\{f^a\}] = S_{\text{prior}} + P - P\int \mathrm{d}\mu(x)\, e^{-\tilde{U}(x)}\left[1 + \frac{\sqrt{\beta}}{\kappa^2}\sum_{a=1}^{n}\nabla\cdot f^a(x)\right] + O(\beta), \tag{25}$$

with $\tilde{U}(x) = \frac{1}{2\kappa^2}\sum_{b=1}^{n}|f^b(x)|^2$. We define the $O(\sqrt{\beta})$ vertex and the reference action

$$S_{\text{v}} = -\frac{P\sqrt{\beta}}{\kappa^2}\int \mathrm{d}\mu(x)\, e^{-\tilde{U}(x)}\sum_{a=1}^{n}\nabla\cdot f^a(x), \qquad S_{\text{ref}} := S - S_{\text{v}}, \tag{26}$$

so that S_{ref} is even under $f \mapsto -f$ and S_{v} is odd. Note that $\tilde{U}(x)$ involves *all* n replicas at the point x; this matters for the damping block in Lemma 2.

[3] The constant $e^{-nd/2\kappa^2}$ is field-independent and tends to 1 as $n \to 0$; it is dropped from here on.

A.6 GAUSSIAN-EQUIVALENCE REFERENCE MEASURE ON FIBRES

Convention. The kernel is diagonal in the output index, $K_{ij}(x,y) = k(x,y)\,\delta_{ij}$, and the field is expanded with one coefficient per (mode, output index),

$$f_i^a(x) = \sum_\rho c_{\rho i}^a\,\phi_\rho(x), \qquad \langle c_{\rho i}^a\, c_{\sigma j}^a\rangle_{\text{prior}} = \lambda_\rho\,\delta_{\rho\sigma}\delta_{ij}, \tag{27}$$

which reproduces the GP prior $\langle f_i(x) f_j(y)\rangle = k(x,y)\,\delta_{ij}$. Explicitly,

$$S_{\text{ref}} = \frac{1}{2}\sum_{a=1}^{n}\langle f^a, K^{-1} f^a\rangle \;+\; P \;-\; P\int \mathrm{d}\mu(x)\; e^{-\tilde U(x)}. \tag{28}$$

The data term couples the nd field values at a single point. Define, at a fixed point x, $Y_i^a(x) := f_i^a(x)$, $Y(x) \in \mathbb{R}^{nd}$, and the pointwise overlap

$$Q_x[\{c\}]_{ab} := \frac{1}{d}\sum_{i=1}^{d} f_i^a(x)\, f_i^b(x), \qquad \tilde U(x) = \frac{d}{2\kappa^2}\;\mathrm{Tr}\, Q_x[\{c\}]. \tag{29}$$

Fibres. Let $E \subset \mathbb{R}^d$ be the subspace of the few dominant directions of the data and $E^\perp$ the directions along which the population has small spread $\sigma_\perp$. Write $x = (s,u)$ with $s = \Pi_E x \in E$ and $u = (1-\Pi_E)x \in E^\perp$, and take as the neighbourhood of x its fibre,

$$N(x) := N_{s(x)}, \qquad N_s := \{x' : \Pi_E x' = s\}. \tag{30}$$

Distinct fibres are disjoint and the fibres cover $\mathbb{R}^d$. With $\mathrm{d}\mu(x) = p(s,u)\,\mathrm{d}s\,\mathrm{d}u$, define the marginal density of the dominant coordinates and the conditional density on the fibre,

$$p_E(s) := \int_{E^\perp} p(s,u)\,\mathrm{d}u, \qquad p_s(u) := \frac{p(s,u)}{p_E(s)}, \qquad \mathrm{d}\nu(s) := p_E(s)\,\mathrm{d}s, \qquad \mathrm{d}\mu_s(u) := p_s(u)\,\mathrm{d}u,$$

so that $\int_{E^\perp}\mathrm{d}\mu_s = 1$ and $\mathrm{d}\nu(s)\,\mathrm{d}\mu_s(u) = \mathrm{d}\mu(x)$. For $x \in \mathbb{R}^d$ we write $\mu_{s(x)}$ for the conditional measure on the fibre through x. Define the fibre moments of the field configuration,

$$\bar Y_{s,i}^a := \int \mathrm{d}\mu_s(x')\, f_i^a(x') \qquad \text{(linear in the fields)}, \tag{31}$$

$$S_s^{(a,i),(b,j)} := \int \mathrm{d}\mu_s(x')\, f_i^a(x') f_j^b(x') \qquad \text{(uncentred second moment, quadratic)}, \tag{32}$$

$$C_s^{(a,i),(b,j)} := S_s^{(a,i),(b,j)} - \bar Y_{s,i}^a \bar Y_{s,j}^b \qquad \text{(centred covariance)}. \tag{33}$$

The supervised theory of Canatar et al. (2021); Cohen et al. (2021) uses a single fibre, $E = \{0\}$, $N(x) = \mathbb{R}^d$, with $\bar Y = 0$ by symmetry and $S = \int \mathrm{d}\mu\, f f^\top$; the present construction refines it by conditioning on the dominant coordinates.

Gaussian-equivalence claim. Each component $Y_i^a(x')$, $x' \sim \mu_s$, is a sum over many kernel modes. We claim that, conditionally on the fibre, the vector Y is approximately Gaussian with the matched moments,

$$Y \,\big|\, N(x) \;\sim\; \mathcal{N}\big(\bar Y_{s(x)},\, C_{s(x)}\big). \tag{34}$$

Output isotropy. The reference measure is even under $f \mapsto -f$ and invariant under rotations of the output index. Hence the mean-field values of the fibre moments are

$$\big\langle \bar Y_{s,i}^a \big\rangle_{\text{ref}} = 0, \qquad \Big\langle S_s^{(a,i),(b,j)} \Big\rangle_{\text{ref}} = Q_{s,ab}^{\text{MF}}\,\delta_{ij}. \tag{35}$$

Step 1: the source-free data term as a position-resolved Gaussian integral.

Lemma 1 (Shifted Gaussian damping)**.** *For $Y \sim \mathcal{N}(\bar Y, C)$ on $\mathbb{R}^N$,*

$$\mathbb{E}\big[e^{-|Y|^2/(2\kappa^2)}\big] = \det\big(I + \kappa^{-2}C\big)^{-1/2}\exp\Big[-\tfrac{1}{2}\,\bar Y^\top\big(\kappa^2 I + C\big)^{-1}\bar Y\Big]. \tag{36}$$

Proof. Complete the square: the exponent of $e^{-\frac{1}{2}(Y-\bar{Y})^\top C^{-1}(Y-\bar{Y})}e^{-|Y|^2/2\kappa^2}$ is $-\frac{1}{2}Y^\top \mathsf{P} Y + \bar{Y}^\top C^{-1}Y - \frac{1}{2}\bar{Y}^\top C^{-1}\bar{Y}$ with $\mathsf{P} = C^{-1} + \kappa^{-2}I$. Shifting $Y = \mathsf{P}^{-1}C^{-1}\bar{Y} + Z$ and integrating Z gives $\det(C\mathsf{P})^{-1/2} = \det(I + \kappa^{-2}C)^{-1/2}$ times $\exp[-\frac{1}{2}\bar{Y}^\top(C^{-1} - C^{-1}\mathsf{P}^{-1}C^{-1})\bar{Y}]$, and by Woodbury $C^{-1} - C^{-1}(C^{-1} + \kappa^{-2}I)^{-1}C^{-1} = (\kappa^2 I + C)^{-1}$. At $\bar{Y} = 0$ this is the normalization identity (56) of Lemma 2. □

Decompose the population into fibres and apply the claim (34) on each:

$$\begin{aligned}\int \mathrm{d}\mu(x)\, e^{-\tilde{U}(x)} &= \int \mathrm{d}\nu(s)\, \mathbb{E}_{x'\sim\mu_s}\Big[e^{-\frac{1}{2\kappa^2}\sum_a |f^a(x')|^2}\Big] \\ &\approx \int \mathrm{d}\nu(s)\, \det\big(I_{nd} + \kappa^{-2}C_s\big)^{-1/2} \exp\Big[-\tfrac{1}{2}\,\bar{Y}_s^\top\big(\kappa^2 I + C_s\big)^{-1}\bar{Y}_s\Big].\end{aligned} \tag{37}$$

The data part of the action is therefore

$$S_{\mathrm{data}} = P - P\int \mathrm{d}\nu(s)\, e^{\Phi_s}, \qquad \Phi_s := -\tfrac{1}{2}\,\mathrm{Tr}\log\big(I + \kappa^{-2}C_s\big) - \tfrac{1}{2}\,\bar{Y}_s^\top\big(\kappa^2 I + C_s\big)^{-1}\bar{Y}_s. \tag{38}$$

Step 2: mean-field split and linearization. Split $\bar{Y}_s = \bar{Y}_s^{\mathrm{MF}} + \Delta\bar{Y}_s$ and $S_s = S_s^{\mathrm{MF}} + \Delta S_s$, with $C_s = S_s - \bar{Y}_s\bar{Y}_s^\top$. By the symmetry (35), $\bar{Y}_s^{\mathrm{MF}} = 0$, so $\bar{Y}_s = \Delta\bar{Y}_s$ is pure fluctuation, and $S_s^{\mathrm{MF}} = Q_s^{\mathrm{MF}} \otimes I_d$. Define

$$M_s := I_n + \kappa^{-2}\, Q_s^{\mathrm{MF}}, \qquad W_s := \kappa^{-2}\,(M_s \otimes I_d)^{-1} = \big(\kappa^2 I + Q_s^{\mathrm{MF}} \otimes I_d\big)^{-1}.$$

For the first term of Φ_s use $e^{-\frac{1}{2}\,\mathrm{Tr}\log(M+X)} = e^{-\frac{1}{2}\,\mathrm{Tr}\log M}(1 - \frac{1}{2}\,\mathrm{Tr}[M^{-1}X]) + O(X^2)$ with $M = M_s \otimes I_d$ and $X = \kappa^{-2}(\Delta S_s - \Delta\bar{Y}_s\Delta\bar{Y}_s^\top)$, and $\mathrm{Tr}\log(M_s \otimes I_d) = d\ \mathrm{Tr}\log M_s$:

$$\begin{aligned}e^{-\frac{1}{2}\,\mathrm{Tr}\log(I+\kappa^{-2}C_s)} = e^{-\frac{d}{2}\,\mathrm{Tr}\log M_s}\Big(1 - \tfrac{1}{2}\,\mathrm{Tr}\big[W_s\,\Delta S_s\big] + \tfrac{1}{2}\,\Delta\bar{Y}_s^\top W_s\,\Delta\bar{Y}_s\Big) \\ + O\big(\Delta S_s^2,\ \Delta\bar{Y}_s^2\Delta S_s,\ \Delta\bar{Y}_s^4\big).\end{aligned} \tag{39}$$

For the second term use $(A+\varepsilon)^{-1} = A^{-1} - A^{-1}\varepsilon A^{-1} + O(\varepsilon^2)$ with $A = \kappa^2 I + Q_s^{\mathrm{MF}} \otimes I_d = W_s^{-1}$ and $\varepsilon = \Delta S_s - \Delta\bar{Y}_s\Delta\bar{Y}_s^\top$; sandwiching between $\Delta\bar{Y}_s^\top$ and $\Delta\bar{Y}_s$ adds two more powers of $\Delta\bar{Y}_s$, so every correction beyond the leading W_s is of the order already being neglected, and

$$e^{-\frac{1}{2}\,\Delta\bar{Y}_s^\top(\kappa^2 I + C_s)^{-1}\Delta\bar{Y}_s} = 1 - \tfrac{1}{2}\,\Delta\bar{Y}_s^\top W_s\,\Delta\bar{Y}_s + O\big(\Delta\bar{Y}_s^2\Delta S_s,\ \Delta\bar{Y}_s^4\big). \tag{40}$$

Multiplying (39) and (40), the two mean terms, $+\frac{1}{2}\Delta\bar{Y}_s^\top W_s\Delta\bar{Y}_s$ from the $-\bar{Y}_s\bar{Y}_s^\top$ inside C_s and $-\frac{1}{2}\Delta\bar{Y}_s^\top W_s\Delta\bar{Y}_s$ from the shift, are equal and opposite:

$$e^{\Phi_s} = A_s\Big(1 - \tfrac{1}{2}\,\mathrm{Tr}\big[W_s\,\Delta S_s\big]\Big) + O\big(\Delta S_s^2,\ \Delta\bar{Y}_s^2\Delta S_s,\ \Delta\bar{Y}_s^4\big), \qquad A_s := e^{-\frac{d}{2}\,\mathrm{Tr}\log M_s}. \tag{41}$$

To this order the fibre mean has dropped out and the data factor depends on the configuration only through ΔS_s. Substituting $(\Delta S_s)^{(a,i),(b,i)} = \int \mathrm{d}\mu_s\, f_i^a f_i^b - Q_{s,ab}^{\mathrm{MF}}$,

$$\begin{aligned}S_{\mathrm{data}} = P - P\int \mathrm{d}\nu(s)\, A_s\Big(1 + \frac{d}{2\kappa^2}\, B_s\Big) \\ + \frac{P}{2\kappa^2}\int \mathrm{d}\nu(s)\, A_s \sum_{i=1}^{d}\sum_{a,b=1}^{n}\big(M_s^{-1}\big)_{ab}\int \mathrm{d}\mu_s\, f_i^a f_i^b \;+\; O(\cdots),\end{aligned} \tag{42}$$

with $B_s := \mathrm{Tr}(M_s^{-1}Q_s^{\mathrm{MF}})$. Dropping the $O(\cdots)$ terms is the quadratic (Gaussian-equivalence) truncation; the terms that still contain $\Delta\bar{Y}_s$ are all in it.

Step 3: RS ansatz on each fibre; $A_s \to 1$ and $B_s \to 0$ in the replica limit. Take the replica-symmetric ansatz fibre by fibre,

$$Q_s^{\mathrm{MF}} = R_s\, I_n + T_s\, \Gamma \quad\Longrightarrow\quad M_s = \alpha_s\, I_n + b_s\, \Gamma, \qquad \alpha_s := 1 + \frac{R_s}{\kappa^2}, \quad b_s := \frac{T_s}{\kappa^2},$$

with two scalar *functions* R_s, T_s on E to be determined. Γ has eigenvalue n once and 0 with multiplicity $n-1$, so M_s has eigenvalues $\alpha_s + nb_s$ (once) and α_s ($n-1$ times). Hence, exactly,

$$\operatorname{Tr}\log M_s = (n-1)\log\alpha_s + \log(\alpha_s + nb_s) = n\log\Big(1 + \frac{R_s}{\kappa^2}\Big) + \frac{nT_s}{\kappa^2 + R_s} + O(n^2) \tag{43}$$

so that $A_s = 1 + O(n) \to 1$ as $n \to 0$. The same eigenstructure gives the exact inverse

$$M_s^{-1} = \frac{1}{\alpha_s} I_n - \frac{b_s}{\alpha_s(\alpha_s + nb_s)}\,\Gamma \;\xrightarrow[n\to 0]{}\; \gamma_s\, I_n - \frac{T_s\gamma_s^2}{\kappa^2}\,\Gamma, \qquad \gamma_s = \frac{1}{\alpha_s} = \frac{\kappa^2}{\kappa^2 + R_s}, \tag{44}$$

from which $\operatorname{Tr} M_s^{-1} = n\gamma_s - n\,T_s\gamma_s^2/\kappa^2 + O(n^2)$ and $\operatorname{Tr}(M_s^{-1}\Gamma) = n\gamma_s + O(n^2)$, so

$$B_s = R_s\,\operatorname{Tr} M_s^{-1} + T_s\,\operatorname{Tr}\big(M_s^{-1}\Gamma\big) = n\big[\gamma_s R_s + \gamma_s^2 T_s\big] + O(n^2) \xrightarrow[n\to 0]{} 0, \tag{45}$$

using $1 - \gamma_s R_s/\kappa^2 = \gamma_s$. Both $A_s - 1$ and B_s are $O(n)$ for each s, so the field-independent block of (42) is $O(n)$ times constants and, like the $O(n^2)$ terms already discarded at the Poisson step (17), contributes nothing to quenched observables; the prefactor A_s of the quadratic term is likewise 1 in the limit. The damping factor that survives is $\gamma_s = \kappa^2/(\kappa^2 + R_s)$.

Step 4: the quadratic reference action. The prior term is diagonal in the modes: with $(K^{-1})_{ij} = k^{-1}\delta_{ij}$ and $\iint \mathrm{d}\mu_x \mathrm{d}\mu_y\, \phi_\rho(x)\, k^{-1}(x,y)\, \phi_\sigma(y) = \delta_{\rho\sigma}/\lambda_\rho$, $S_{\text{prior}} = \frac{1}{2}\sum_a \sum_\rho \sum_i (c^a_{\rho i})^2/\lambda_\rho$. Adding the field-dependent part of (42) with $A_s \to 1$ and using $\mathrm{d}\nu(s)\mathrm{d}\mu_s(u) = \mathrm{d}\mu(x)$,

$$S_{\text{ref}} \;\approx\; \frac{1}{2}\sum_{a=1}^{n}\sum_{\rho}\sum_{i=1}^{d}\frac{\big(c^a_{\rho i}\big)^2}{\lambda_\rho} \;+\; \frac{P}{2\kappa^2}\int \mathrm{d}\mu(x)\sum_{i=1}^{d}\sum_{a,b=1}^{n} f_i^a(x)\,\big(M^{-1}_{s(x)}\big)_{ab}\, f_i^b(x) \;+\; \text{const.} \tag{46}$$

The reference measure is Gaussian and block diagonal in the output index i, but *not* diagonal in the modes. Let $\hat{g}$ denote multiplication by a function g, $(\hat{g}\,\varphi)(x) = g(x)\,\varphi(x)$, with matrix elements $g_{\rho\sigma} := \int \mathrm{d}\mu\; g\,\phi_\rho\phi_\sigma$ in the kernel eigenbasis. Inserting the $n \to 0$ inverse (44), the replica inverse propagator is the operator-valued $n \times n$ matrix

$$G^{-1} = \Big(k^{-1} + \frac{P}{\kappa^2}\,\hat{\gamma}\Big) \otimes I_n \;-\; \frac{P}{\kappa^4}\,\widehat{T\gamma^2} \otimes \Gamma, \tag{47}$$

where $\hat{\gamma}$ is multiplication by $\gamma_{s(x)}$ and $\widehat{T\gamma^2}$ multiplication by $T_{s(x)}\gamma^2_{s(x)}$. The local damping mixes the kernel eigenfunctions: G^{-1} is diagonal in the modes only if $\gamma_{s(x)}$ and $T_{s(x)}\gamma^2_{s(x)}$ are constant over the population.

Explicit inversion. Write $G^{-1} = \hat{\alpha} \otimes I_n + \hat{b} \otimes \Gamma$ with the operators $\hat{\alpha} = k^{-1} + \frac{P}{\kappa^2}\,\hat{\gamma}$ and $\hat{b} = -\frac{P}{\kappa^4}\,\widehat{T\gamma^2}$. Make the ansatz $G = u \otimes I_n + v \otimes \Gamma$ with operators u, v (which need not commute with $\hat{\alpha}$ and $\hat{b}$, so the order of factors is kept) and use $\Gamma^2 = n\Gamma$:

$$\begin{aligned}\big(\hat{\alpha} \otimes I_n + \hat{b} \otimes \Gamma\big)\big(u \otimes I_n + v \otimes \Gamma\big) &= \hat{\alpha}u \otimes I_n + \big(\hat{\alpha}v + \hat{b}u + n\,\hat{b}v\big) \otimes \Gamma \overset{!}{=} \mathbf{1} \otimes I_n \\ \Longrightarrow \quad u = \hat{\alpha}^{-1}, &\qquad v = -\big(\hat{\alpha} + n\hat{b}\big)^{-1}\hat{b}\,\hat{\alpha}^{-1} \xrightarrow[n\to 0]{} -\hat{\alpha}^{-1}\hat{b}\,\hat{\alpha}^{-1}.\end{aligned}$$

With $h := \hat{\alpha}^{-1}$, an operator with kernel $h(x,y)$,

$$G^{ab}(x,y) = h(x,y)\,\delta^{ab} + \frac{P}{\kappa^4}\int \mathrm{d}\mu(z)\; h(x,z)\,T_{s(z)}\,\gamma^2_{s(z)}\,h(z,y)\,\Gamma_{ab}, \qquad h = \Big(k^{-1} + \frac{P}{\kappa^2}\,\hat{\gamma}\Big)^{-1}. \tag{48}$$

Step 5: closing the self-consistency, fibre by fibre. The saddle condition fixing Q_s^{MF} is moment matching: the tadpole of the neglected linear term must vanish, $\langle \Delta S_s \rangle_{\text{ref}} = 0$, i.e. by (35) and (32),

$$Q^{\text{MF}}_{s,ab} = \frac{1}{d}\sum_{i=1}^{d}\int \mathrm{d}\mu_s(u)\,\big\langle f_i^a(s,u)\, f_i^b(s,u)\big\rangle_{\text{ref}} = \int \mathrm{d}\mu_s(u)\; G^{ab}\big((s,u),(s,u)\big). \tag{49}$$

Inserting (48) and matching the two matrix structures of the RS ansatz,

$$\begin{cases} R_s = \displaystyle\int \mathrm{d}\mu_s(u)\; h\big((s,u),(s,u)\big) = \mathbb{E}\big[\, R_x \,\big|\, \Pi_E x = s \,\big], \quad R_x := h(x,x), \quad \gamma_s = \dfrac{\kappa^2}{\kappa^2 + R_s}, \\ T_s = \dfrac{P}{\kappa^4} \displaystyle\int \mathrm{d}\mu_s(u) \int \mathrm{d}\mu(z)\; h\big((s,u),z\big)^2\, \gamma_{s(z)}^2\, T_{s(z)}. \end{cases} \tag{50}$$

The first line is the self-consistency for the fibre-averaged variance R_s, which enters the ridge through γ_s; note that h itself depends on $\{R_s\}$ through $\hat{\gamma}$, so this is a fixed-point equation for the function $s \mapsto R_s$. The second line is a homogeneous linear equation $T = \mathcal{K}T$ with a positive kernel; away from the replica-symmetry instability at which $\mathcal{K}$ acquires eigenvalue one, it forces $T_s = 0$ for all s. (In the full $\beta > 0$ theory the $O(\sqrt{\beta})$ source regenerates $T = O(\beta)$ through this same structure, but since the vertex already carries $\sqrt{\beta}$ that T would feed the mean only at $O(\beta^{3/2})$.) We therefore set $T_s = 0$, and the replica-diagonal propagator is exact at this order:

$$\begin{gathered} \big\langle f_i^a(x)\, f_j^b(y) \big\rangle_{\mathrm{ref}} = \delta^{ab}\, \delta_{ij}\, h(x,y), \qquad h = \Big(k^{-1} + \frac{P}{\kappa^2}\,\hat{\gamma}\Big)^{-1}, \\ \big[h^{-1}\big]_{\rho\sigma} = \frac{\delta_{\rho\sigma}}{\lambda_\rho} + \frac{P}{\kappa^2} \int \mathrm{d}\mu\; \gamma_{s(x)}\, \phi_\rho \phi_\sigma. \end{gathered} \tag{51}$$

A.7 Local covariance inventory

At the vertex point x, define

$$Y_m^b := f_m^b(x), \qquad Z_{lm}^b := \partial_{x_l} f_m^b(x), \qquad F := f_i^{a_0}(y).$$

(We reuse the symbol Y from Sec. A.6 for the field values at the vertex point, but the conditioning is now reversed: x is a *fixed* point, and the randomness is the field under $\langle\cdot\rangle_{\mathrm{ref}}$.) Let $\partial_{2,l}$ denote a derivative with respect to the l-th component of the *second* argument of h (and $\partial_{1,l}$ of the first). The undamped covariances all follow from (51):

$$\langle Y_l^a\, Y_m^b \rangle = \delta^{ab}\delta_{lm}\, R_x, \qquad \langle F\, Y_m^b \rangle = \delta^{a_0 b}\delta_{im}\, h(y,x), \tag{52}$$

and differentiating (51) once with respect to the second argument,

$$\langle Y_r^c\, Z_{lm}^b \rangle = \delta^{cb}\delta_{rm}\, \partial_{2,l} h(x,x), \qquad \langle F\, Z_{lm}^b \rangle = \delta^{a_0 b}\delta_{im}\, \partial_{2,l} h(y,x). \tag{53}$$

The coincident-point identity is $\partial_l R_x = \partial_l h(x,x) = (\partial_{1,l} + \partial_{2,l}) h(x,x) = 2\,\partial_{2,l} h(x,x)$, the last step by the symmetry $h(x,y) = h(y,x)$. Hence

$$\partial_{2,l} h(x,x) = \frac{1}{2}\,\partial_l R_x. \tag{54}$$

No $\langle ZZ \rangle$ covariance is needed: it never appears at one vertex insertion.

A.8 Exact local damping lemma

Lemma 2 (Gaussian damping). *Let $V = (Y, W) \in \mathbb{R}^{N_Y + N_W}$ be jointly Gaussian, centred, with covariance C (blocks $C_{YY}, C_{YW}, C_{WY}, C_{WW}$). Let $\Pi_Y := \mathrm{diag}(I_{N_Y}, 0)$ be the orthogonal projector onto the Y coordinates, so that $|Y|^2 = V^\top \Pi_Y V$. Then for every polynomial G,*

$$\Big\langle e^{-|Y|^2/(2\kappa^2)}\, G(V) \Big\rangle = \Big\langle e^{-|Y|^2/(2\kappa^2)} \Big\rangle\, \langle G(V) \rangle', \tag{55}$$

where $\langle\cdot\rangle'$ is Gaussian, centred, with covariance $C' = (C^{-1} + \kappa^{-2}\Pi_Y)^{-1}$. The normalization is

$$\Big\langle e^{-|Y|^2/(2\kappa^2)} \Big\rangle = \det\big(I + \kappa^{-2} C_{YY}\big)^{-1/2}, \tag{56}$$

and, for arbitrary components U, V,

$$C'_{UV} = C_{UV} - C_{UY}\big(\kappa^2 I + C_{YY}\big)^{-1} C_{YV}. \tag{57}$$

Proof. *(i)* Multiplying the Gaussian density $\propto e^{-V^\top C^{-1}V/2}$ by $e^{-V^\top \Pi_Y V/(2\kappa^2)}$ changes the quadratic form from C^{-1} to $C^{-1}+\kappa^{-2}\Pi_Y$; the reweighted measure is Gaussian, centred, with covariance C', times a constant, and dividing by that constant is (55). *(ii)* The constant is $\sqrt{\det C'/\det C} = \det(I+\kappa^{-2}\,C\,\Pi_Y)^{-1/2}$. *(iii)* In the ordering (Y,W), $I+\kappa^{-2}C\,\Pi_Y$ is block lower triangular with diagonal blocks $I+\kappa^{-2}C_{YY}$ and I, so the determinant reduces to $\det(I+\kappa^{-2}C_{YY})$, giving (56). *(iv)* The Woodbury identity applied to C', with the perturbation supported on the Y block, gives $C' = C - C\,\Pi_Y(\kappa^2 I+\Pi_Y C\,\Pi_Y)^{-1}\Pi_Y\,C$, which componentwise is (57). □

Application at the vertex. The damped block is $Y=\{Y^b_m : b=1,\dots,n;\ m=1,\dots,d\}$, all n replicas and all d output components at the vertex point x (recall that $\tilde U(x)$ in (26) sums over all replicas), so $N_Y = nd$ and, by (52), $C_{YY} = R_x\,I_{nd}$ and $(\kappa^2 I + C_{YY})^{-1} = (\gamma_x/\kappa^2)\,I_{nd}$. The damping normalization is

$$\left\langle e^{-\tilde U(x)}\right\rangle = \left(1+R_x/\kappa^2\right)^{-nd/2} \xrightarrow[n\to 0]{} 1. \tag{58}$$

Each damped covariance follows from (57). With one leg in the Y block, $\langle F\,Y^b_m\rangle' = \gamma_x\,\langle F Y^b_m\rangle$ and $\langle Y^c_r\,Z^b_{lm}\rangle' = \gamma_x\,\langle Y^c_r Z^b_{lm}\rangle$; with both legs in the Y block, $\langle Y^a_l\,Y^b_m\rangle' = \gamma_x R_x\,\delta^{ab}\delta_{lm}$. The only covariance with *both* legs outside the damped block is

$$\langle F Z^b_{lm}\rangle' = \langle F Z^b_{lm}\rangle - \sum_{c,r}\langle F Y^c_r\rangle\,\frac{\gamma_x}{\kappa^2}\,\langle Y^c_r Z^b_{lm}\rangle;$$

inserting (52), (53), the first factor forces $c=a_0$, $r=i$, and the second then forces $b=a_0$, $m=i$:

$$\langle F Z^b_{lm}\rangle' = \delta^{a_0 b}\,\delta_{im}\left[\partial_{2,l}h(y,x) - \frac{h(y,x)\,\partial_{2,l}h(x,x)}{\kappa^2+R_x}\right]. \tag{59}$$

A.9 Mean predictor

Strategy. Every quantity of interest is an average $\langle F[f]\rangle = \int Df\ F\,e^{-S}/\int Df\ e^{-S}$. Our action is not quadratic; we split $S = S_{\rm ref}+S_{\rm v}$, where $S_{\rm ref}$ defines the solvable reference measure and $S_{\rm v}$ is small. Exactly,

$$\langle F\rangle = \frac{\left\langle F\,e^{-S_{\rm v}}\right\rangle_{\rm ref}}{\left\langle e^{-S_{\rm v}}\right\rangle_{\rm ref}}. \tag{60}$$

Writing $\langle\cdot\rangle$ for $\langle\cdot\rangle_{\rm ref}$ in this subsection, expand numerator and denominator, $N = \langle F\rangle - \langle F S_{\rm v}\rangle + \frac12\langle F S_{\rm v}^2\rangle - \dots$ and $D = 1-\langle S_{\rm v}\rangle + \frac12\langle S_{\rm v}^2\rangle - \dots$, invert the denominator as a geometric series, and collect orders:

$$\begin{aligned}\langle F\rangle_{\rm full} = ND^{-1} = \langle F\rangle\ &-\ \Big[\langle F S_{\rm v}\rangle - \langle F\rangle\langle S_{\rm v}\rangle\Big]\\ &+\ \Big[\tfrac12\langle F S_{\rm v}^2\rangle - \langle F S_{\rm v}\rangle\langle S_{\rm v}\rangle - \tfrac12\langle F\rangle\langle S_{\rm v}^2\rangle + \langle F\rangle\langle S_{\rm v}\rangle^2\Big]\ +\ O(S_{\rm v}^3).\end{aligned} \tag{61}$$

The brackets are the *joint cumulants* of F with one and two insertions of $S_{\rm v}$,

$$\langle F S_{\rm v}\rangle_c := \langle F S_{\rm v}\rangle - \langle F\rangle\langle S_{\rm v}\rangle, \tag{62}$$

$$\langle F S_{\rm v}^2\rangle_c := \langle F S_{\rm v}^2\rangle - 2\langle S_{\rm v}\rangle\langle F S_{\rm v}\rangle - \langle F\rangle\langle S_{\rm v}^2\rangle + 2\langle F\rangle\langle S_{\rm v}\rangle^2, \tag{63}$$

and the pattern continues to all orders (linked-cluster theorem):

$$\langle F\rangle_{\rm full} = \sum_{k\ge 0}\frac{(-1)^k}{k!}\,\left\langle F\,S_{\rm v}^k\right\rangle_{{\rm ref},\,c}. \tag{64}$$

Parity, term by term. $S_{\rm ref}$ is even under $f\mapsto -f$ while $F = f_i^{a_0}(y)$ and $S_{\rm v}$ are odd. Consequently $\langle F\rangle_{\rm ref} = 0$ and $\langle S_{\rm v}\rangle_{\rm ref} = 0$; at first order the connected correlator equals the plain moment, $\langle F S_{\rm v}\rangle_c = \langle F S_{\rm v}\rangle$; at second order *every term* of (63) vanishes individually ($\langle F S_{\rm v}^2\rangle$ is an odd moment, and each subtraction term contains a factor $\langle F\rangle$ or $\langle S_{\rm v}\rangle$). The first nonvanishing correction is therefore the three-insertion cumulant $-\frac{1}{3!}\langle F S_{\rm v}^3\rangle_c = O(\beta^{3/2})$, and the even $O(\beta)$ terms dropped from (25) likewise contribute to the mean only when paired with one odd insertion, so

$$\langle F\rangle_{\rm full} = -\,\langle F\,S_{\rm v}\rangle_{\rm ref} + O(\beta^{3/2}). \tag{65}$$

A.9.1 ONE-INSERTION EVALUATION

Using $\nabla\cdot f^a = \sum_m Z^a_{mm}$ in the vertex (26),

$$\langle f_i^{a_0}(y)\rangle = -\langle F S_{\rm v}\rangle = \frac{P\sqrt{\beta}}{\kappa^2}\int \mathrm{d}\mu(x)\sum_{a=1}^{n}\sum_{m=1}^{d}\Big\langle F\, e^{-\tilde U(x)}\, Z^a_{mm}\Big\rangle. \tag{66}$$

By Lemma 2 and (58), $\langle F\, e^{-\tilde U}\, Z^a_{mm}\rangle = \langle e^{-\tilde U}\rangle\,\langle F Z^a_{mm}\rangle' \to \langle F Z^a_{mm}\rangle'$. Setting $l = m$ in (59) and summing over m (the δ_{im} picks $m = i$), then using (54) and $1/(\kappa^2 + R_x) = \gamma_x/\kappa^2$,

$$\sum_m \langle F Z^a_{mm}\rangle' = \delta^{a_0 a}\left[\partial_{2,i}h(y,x) - \frac{\gamma_x}{2\kappa^2}\,h(y,x)\,\partial_i R_x\right]. \tag{67}$$

The replica sum collapses on a_0. Integrating the first term by parts against $\mathrm{d}\mu = p\,\mathrm{d}x$ (surface term vanishes; $\partial_i p = p\,s_i$),

$$\int \mathrm{d}x\; p(x)\,\partial_{2,i}h(y,x) = -\int \mathrm{d}x\;\big(\partial_i p(x)\big)\,h(y,x) = -\int \mathrm{d}\mu(x)\;h(y,x)\,s_i(x). \tag{68}$$

Thus

$$\boxed{\langle f_i(y)\rangle = -\frac{P\sqrt{\beta}}{\kappa^2}\int \mathrm{d}\mu(x)\;h(y,x)\left[s_i(x) + \frac{\gamma_x}{2\kappa^2}\,\partial_i R_x\right] + O(\beta^{3/2}),} \tag{69}$$

which is Eq. (7) of the main text, with h from (51). The score source has *no* residual γ_x damping; the damping only produces the drift term, through the correlation between the derivative field Z and the local field Y. The first term of (67) is a direct $F \to Z$ contraction, while the second is an $F \to Y \to Z$ contraction; at this order in β the f-correlations are those of the reference measure.

A.10 VARIANCE

Take the *even* two-point observable $F = f_i^{a_0}(y) f_j^{a_0}(y)$ in the linked-cluster expansion (64). Since $S_{\rm ref}$ is even and $S_{\rm v}$ is odd, $F S_{\rm v}^k$ is odd for odd k and its reference average vanishes, while the $k = 0$ term survives:

$$\langle f_i(y) f_j(y)\rangle_{\rm full} = \langle f_i(y) f_j(y)\rangle_{\rm ref} + \tfrac{1}{2}\langle f_i(y) f_j(y)\, S_{\rm v}^2\rangle_{\rm ref,c} + O(\beta^2) = \delta_{ij}\,h(y,y) + O(\beta), \tag{70}$$

using Wick's theorem in the reference measure and $S_{\rm v}^2 = O(\beta)$. Subtracting $\langle f_i(y)\rangle_{\rm full}\langle f_j(y)\rangle_{\rm full} = O(\beta)$ (the mean being $O(\sqrt{\beta})$) leaves

$$\mathrm{Var}[f_i(y)] \;=\; h(y,y) \;=\; R_y \;+\; O(\beta), \tag{71}$$

which is Eq. (8): at this order an ensemble of denoising networks has the predictor variance of the source-free reference measure, independent of β and of the data score.

B NUMERICAL SOLUTION OF THE SELF-CONSISTENCY

Equation (5) is a fixed-point problem for the function $x \mapsto R_x$. We solve on a quadrature sample $X_q = \{x_q\}_{q=1}^{N_q}$ of the population, i.e. with μ replaced by the empirical measure of X_q. For a multiplication operator $\hat m$ with $m_x = P\gamma_x/\kappa^2 = P/(\kappa^2 + R_x)$ one has $h = (K^{-1} + \hat m)^{-1} = K - K(\hat m^{-1} + K)^{-1}K$, and on the sample this becomes

$$h(y,x) = k(y,x) - k(y,X_q)\,B^{-1}\,k(X_q,x), \qquad B = K_{qq} + \mathrm{diag}(\rho_q), \qquad \rho_q = \frac{N_q\,(\kappa^2 + R_{x_q})}{P}, \tag{72}$$

where $K_{qq} = k(X_q, X_q)$. Given a current iterate $R_{x_q}^{(t)}$ we form $B^{(t)}$ by one Cholesky factorization, read off $R_{x_q}^{(t+1)} = \rho_q\,(1 - \rho_q\,[B^{(t)-1}]_{qq})$ (which is $h(x_q, x_q)$ by Eq. (72)), and damp, $R^{(t+1)} \leftarrow (1-\eta)R^{(t)} + \eta R^{(t+1)}$ with $\eta = 0.5$-0.6, until the largest relative change over the sample is below 10^{-3} (10-40 iterations). The constant-ridge solution, with R from the scalar self-consistency $R = \sum_\rho \tilde\lambda_\rho$ on the spectrum $\{\lambda_\rho\}$ of K_{qq}/N_q, is the initial condition; the same scalar problem fixes κ^2 from the prescribed R/κ^2. Integrals against $\mathrm{d}\mu$ are sums over the sample:

$\int \mathrm{d}\mu(x)\, h(y,x)\, g(x) = \frac{1}{P}\, k(y, X_q)\, B^{-1}\, \mathrm{diag}(\kappa^2 + R_{x_q})\, g(X_q)$, exact for the empirical measure. The variance (8) is $R_y = h(y,y)$ at the probes, the score bias is $D(x) = \frac{1}{2}\nabla \log(\kappa^2 + R_x)$ with R_x from Eq. (72) differentiated by automatic differentiation, and for the synthetic distributions the mean (7) uses the analytic score at the sample points. On image data the score is unknown and is removed by parts, $\int \mathrm{d}\mu(x)\, h(y,x)\, s_i(x) = -\int \mathrm{d}\mu(x)\, \partial_{2,i} h(y,x)$, evaluated as $-\frac{1}{N_q}\sum_q \partial_{2,i} h(y, x_q)$ with the kernel gradient sums $\sum_q \partial_{x_q} k(\cdot, x_q)$ in closed form. For the fibre-averaged ridge (50) on synthetic data with $E = \mathrm{span}(e_1)$.

A rank-truncated eigenbasis solve is *not* adequate for the heteroscedastic ridge: the operator $\hat{\gamma}$ couples the retained modes to the discarded tail, and truncating it distorts the score bias.

C THE $M \to \infty$ POSTERIOR OF A FIXED DATASET, AND THE FINITE-M CHECK

The theory is derived for the noise-averaged loss, the $M \to \infty$ limit of Eq. (2), while a trained network sees M fixed noise draws per clean point. This appendix gives the closed form of the $M \to \infty$ posterior for a fixed dataset and a fixed hidden layer, the finite-M posterior it is compared with, and the comparison behind the value of M used in Sec. 6.

Posterior of the readout for a fixed hidden layer. Write the network of Appendix D as $f_i(x) = \sum_{j=1}^{N_\mathrm{w}} a_{ji}\, \phi_j(x)$ with features $\phi_j(x) = \sqrt{c/N_\mathrm{w}}\, \mathrm{erf}\big(W_j^{(1)} \cdot x/\sqrt{d}\big)$, so that every weight, a_{ji} and $W_{jk}^{(1)}$, has a standard normal prior, the prior that weight decay of unit strength at unit temperature imposes. For a fixed hidden layer $W^{(1)}$ the network is linear in the readout a, and its prior over functions is the Gaussian process with the feature kernel

$$\hat{K}(x,y) = \sum_{j=1}^{N_\mathrm{w}} \phi_j(x)\, \phi_j(y), \tag{73}$$

whose expectation over $W^{(1)}$ is the NNGP kernel K and which converges to K as $N_\mathrm{w} \to \infty$. Under the loss (2), with the $N = MP$ noisy inputs $x_\beta^{(a)}$ and targets $\epsilon^{(a)}$ arranged in the matrices $\Phi_{aj} = \phi_j(x_\beta^{(a)})$ and $\mathcal{E}_{ai} = \epsilon_i^{(a)}$, the posterior of a is Gaussian and independent across output indices i:

$$\begin{gathered} \Lambda = I + \frac{\Phi^\top \Phi}{M\kappa^2}, \quad a^* = \Lambda^{-1}\, \frac{\Phi^\top \mathcal{E}}{M\kappa^2}, \\ \mathbb{E}\big[f_i(y) \,\big|\, W^{(1)}\big] = \phi(y)^\top a^*_{\cdot i}, \quad \mathrm{Var}\,\big[f_i(y) \,\big|\, W^{(1)}\big] = \phi(y)^\top \Lambda^{-1} \phi(y). \end{gathered} \tag{74}$$

This is Eq. (3) with $\hat{K}$ in place of K. It is the readout move of the Gibbs sampler of Sec. 6, and the conditional mean and variance in Eq. (74), evaluated along the hidden-layer chain and averaged over it, are the Rao-Blackwellised estimators of a member's mean and variance.

The $M \to \infty$ limit. As $M \to \infty$ the loss becomes the noise average $\frac{1}{2\kappa^2}\sum_\mu \mathbb{E}_\epsilon \|\epsilon - f(x_0^{(\mu)} + \sqrt{\beta}\,\epsilon)\|^2$. Expanding $f(x_0^{(\mu)} + \sqrt{\beta}\,\epsilon) = f(x_0^{(\mu)}) + \sqrt{\beta}\,(\epsilon\cdot\nabla) f(x_0^{(\mu)}) + O(\beta)$ and averaging,

$$\mathbb{E}_\epsilon \big\|\epsilon - f(x_0^{(\mu)} + \sqrt{\beta}\,\epsilon)\big\|^2 = d + \big|f(x_0^{(\mu)})\big|^2 - 2\sqrt{\beta}\,\nabla\cdot f(x_0^{(\mu)}) + O(\beta)\,:$$

the term linear in ϵ averages to zero and $\mathbb{E}[\epsilon_i \epsilon_k] = \delta_{ik}$ turns the cross term into the divergence, the integration by parts of Sec. 5. The negative log posterior over the functions f in the span of the features is therefore, up to a constant,

$$\frac{1}{2}\sum_{i=1}^{d} \big\langle f_i, \hat{K}^{-1} f_i \big\rangle + \frac{1}{2\kappa^2}\sum_{\mu=1}^{P} \big|f(x_0^{(\mu)})\big|^2 - \frac{\sqrt{\beta}}{\kappa^2}\sum_{\mu=1}^{P} \nabla\cdot f(x_0^{(\mu)}) + O(\beta), \tag{75}$$

the quadratic form of Gaussian-process regression at the P clean points with observation noise κ^2, plus a linear source; the $O(\beta)$ terms move the mean at $O(\beta^{3/2})$ and the covariance at $O(\beta)$ and are dropped. A Gaussian with log-density $-\frac{1}{2}\langle f, Af\rangle + \langle J, f\rangle$ has covariance A^{-1} and mean $A^{-1}J$. Here A^{-1} is the posterior covariance given the clean points,

$$h_P(y,x) = \hat{K}(y,x) - \hat{K}(y,X)\big(\hat{\mathbf{K}} + \kappa^2 I_P\big)^{-1} \hat{K}(X,x), \qquad \hat{\mathbf{K}}_{\mu\nu} = \hat{K}\big(x_0^{(\mu)}, x_0^{(\nu)}\big),$$

and J is the divergence source, so that, to this order, the posterior mean and variance at a probe y are

$$\begin{aligned}\bar{f}_i(y) &= \frac{\sqrt{\beta}}{\kappa^2}\sum_{\mu=1}^{P}\partial_{2,i}\,h_P\big(y,x_0^{(\mu)}\big)\\ &= \frac{\sqrt{\beta}}{\kappa^2}\Big[\sum_{\mu}\partial_{2,i}\hat{K}\big(y,x_0^{(\mu)}\big)-\hat{K}(y,X)\big(\hat{\mathbf{K}}+\kappa^2 I_P\big)^{-1}q_i\Big],\qquad \mathrm{Var}\left[f_i(y)\right]=h_P(y,y),\end{aligned}\tag{76}$$

with $(q_i)_\nu=\sum_\mu \partial_{2,i}\hat{K}(x_0^{(\nu)},x_0^{(\mu)})$ and $\partial_{2,i}$ the derivative in the i-th coordinate of the second argument. Both are $P\times P$ linear algebra on the clean points and involve no noise draws. Equations (7) and (8) of the main text are the theory's predictions for the average of these two quantities over draws of the clean dataset; with the infinite-width kernel K in place of $\hat{K}$ they are the $M\to\infty$ posterior of the NNGP itself.

At finite M the same expansion applied to Eq. (3) gives, at zeroth order in $\sqrt{\beta}$, the regression of the mean noise of each clean point,

$$\hat{K}(y,X)\big(\hat{\mathbf{K}}+\kappa^2 I_P\big)^{-1}\bar{\epsilon}_i,\qquad \bar{\epsilon}^{(\mu)}=\frac{1}{M}\sum_{m=1}^{M}\epsilon^{(\mu,m)},\tag{77}$$

a term odd in the noise draws and of order $M^{-1/2}$, which averages to zero over the draws and is absent from the $M\to\infty$ limit; at the small β of the experiments it dominates the $O(\sqrt{\beta})$ signal for any practical M. We therefore draw the M noise vectors of each clean point as $M/2$ antithetic pairs $\pm\epsilon$, which sets $\bar{\epsilon}^{(\mu)}=0$ exactly, leaves Eq. (2) and its $M\to\infty$ limit unchanged, and removes the term. What remains of the finite-M error is the sampling fluctuation of the noise second moments $\frac{1}{M}\sum_m \epsilon^{(\mu,m)}\epsilon^{(\mu,m)\top}$ about I_d.

D Experimental details

The protocol details the protocol section of 6 states experiments done in $M=64$ were checked against the $M\to\infty$ limit. These details are brought forward here. For a fixed training set and a fixed hidden layer the network is linear in its readout and the posterior is Gaussian; as $M\to\infty$ the average over the noise turns the target ϵ into the divergence of the network output at the clean points, so that to leading order in $\sqrt{\beta}$ the posterior is Gaussian-process regression at the P clean points with observation noise κ^2 and a source proportional to $\sqrt{\beta}\,\nabla\cdot f$, whose mean and variance at the probes are closed forms in the $P\times P$ kernel matrix of the clean points and involve no noise draws at all (Appendix C). We evaluated this closed form for the same training sets and the same initial hidden layers as the trained ensembles, against both the trained members and their finite-M readout posterior at the initial hidden layer, and found close agreement, so the finite-M ensembles below stand in for the $M\to\infty$ object the theory describes. Each member is trained by the weight-decay Langevin dynamics of Sec. 3 at unit temperature, run to equilibrium, in the blocked form that samples the same posterior but mixes faster: Langevin steps on the hidden-layer weights alternate with exact draws of the readout weights from their Gaussian conditional given the hidden layer, a Gibbs scheme in which both moves leave the posterior invariant. The conditional is available because for a fixed hidden layer the network is linear in its readout, so the readout posterior is Bayesian linear regression on the hidden features in closed form (Appendix C, Eq. (74)). Because the posterior mean is of order $\sqrt{\beta}$ while a single posterior sample fluctuates by the much larger $\sqrt{R_x}$, the member's mean and variance at the probes are not read off the sampled outputs but off the readout's conditional posterior along the chain, we use the Rao-Blackwellised estimators of Gibbs sampling (Casella & Robert, 1996): the time average over the equilibrium phase of the conditional mean is the member's mean, and the time average of the conditional variance plus the variance of the conditional mean over the chain is its posterior variance.

Heteroscedastic ridge of real images. To apply our theory to real datasets, where we do not have a clear definition of the base manifold and the fiber, we note that the diagonal covariance R_x, typically changes on scales of the kernel. Since the score $s(x)$ and x are related by low-variance directions, we approximate $R_x=R_{s(x)}$. Given the latter equality, we may replace any instance of R_s in our theory

by R_x and obtain fiber-agnostic equations. We use this fiber-agnostic version to compute our theory for MNIST and CIFAR-10.

Networks and prior. One-hidden-layer fully connected erf networks without biases, $f_i(x) = \sum_{j=1}^{N_w} W_{ij}^{(2)} \operatorname{erf}\big(W_j^{(1)} \cdot x/\sqrt{d}\big)$, in the NNGP parameterization, $W_{jk}^{(1)} \sim \mathcal{N}(0,1)$ and $W_{ij}^{(2)} \sim \mathcal{N}(0, c/N_w)$, with c fixed so that the NNGP kernel has unit trace, $\mathbb{E}_x K(x,x) = 1$.

Image data. MNIST (60 000 training images, 28×28) is area-downsampled by 2×2 block averaging to 14×14, $d = 196$; CIFAR-10 (50 000 training images, 32×32 RGB) is area-downsampled to 16×16 and converted to grey scale with the ITU-R BT.601 luminance weights $(0.299, 0.587, 0.114)$, $d = 256$; pixel values are rescaled to $[-1, 1]$. The population is the full training set (pooled) or one class (5421–6742 MNIST digits, 5000 CIFAR-10 images); each dataset draw is $P = 256$ images drawn with replacement from it; the 17 probes lie on the geodesic of Appendix E.3 built from the same population, and the mean and variance are projected on its local tangent.

Synthetic distributions. All densities are of the form $p(x) = q(x_1) \prod_{j \geq 2} \mathcal{N}(x_j; 0, \sigma_j^2)$ in $d = 64$. In the isotropic embedding $\sigma_j = 1$ and the profiles q are: the standard Gaussian; Student-t with $\nu = 3$ degrees of freedom and unit scale; the two-component mixture $\frac{1}{2}\mathcal{N}(-2,1) + \frac{1}{2}\mathcal{N}(2,1)$; and the heavy-tailed profile $\frac{1}{3}\sum_k \mathcal{N}(0, \sigma_k^2)$ with $\sigma_k \in \{1, 3, 10\}$, a zero-mean scale mixture with three length scales. The probes are $y = y_1 e_1$ on a grid of 41 points, $|y_1| \leq L$ with $L = 3, 4, 5, 7$ respectively. In the anisotropic embedding $\sigma_j = \sigma_\perp = 0.15$ and every profile is scaled to variance $\sigma_1^2 = 16$: $q = \mathcal{N}(0, \sigma_1^2)$; Student-$t$ with $\nu = 3$ and scale $\sigma_1\sqrt{(\nu-2)/\nu}$; and the three-component mixture with weights $0.30/0.45/0.25$, means $-5.16/0.40/4.57$ and standard deviations $0.87/2.18/1.27$ (a narrow left lobe, a broad centre and a medium right lobe). Its probes are $y = y_1 e_1 + \sigma_\perp \sqrt{d-1}\, e_2$, on 241 points with $|y_1| \leq 3\sigma_1$, of which $|y_1| \leq 1.5\sigma_1$ is shown. In both embeddings the network output is projected onto e_1. Fig. 4 repeats the experiment at $d = 16$.

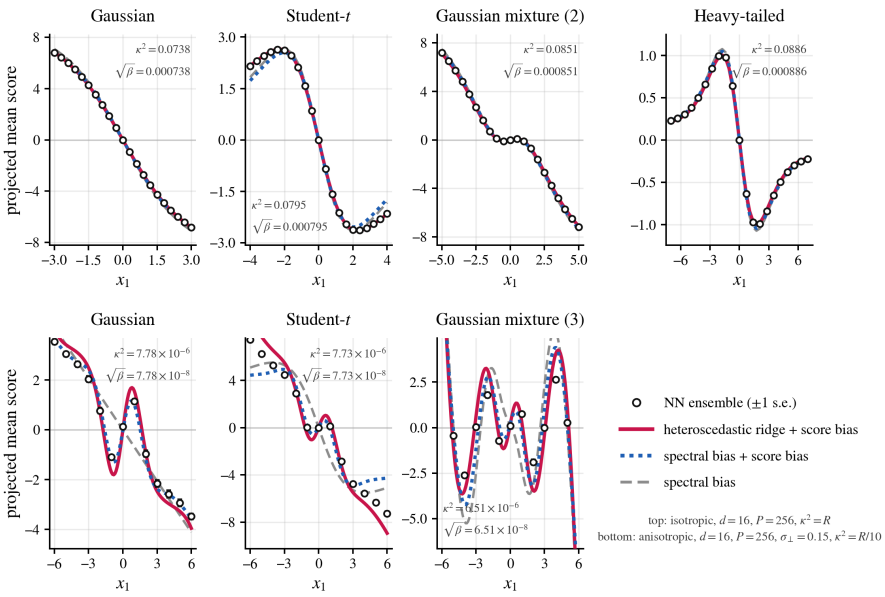

Figure 4: **Synthetic distributions: the theory predicts the ensemble mean.** Projected mean output of a denoising ensemble against the non-Gaussian coordinate x_1. *Top row (isotropic embedding, $d = 16$, $P = 256$, $\kappa^2 = R$):* Gaussian, Student-t, two-component Gaussian mixture, and heavy-tailed profiles, with the remaining coordinates standard Gaussian. *Bottom row (anisotropic embedding, $d = 16$, $P = 256$, $\sigma_\perp = 0.15$, $\kappa^2 = R/10$):* Gaussian, Student-t, and three-component Gaussian mixture profiles, with the remaining coordinates Gaussian of variance $\sigma_\perp^2$. *Circles*: ensemble mean over 4096 dataset realizations, one width-5000 network per realization with its readout posterior evaluated in closed form in the $M \to \infty$ limit (Appendix C), ± 1 s.e. *Solid*: Eq. (7) with the heteroscedastic ridge (5) in its pointwise form, $R_{s(x)} = R_x = h(x, x)$. *Dotted*: Eq. (7) with the constant ridge h_0 of Eq. (4). *Dashed*: the score term of Eq. (7) with the constant ridge, $-\frac{P\sqrt{\beta}}{\kappa^2} \int \mathrm{d}\mu \, h_0(y, x) \, s(x)$, i.e. kernel ridge regression on the true score with the same kernel, ridge and prefactor; it differs from the dotted curve only by the score bias D. *Dash-dotted (bottom row)*: Eq. (7) with the fibre-averaged ridge of Eq. (5), $E = \mathrm{span}(e_1)$. Variance panels are in Appendix E.1 (Fig. 5); the same experiment at $d = 64$ is in Fig. 1.

E ADDITIONAL FIGURES

E.1 VARIANCE PANELS FOR THE SYNTHETIC DISTRIBUTIONS

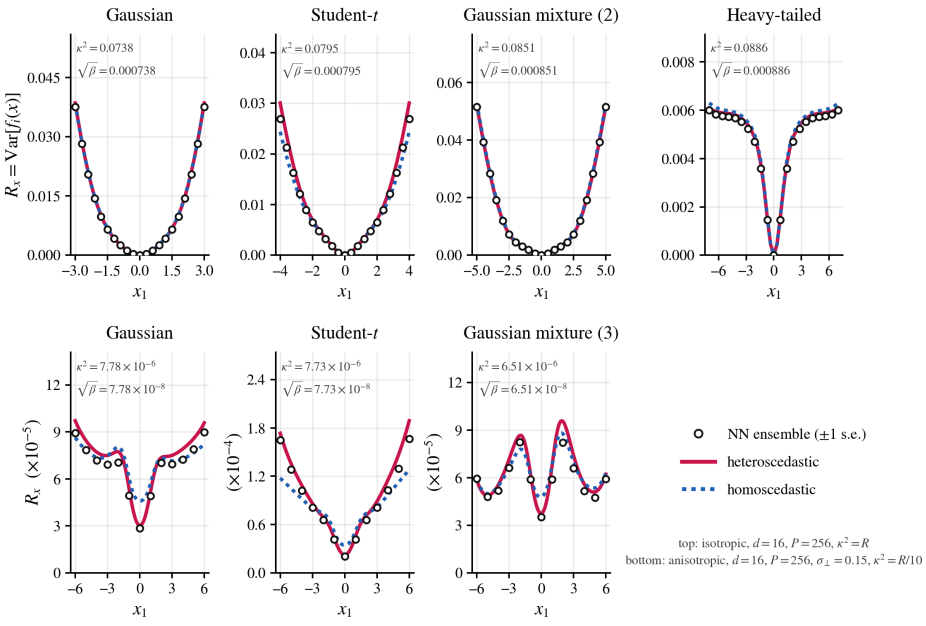

Figure 5: **Ensemble variance on the synthetic distributions.** Across-realization variance of the ensemble output along the probe line, $\mathrm{Var}[f_i(x)]$, for the seven distributions and settings of Fig. 4. *Circles*: ensemble, ± 1 s.e.; *solid*: heteroscedastic $R_x = h(x, x)$; *dotted*: homoscedastic. In the isotropic embedding, the two coincide; in the anisotropic embedding, only the heteroscedastic ridge reproduces the dip of the Gaussian at the center, the tails of the Student-t, and the peaks of the mixture.

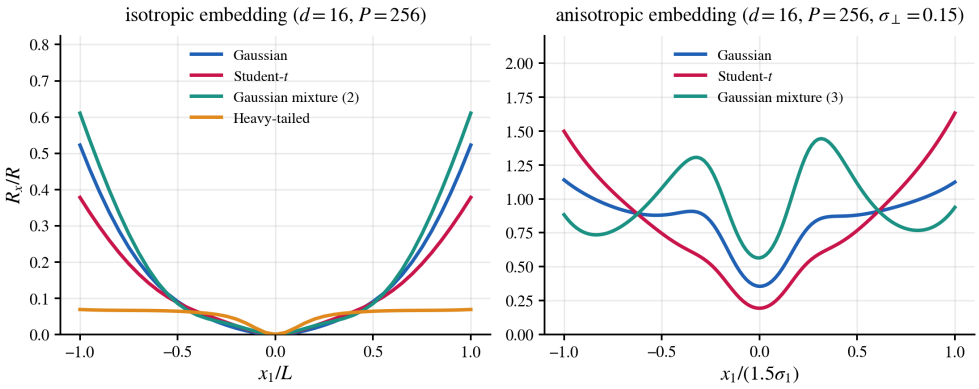

Figure 6: $d = 16$**: Variation of R_x along the first coordinate axes.** posterior variance of the heteroscedastic ridge, normalized by $R = \int \mathrm{d}\mu\, R_x$, for the seven distributions and settings of Fig. 4 ($d = 16$, $P = 256$ in both embeddings), along its probe lines: isotropic embedding (left; $y = y_1 e_1$, $|y_1| \le L$ with $L = 3, 4, 5, 7$ for the Gaussian, Student-t, mixture (2) and heavy-tailed respectively) and anisotropic embedding (right; $y = y_1 e_1 + \sigma_\perp \sqrt{d-1}\, e_2$, $|y_1| \le 1.5\sigma_1$). In the isotropic embedding the probe line passes through the origin, where the bias-free kernel vanishes, so R_x varies along it through the input norm even though its spread over the data is small, less so than at $d = 64$; in the anisotropic embedding R_x varies by $O(1)$ across the data, with a dip at the center of the Gaussian and of the Student-t and oscillations that follow the components of the mixture.

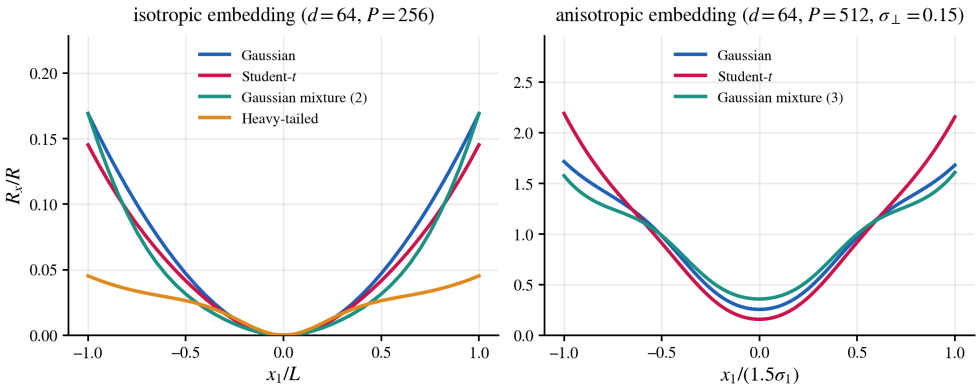

Figure 7: $d = 64$ **Variation of R_x along the first coordinate axes.** Pointwise posterior variance of the heteroscedastic ridge, normalized by $R = \int \mathrm{d}\mu\, R_x$, for the seven distributions and settings of Fig. 1, along its probe lines: isotropic embedding (left; $y = y_1 e_1$, $|y_1| \le L$ as in Fig. 6) and anisotropic embedding (right; $y = y_1 e_1 + \sigma_\perp \sqrt{d-1}\, e_2$, $|y_1| \le 1.5\sigma_1$). In the isotropic embedding the probe line passes through the origin, where the bias-free kernel vanishes, so R_x varies along it through the input norm even though its spread over the data is small; in the anisotropic embedding R_x varies by $O(1)$ across the data, with a dip at the center of all three profiles.

E.2 Full per-class breakdown on MNIST and CIFAR-10

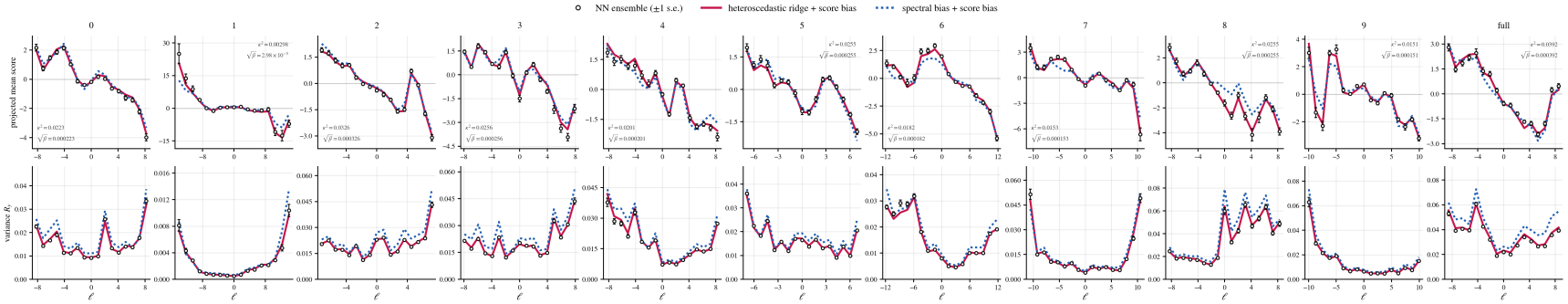

Figure 8: MNIST: projected mean score (top) and variance (bottom) vs. geodesic arc length ℓ, per digit class and pooled ("full"). *Circles*: ensemble; *solid*: theory with heteroscedastic ridge; *dotted*: spectral bias. κ^2 and $\sqrt{\beta}$ are printed in each panel.

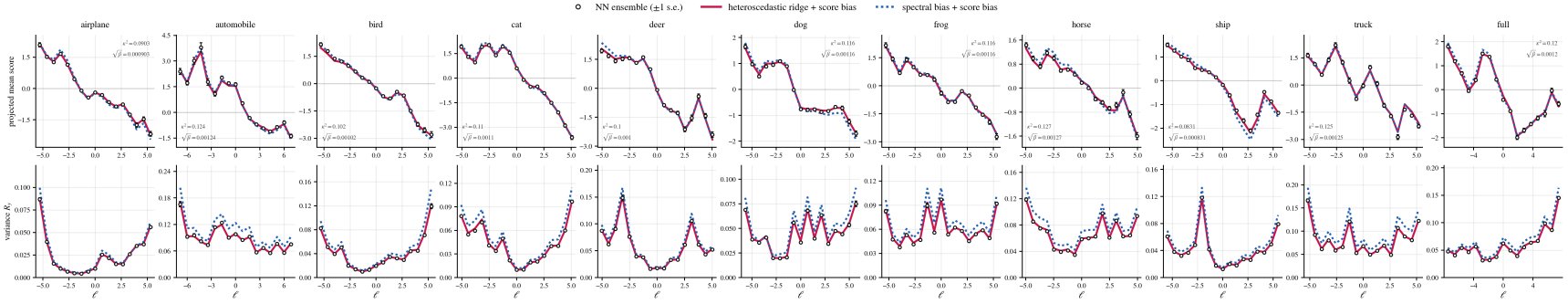

Figure 9: CIFAR-10: as the previous figure, per class and pooled ("full"). Heteroscedastic ridge and spectral bias nearly coincide on this dataset.

E.3 GEODESIC PROBE CONSTRUCTION

The scalar coordinate ℓ used in Sec. 6 for MNIST and CIFAR-10 is built directly from the image population, not from the theory, as follows. We write it as ℓ, an arc length, rather than s, which is reserved for the score $s(x) = \nabla \log p(x)$ and (in Appendix A) for the dominant coordinates.

1. **Whitened PCA subspace.** Project the image population onto its top 10 principal components and whiten (divide each coordinate by the square root of its eigenvalue).

2. **Local density.** For each point, estimate a local density radius as the distance to its 12th nearest neighbour in this whitened subspace (small radius = dense region).

3. **Density-restricted kNN graph.** Build a $k = 12$ nearest-neighbour graph in the whitened subspace, retain only the vertices whose density radius falls below a quantile threshold (starting at the 70th percentile and relaxing to 80th/90th/100th if this leaves too few points or too small a connected component), and keep its largest connected component.

4. **Endpoints via double-sweep search.** Within this component, start from the single densest point and run Dijkstra's algorithm to find the farthest reachable vertex; run Dijkstra again from that vertex to find a second farthest vertex. This standard graph-diameter heuristic gives two near-maximally-separated, well-connected endpoints.

5. **Geodesic.** The shortest path between the two endpoints in the density-restricted graph is the geodesic: a chain of real dataset images connected by short, dense-region hops, which by construction cannot take a low-density shortcut through a region the data manifold does not populate.

6. **Probes and coordinate.** Parametrize the path by cumulative Euclidean arc length (in whitened PCA space) and place 17 probes evenly spaced in this arc length; ℓ is this arc length re-centred on the path's midpoint. Each probe's pixel-space image is the linear interpolation of the two real images bounding its graph edge; the resulting probe sequences are shown in Fig. 10.

7. **Tangent direction.** At each probe, a local tangent direction in pixel space is obtained from a centred finite difference of the PCA scores along the path, mapped back through the (unwhitened) PCA basis and normalized; the network's d-dimensional mean and variance are projected onto this direction to obtain the scalar "projected mean" and "total variance" plotted against ℓ in Fig. 2 and Appendix E.2.

Figure 10 shows the probe images themselves for the two pooled runs of Fig. 2. They make concrete what the coordinate ℓ parametrizes: a path of real, in-distribution images that migrates between classes ($2 \to 6 \to 0$ on MNIST; deer/airplane/ship $\to$ frog on CIFAR-10) without leaving the density-supported region, rather than a straight line in pixel space, which would pass through images the data manifold does not populate.

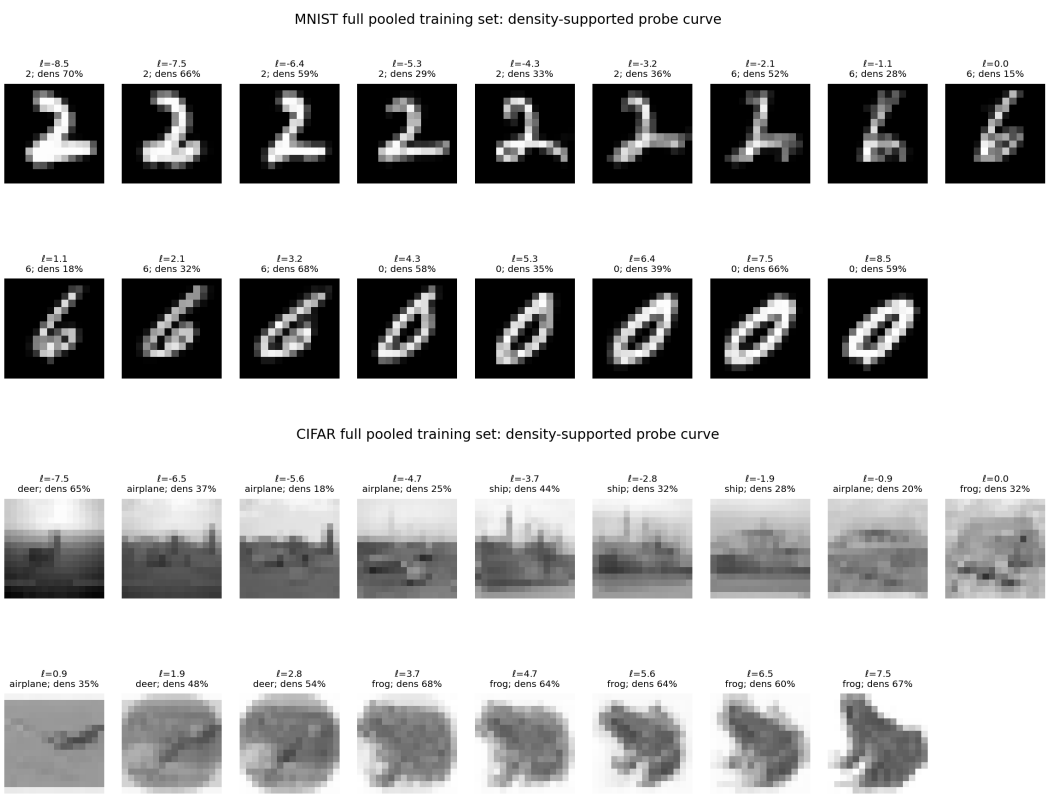

Figure 10: The 17 probe images along the pooled MNIST (top) and CIFAR-10 (bottom) geodesics, in order of arc length ℓ. Each panel is annotated with its coordinate ℓ, the class of the nearest real image, and the local density statistic of the underlying graph vertex. These are the probe locations at which the mean and variance of Fig. 2 are evaluated.