WHERE THE NOISE HIDES: BLIND DENOISING FROM THE SCORE JACOBIAN'S SPECTRUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Blind denoising, i.e., recovering a clean signal when the noise level is unknown, is a long-standing challenge in signal processing and machine learning. Differentiating Tweedie's formula ties the Jacobian of the score function to the posterior covariance of the clean signal, and hence to the noise variance. However, conventional methods built on it assume the true score function, ignoring the approximation error that neural network training introduces. In this work, we target the noise estimation and denoising problem by analysing the effect of that error on the spectrum. We first show that when the clean signal is supported on a low-dimensional manifold, the true score Jacobian has a cluster of eigenvalues at $-1/\sigma^2$ displaced only slightly by signal manifold curvature, making the smallest eigenvalue an optimal noise estimator with the exact score. Using neural tangent kernel theory, we illustrate that score matching perturbs the learned score Jacobian with zero mean initialization randomness, even for a network that interpolates its training samples exactly. In an idealized case the resulting distribution of the Jacobian noise eigenvalues is symmetric about $-1/\sigma^2$ for any covariance of the perturbation, implying that the centre of the eigenvalue cluster is robust to learning perturbation while the smallest eigenvalue drifts to underestimate noise. A practical algorithm, JAcobian-spectrum Noise Estimation (JANE), is hence proposed to estimate σ^2 from the leftmost mode of the eigenvalue density, with no knowledge of the noise level, the signal dimension, or a clean reference. Across six dataset configurations and different noise levels, JANE attains 3.3% mean relative error in σ^2 estimation against 9.4% for Noise2Score and 28.1% for UNSURE, and matches an oracle step-size denoising to within 0.05 dB PSNR in all dataset and noise level settings.

1 INTRODUCTION

Recovering a clean signal from corrupted observations is a central problem across signal processing, imaging, sensing, communications, and machine learning, and the statistical properties of the noise are often partially or entirely unknown. For the additive model

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{1}$$

with $\mathbf{x} \in \mathbb{R}^D$ the signal and σ^2 the noise variance, Tweedie's formula (Robbins, 1992; Efron, 2011) expresses the minimum mean-square error estimator as $\hat{\mathbf{x}} = \mathbf{y} + \sigma^2 \nabla_{\mathbf{y}} \log p(\mathbf{y})$, decomposing the denoising problem into the noise level σ^2 and the estimation of score function $\mathbf{s}(\mathbf{y}) \triangleq \nabla_{\mathbf{y}} \log p(\mathbf{y})$. The score function estimation has been well addressed by using score matching to learn $\mathbf{s}(\mathbf{y})$ from noisy data alone (Hyvärinen, 2005; Vincent, 2011; Song et al., 2020). The noise level, although required by many classical denoisers (Dabov et al., 2007) including SURE-based training (Stein, 1981) and Tweedie-type denoising as an input, is rarely known a priori and clean references are scarce in many applications. Therefore, blind denoising is required in such use cases and noise level has to be extracted from the noisy observations with little prior knowledge.

Some recent works have estimated σ^2 from a learned score network based on Tweedie's formula, aggregating statistics over all dimensions while mixing signal and noise contributions (Tachella et al., 2025; Kim et al., 2022). Gradient expression of Tweedie's formula, i.e., Miyasawa identity (Raphan & Simoncelli, 2011; Shoushtari et al., 2026), connects the *full* score Jacobian $\nabla^2_{\mathbf{y}} \log p(\mathbf{y})$ to the posterior covariance $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$, and has started to be exploited for uncertainty quantification, Hessian

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.

learning and out-of-distribution detection (Manor & Michaeli, 2024; Meng et al., 2021; Shoushtari et al., 2026). However, these analyses operate on the *true* score that is assumed to be learnt perfectly. In practice the score function is a neural network trained by score matching loss, which inevitably perturbs the Jacobian. It has not been investigated yet how that perturbation affects the Jacobian spectrum, and how an estimator should be designed around it.

In this work, we bridge this gap by showing that the eigenvalue spectrum of the score Jacobian directly encodes the noise variance, and providing analysis and algorithm for addressing the network learning perturbation introduced in score matching. Our contributions are:

1. **Dependence of score Jacobian spectrum on noise and manifold curvature.** The Miyasawa identity gives the relationship between posterior covariance $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$ and score Jacobian. Based on manifold curvature analysis, we derived the dependence of score Jacobian eigenvalues on noise and manifold reach ρ, demonstrating the feasibility of unconditional noise upperbound estimation using the minimum score Jacobian eigenvalue and the relative estimation error bound that scales in the order of $(\sigma/\rho)^2$. The analysis implies an exact analytical noise estimator from clustered eigenvalues at $-1/\sigma^2$ when signals are restricted to a low-dimensional linear subspace.
2. **Analysis and modeling of perturbation introduced by learning.** We demonstrate in experiments and theoretical analysis that the analytical score Jacobian properties can be violated in self-supervised score matching regardless of network capacity, training length, and dimensions. Modeling the training perturbation with neural tangent kernel (NTK) theory (Yang, 2020; Yang & Littwin, 2021), we show that the noise eigenvalue law is *exactly* symmetric about $-1/\sigma^2$ for any covariance of the perturbation when ideally the deterministic learning error vanishes. Such a property implies the noise eigenvalue cluster center being invariant to the perturbation while exposing the sensitivity of analytically exact estimator that relies on minimum Jacobian eigenvalue.
3. **JAcobian-spectrum Noise Estimation (JANE) with oracle-level accuracy and denoising performance.** Based on the analysis, we propose the JANE algorithm that blindly extracts noise variance from the leftmost mode of the pooled eigenvalue density, requiring no prior knowledge of noise / signal or reference. Across six dataset configurations and different noise levels, it reduces mean relative error in σ^2 by $2.9\times$ over Noise2Score and $8.6\times$ over UNSURE, and the resulting denoiser matches an oracle step-size denoiser to within 0.05 dB.

2 Background

2.1 Related Work

Blind noise estimation and denoising. Score-based estimators read σ^2 from the score norm or from finite-difference score differences (Tachella et al., 2025; Kim et al., 2022); covariance-based ones read it from wavelet coefficients, the low-variance tail of local PCA spectra, or the trailing eigenvalues of a sample covariance (Donoho & Johnstone, 1994; Liu et al., 2013; Chen et al., 2015; Wax & Kailath, 1985; Kritchman & Nadler, 2009). Learned alternatives predict a noise level with a dedicated subnetwork (Guo et al., 2019; Yue et al., 2019), while SURE-based unsupervised training takes σ as an input (Soltanayev & Chun, 2018; Metzler et al., 2018), so it does not address the blind setting. Score-based estimators aggregate over all dimensions, so signal and noise contributions are mixed. The covariance-based estimators do isolate a noise subspace, but a covariance formed over a dataset has a single eigenbasis and so a single signal subspace, whereas the score Jacobian is a distinct matrix at each observation (Section 3). Besides, some self-supervised denoisers exploit statistical structure in the noise itself and never model the noise level explicitly (Lehtinen et al., 2018; Krull et al., 2019; Batson & Royer, 2019). This work mainly concerns Tweedie-type denoising, where the noise statistics enters the estimator directly.

Signal-subspace analyses of the Jacobian spectrum. The posterior covariance exposed by the Miyasawa identity has been used for uncertainty quantification, for learning the score Hessian through higher-order Tweedie formulas, and for out-of-distribution detection from the leading eigenvalues (Manor & Michaeli, 2024; Meng et al., 2021; Shoushtari et al., 2026). Mohan et al. (2020) and Kadkhodaie et al. (2024) analyze the Jacobian of a *denoiser* and observe a signal/noise eigenvalue split for signals on a low-dimensional manifold, while Stanczuk et al. (2024) and Ventura

et al. (2025) estimate the intrinsic dimension and track the spectral gap across the diffusion process. They characterize the signal subspace but do not model the randomness in the Jacobian. Kadkhodaie et al. (2024) do report unexpected eigenvalues outside that subspace, which they attribute to the network's inductive bias. We identify a further contribution of neural network randomness that longer training and larger capacity do not remove, and build the noise estimator on that account (Section 4).

2.2 Score Function and Tweedie's Formula

We assume the signal is contaminated by Gaussian noise

$$\mathbf{y} = \mathbf{x} + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}), \tag{2}$$

with $\mathbf{C}$ a known positive-definite covariance, and specialize to the isotropic case $\mathbf{C} = \sigma^2\mathbf{I}$ of (1) throughout; Appendix A treats the anisotropic case. The multivariate Tweedie's formula (Robbins, 1992; Efron, 2011; Raphan & Simoncelli, 2011) gives the MMSE estimator:

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = \mathbf{y} + \mathbf{C}\,\nabla_{\mathbf{y}} \log p(\mathbf{y}) = \mathbf{y} + \mathbf{C}\,\mathbf{s}(\mathbf{y}). \tag{3}$$

Score matching (Hyvärinen, 2005; Vincent, 2011; Song et al., 2020) provides practical methods for learning $\mathbf{s}(\mathbf{y})$ from noisy samples, which is also used for generative models (Song & Ermon, 2019; Song et al., 2021) and diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015). The *score Jacobian* is the Hessian of the log-marginal:

$$\mathbf{H}_{\mathbf{y}} \triangleq \nabla_{\mathbf{y}}\mathbf{s}(\mathbf{y}) = \nabla^2_{\mathbf{y}} \log p(\mathbf{y}). \tag{4}$$

The generalized Miyasawa identity (Raphan & Simoncelli, 2011; Shoushtari et al., 2026) connects $\mathbf{H}_{\mathbf{y}}$ to the posterior covariance, revealing that the curvature of the log-density encodes uncertainty in the posterior distribution

$$\mathrm{Cov}[\mathbf{x}|\mathbf{y}] = \mathbf{C} + \mathbf{C}\,\mathbf{H}_{\mathbf{y}}\,\mathbf{C} = \mathbf{C}(\mathbf{I} + \mathbf{H}_{\mathbf{y}}\,\mathbf{C})\,. \tag{5}$$

For the isotropic case $\mathbf{C} = \sigma^2\mathbf{I}$, this simplifies to $\mathrm{Cov}[\mathbf{x}|\mathbf{y}] = \sigma^2\mathbf{I} + \sigma^4\mathbf{H}_{\mathbf{y}}$; see also Kadkhodaie et al. (2024) for a recent derivation of this identity in denoiser form.

3 Noise Estimation from Score Jacobian Eigenvalues

We derive bounds on the eigenvalues of the score Jacobian $\mathbf{H}_{\mathbf{y}}$ directly from the generalized Miyasawa identity (5), and show how these bounds enable noise estimation. For isotropic noise $\mathbf{C} = \sigma^2\mathbf{I}$ the Miyasawa identity (5) reads $\mathrm{Cov}[\mathbf{x}|\mathbf{y}] = \sigma^2\mathbf{I} + \sigma^4\mathbf{H}_{\mathbf{y}}$, so the two matrices share eigenvectors and their spectra are in exact correspondence:

$$\lambda_i = \frac{\nu_i - \sigma^2}{\sigma^4}, \qquad \nu_i = \sigma^2 + \sigma^4\lambda_i, \qquad i = 1, \ldots, D, \tag{6}$$

where $\nu_1 \geq \cdots \geq \nu_D \geq 0$ are the eigenvalues of $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$. Since (6) holds for any observation, with the same noise distribution eigenvalues at different data samples $\mathbf{y}$ can be pooled and analyzed together without assuming a common subspace. Therefore $\lambda_i \geq -1/\sigma^2$ and

$$\sigma^2 \;\leq\; -1/\lambda_{\min}(\mathbf{H}_{\mathbf{y}}), \qquad \text{whenever } \lambda_{\min}(\mathbf{H}_{\mathbf{y}}) < 0. \tag{7}$$

This inequality actually provides an upper bound estimation of the noise variance, without requiring any structural assumption on the signal. The general anisotropic case is treated in Appendix A, where the bound only applies to the smallest noise variance in all dimensions. Natural signals typically have intrinsic dimension lower than the ambient dimension (Mohan et al., 2020). This allows us to quantify the estimation accuracy in most common scenarios as follows.

Proposition 1. *Let the clean signal be supported on a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ of reach ρ, with $d < D$, under the regularity conditions (B1)–(B3) and the proximity condition (T) of Appendix B, which simplifies the problem by assuming noise locality. Even with low-dimensional signals, the $D - d$ noise-dimension eigenvalues of $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$ do not vanish due to manifold curvature. However, they are bounded as $\nu_i \leq \varepsilon$ with $\varepsilon = \mathcal{O}(\sigma^4/\rho^2)$ (Appendix B). By (6) all $D - d$ trailing Jacobian eigenvalues then lie in $[-1/\sigma^2,\ -1/\sigma^2 + \eta]$, an interval of width $\eta := \varepsilon/\sigma^4 = \mathcal{O}(1/\rho^2)$, and*

$$\sigma^2 \;\leq\; -1/\lambda_{\min}(\mathbf{H}_{\mathbf{y}}) \;\leq\; \frac{\sigma^2}{1 - \varepsilon/\sigma^2} \;=\; \sigma^2\Big(1 + \frac{\varepsilon/\sigma^2}{1 - \varepsilon/\sigma^2}\Big) \;=\; \sigma^2\big(1 + \mathcal{O}((\sigma/\rho)^2)\big), \tag{8}$$

so the bound (7) *overestimates* σ^2 *by a relative amount that is at most second order in* σ/ρ*. The upper bound requires* $\varepsilon < \sigma^2$*, and is tight only when* $\varepsilon \ll \sigma^2$*. In particular, if the clean signal lies in a linear subspace which has infinite reach, so* $\rho = +\infty$ *and* $\varepsilon = 0$*, and the gap closes entirely:*

$$\nu_i = 0, \qquad \lambda_i = -1/\sigma^2, \qquad i = d+1, \ldots, D, \tag{9}$$

(7) *is* exact *as a degenerate case (Mohan et al., 2020; Kadkhodaie et al., 2024; Ventura et al., 2025).*

Although it is established in Proposition 1 that the minimum eigenvalue of the score Jacobian is always an over-estimator of the noise variance with up to second-order gap and is therefore the tightest reading the eigenspectrum offers, violations have been observed pervasively in most practical scenarios that require learning. In this section, we use a low-dimensional signal as an example to show this universal violation as well as its properties, and more analysis on image dataset will be presented in the result section. Here we use a product of two unit circles, i.e., a 2-torus, in $\mathbb{R}^4$ under isotropic noise, with $\rho = 1$ and $D = 4$. The density factorizes over the two planes, so the score is available in closed form through modified Bessel functions and the exact spectrum is

$$\operatorname{spec}(\mathbf{H}_\mathbf{y}) = \Big\{ \underbrace{g(r_j)/r_j}_{\text{tangential: signal}} \;,\; \underbrace{g'(r_j)}_{\text{radial: noise}} \Big\}_{j=1,2},\; g(r) = -\frac{r}{\sigma^2} + \frac{\rho}{\sigma^2}\frac{I_1(z)}{I_0(z)},\;\; z = \frac{r\rho}{\sigma^2}. \tag{10}$$

Here r_j is the radius of $\mathbf{y}$ distribution on the j-th plane and $g(r) \coloneqq \frac{\mathrm{d}}{\mathrm{d}r}\log p(r)$ is the radial score of a single noisy circle, whose density at radius r is proportional to $e^{-(r^2+\rho^2)/(2\sigma^2)} I_0(r\rho/\sigma^2)$, where I_0 and I_1 are the modified Bessel functions of the first kind of orders 0 and 1; the closed form follows from $I_0' = I_1$. Each plane contributes the radial field $g(r)\,\mathbf{y}/r$, whose Jacobian is diagonal in the radial–tangential frame.

Figure 1(a) shows the eigenspectrum of a self-supervised network trained to minimize the matching error with the *analytical* score at $\sigma = 0.05$, which does not involve randomness in the loss function. Near the eigenvalue bound $-1/\sigma^2$, an eigenvalue cluster could be observed, and a large portion of this cluster lies below the bound. The peak of the cluster is within 0.014% of the analytic noise eigenvalue median, while the minimum eigenvalue $\lambda_{\min}$ is smaller than $-1/\sigma^2$ and the corresponding noise variance estimation (7) is in error by 4.5%. The violation of (7) is also observed with denoising score matching (DSM) in Figure 1(b), which shows noise estimation accuracy with $-1/\lambda_{\min}$ and the center of the cluster as is located in Figure 1(a), for both the analytical and the learned Jacobian spectrum across noise levels, tabulated in Table 3 of Appendix D.5. Both analytical curves over-estimate the noise at every level and never violate the bound, with estimation error growing with σ and bounded by $\frac{1}{2}(\sigma/\rho)^2$, confirming the second-order gap of Proposition 1 up to the constant. The learned curves replace the closed form score by an MLP with ample capacity and data trained by DSM, with the scaling of the additive DSM noise accounted for in the estimate. Then the learned $-1/\lambda_{\min}$ *under*-estimates σ^2 at every noise level, which universally violates the bound and the curvature effect enforced in Proposition 1. Using the cluster center also gives underestimation of noise but actually has better accuracy in all noise levels. The accuracy of cluster center estimation also degrades in the high-noise regime, where the noise ball becomes comparable to the manifold's radius of curvature and the posterior no longer concentrates on an approximately flat patch. Such a general violation might not be an artifact caused just by a lack of data, MLP capacity, training bias, or by the randomness of the score-matching objective. Instead, the randomness in neural network, existing in both analytical score matching and DSM, could contribute to the spread. Figure 4 of Appendix D.4 shows that even with large capacity / training epochs for analytical score matching, the eigenvalue spread in Figure 1(a) persists while score error is reduced significantly.

4 Network-Induced Score Jacobian Eigenvalue Spread

The eigenvalue bound violation of the learned score Jacobian is caused by the significantly increased spread of the eigenspectrum compared to that of the analytical score. This section aims to show the source of the spread and provides a theoretical basis for robust noise estimation in this scenario. We establish with NTK theory that the perturbation of the Jacobian is *centred*, and that a centred perturbation leaves the noise cluster's centre where it is while displacing its extremum.

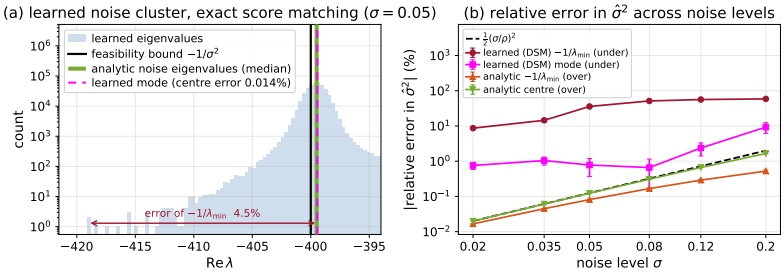

Figure 1: The bound is valid for analytical score Jacobian eigenspectrum but fails with score learning on a 2-torus in $\mathbb{R}^4$. **(a)** Jacobian eigenspectrum of a network trained to match the *analytical* score at $\sigma = 0.05$. Analytically the $D - d$ noise eigenvalues have minimal spread, while the learned spectrum spreads them over a far wider cluster and violates the eigenvalue bound. **(b)** Relative noise estimation error with DSM across noise levels in the observations $\mathbf{y}$. The analytical score Jacobian over-estimates the noise at every σ, following the second-order gap $\frac{1}{2}(\sigma/\rho)^2$ (dashed) that Proposition 1 predicts, while the learned score Jacobian always under-estimates it. The noise variance read from the minimum eigenvalue is in error by 8.7% at the lowest noise level, rising to 58.7% at the highest, while the estimate read from the cluster mode stays much closer to the true value throughout.

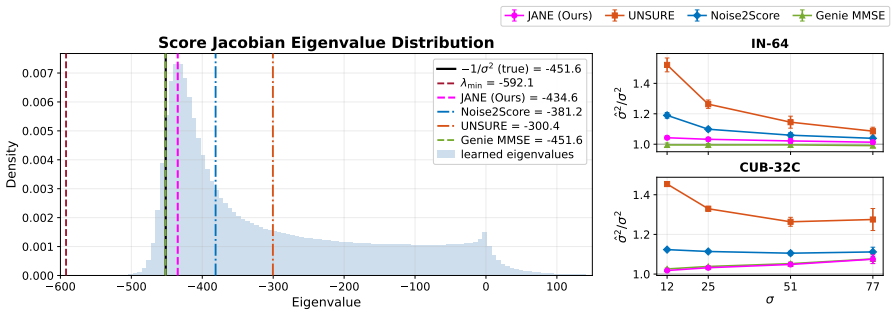

Figure 2: **Left:** eigenvalue histogram of the learned score Jacobian on IN-64 at $\sigma = 12$. The leftmost peak (magenta) sits close to the *Genie* step size (green) and the true $-1/\sigma^2$ (black), while the minimum eigenvalue (brown) falls below the bound. Noise2Score (blue) and UNSURE (orange) are contaminated by the signal subspace. **Right:** noise estimation accuracy $\hat{\sigma}^2/\sigma^2$ across noise levels on two datasets; JANE (magenta) stays closest to 1.0 and the Genie value.

4.1 The Perturbation of the Learned Jacobian

We use NTK theory to explain why the learned Jacobian carries a *centred* perturbation, and to identify what governs its spread. The argument is idealized: our networks have finite width, and kernel evolution during training also contributes to the spread without being covered here (Hanin & Nica, 2020; Yaida, 2020; Roberts et al., 2022). What it establishes is that the perturbation survives even in the ideal case, so the mismatch between the learned eigenspectrum and the analytical one is intrinsic rather than an artifact of imperfect training. In the infinite-width limit the trained output is Gaussian and splits into a deterministic part and a zero-mean Gaussian-process residual $\mathbf{g}$, which is the initialization randomness after training (Lee et al., 2019); $\mathbb{E}[\mathbf{g}] = \mathbf{0}$ because the output is affine in that initialization (Appendix C). The deterministic part is the true score plus an interpolation residual $\mathbf{r}$, so differentiating in the input decomposes the learned Jacobian as

$$\hat{\mathbf{J}} \;=\; \mathbf{H}_{\mathbf{y}} + \nabla\mathbf{r} + \mathbf{\Delta}, \tag{11}$$

where $\hat{\mathbf{J}}$ is the learned Jacobian, $\mathbf{H}_{\mathbf{y}}$ the true score Jacobian there, $\nabla\mathbf{r}$ the derivative of the interpolation residual, and $\mathbf{\Delta} = \nabla\mathbf{g}$ a zero-mean Gaussian random matrix, since the derivative of a Gaussian process is again a Gaussian process when the residual kernel $\tilde{K}$ is mean-square differentiable for smooth activations. Its covariance is the mixed second derivative of that kernel, which we collect as the *cross-output derivative covariance* $\mathbf{\Sigma}_{\partial}^{(mm')} \in \mathbb{R}^{D\times D}$, one block for each pair of output dimen-

sions, so that $\{\mathbf{\Sigma}_\partial^{(mm')}\}$ carries the full covariance of $\mathrm{vec}(\mathbf{\Delta})$; Appendix C, (63), gives it explicitly. For independent readouts the blocks are diagonal; with weight sharing across output positions, as in convolutional networks and U-Nets, the off-diagonal blocks do not vanish. The zero-mean property holds for any architecture meeting the NTK conditions and only the covariance changes. Since $\mathbf{\Delta}$ originates in the randomness of the initialization rather than in any approximation error, it persists even for a network that interpolates its training targets exactly. Next we will show $\mathbf{\Delta}$ could introduce unavoidable spread of the noise eigenvalue cluster and its impact on the spectrum.

4.2 Structure of Eigenspectrum under Perturbation

As is shown in Section 3, the eigenspectrum of score Jacobian contains information about noise that is necessary for blind denoising. In self-supervised problems the Jacobian is read at training observations and (11) applies there. The estimator is assumed to have access to the eigenvalues pooled over the subset $\mathcal{S}$ and the pooled law is analyzed. We begin from the special case in which the deterministic part vanishes: the signal is locally flat, so the noise eigenvalues are exactly degenerate as in (9), and we set the deterministic residual aside, $\nabla \mathbf{r} = \mathbf{0}$. What remains is the $\mathbf{\Delta}$ that persists even then. With $\mathbf{P}$ projecting onto the noise eigenspace of $\mathbf{H}_\mathbf{y}$, the learned Jacobian restricted to that eigenspace is $\lambda_0 \mathbf{I} + \mathbf{P}\mathbf{\Delta}\mathbf{P}$, $\lambda_0 = -1/\sigma^2$. The general case is discussed at the end of this section.

Proposition 2. *(i)* Symmetry (Proof in Appendix C.5.). *For the block $\lambda_0 \mathbf{I} + \mathbf{P}\mathbf{\Delta}\mathbf{P}$ of the special case above, let λ be the real part of an eigenvalue drawn at random from the pooled noise block. If the law of $\mathbf{\Delta}$ is invariant under $\mathbf{\Delta} \mapsto -\mathbf{\Delta}$, which is satisfied by the centered Gaussian $\mathbf{\Delta}$, then*

$$\lambda - \lambda_0 \overset{d}{=} \lambda_0 - \lambda, \tag{12}$$

so the law of λ is symmetric about λ_0, regardless of the block dimension, joint dependency, and covariance matrix. Both its median and mean are λ_0. It concerns the pooled law and not the spectrum of a single realization, which is in general asymmetric. The learned Jacobian may not be perfectly symmetric, so the block's eigenvalues are complex and real parts are considered.

(ii) Extracting λ_0 without knowledge of d. *Calculating the mean / median of the noise block requires the knowledge of noise eigenspace dimension d, which is a common approach in conventional methods. However, if the noise eigenspace λ density has a global maximum, by symmetry in (i) it has to be equal to the median / mean λ_0. With the signal dimension $d \ll D$, the contribution of signal-space eigenvalues is small compared to the noise block's and only slightly displace the maximum of the noise eigenvalue cluster when the whole Jacobian eigenspectrum is considered. Given* (6) *and the symmetry property under perturbation, the leftmost cluster of the Jacobian spectrum carries least impact from signal and its peak / mode should be approximately λ_0. Although this point is only conceptual, results in Section 3 and 6 show that it is approximately valid in many practical scenarios.*

Note that the minimum eigenvalue of the pooled block is displaced below λ_0 by an amount that grows with the scale of $\mathbf{\Delta}$, so $-1/\lambda_{\min}$ likely under-estimates σ^2.

Remark 3. *Proposition 2 (i) proves that the center of noise eigenspace λ density corresponds to λ_0 in an idealized scenario, while (ii) discusses the extraction of λ_0 without knowing space dimension d. However, (ii) relies on the assumption that the noise eigenspace λ density is unimodal, which lacks theoretical guarantee. In this remark we provide a simplification of the perturbation model and show that intuitively a unimodal noise cluster is expected. If we treat each noise eigenvalue as receiving a scalar perturbation on λ_0 as in Appendix C.3, then the scalar perturbation could be viewed as a linear functional of jointly Gaussian entries of $\mathbf{\Delta}$ and is hence centered Gaussian. In the idealized case, the pooled density of eigenvalues is the mixture*

$$p(\lambda) \;=\; \sum_i w_i \, \varphi_{v_i}(\lambda - \lambda_0)\,, \tag{13}$$

where φ_{v_i} is the centered Gaussian density of variance v_i and w_i the probability of drawing from this distribution. Such a mixture is not *itself Gaussian, but every term of* (13) *is a Gaussian with the same center so p is therefore symmetric with a single peak at λ_0. The covariance of perturbation sets how wide the cluster is rather than where its peak lies. This simplification serves as an intuitive justification rather than a proof to expect unimodality, since treating a degenerate block one eigenvalue at a time is not exact, even though Proposition 2(i) is proved. More results will be discussed in Section 6.*

Table 1: Each cell reports denoising **PSNR in dB** | **noise estimation accuracy** $\hat{\sigma}^2/\sigma^2$ (ideal 1.00). Bold marks the best *blind* method per metric and column, the two bolded independently; ties at the displayed precision are both bolded. Noisy gives the input PSNR. *Genie* tunes step sizes against clean signals rather than estimating σ^2, so its ratio can be below 1. Results are averaged over 10 seeds.

σ	Method	CIFAR-10	IN-32	IN-64	CUB-64	Channel	CUB-32C
12	Noisy	26.68 | –	26.72 | –	26.71 | –	26.62 | –	26.78 | –	26.65 | –
	JANE	**31.30** | **1.03**	**31.66** | **1.06**	**31.92** | **1.04**	**31.96** | **1.03**	**32.82** | **1.00**	**32.46** | **1.02**
	Noise2Score	31.06 | 1.17	31.16 | 1.21	31.48 | 1.19	31.75 | 1.14	32.77 | 1.06	32.35 | 1.12
	UNSURE	29.33 | 1.54	29.15 | 1.56	29.53 | 1.52	29.99 | 1.47	31.37 | 1.35	30.64 | 1.45
	Genie	*31.31* | *1.00*	*31.71* | *0.99*	*31.94* | *1.00*	*31.97* | *1.00*	*32.82* | *1.00*	*32.46* | *1.03*
25	Noisy	20.45 | –	20.54 | –	20.52 | –	20.36 | –	20.66 | –	20.40 | –
	JANE	**26.73** | **1.03**	**27.31** | **1.04**	**27.71** | **1.03**	**27.42** | **1.03**	**27.12** | **1.00**	**27.67** | **1.03**
	Noise2Score	26.58 | 1.11	27.06 | 1.12	27.50 | 1.10	27.28 | 1.09	27.05 | 1.06	27.57 | 1.11
	UNSURE	25.55 | 1.30	25.75 | 1.30	26.33 | 1.26	26.24 | 1.27	26.00 | 1.27	26.31 | 1.33
	Genie	*26.75* | *1.00*	*27.34* | *1.00*	*27.73* | *1.00*	*27.43* | *1.00*	*27.12* | *1.00*	*27.67* | *1.04*
51	Noisy	14.73 | –	14.85 | –	14.84 | –	14.60 | –	15.02 | –	14.67 | –
	JANE	**22.66** | **1.03**	**23.18** | **1.04**	**23.59** | **1.02**	**23.29** | **1.04**	**21.19** | **1.01**	**23.35** | **1.05**
	Noise2Score	22.57 | 1.07	23.09 | 1.07	23.48 | 1.06	23.22 | 1.07	21.17 | 1.04	23.28 | 1.11
	UNSURE	21.99 | 1.17	22.36 | 1.18	22.86 | 1.14	22.64 | 1.16	20.13 | 1.25	22.21 | 1.26
	Genie	*22.68* | *1.00*	*23.21* | *1.00*	*23.60* | *1.00*	*23.33* | *1.01*	*21.20* | *1.00*	*23.35* | *1.05*
77	Noisy	11.84 | –	11.94 | –	11.93 | –	11.73 | –	12.09 | –	11.79 | –
	JANE	**20.44** | **1.03**	**20.95** | **1.03**	**21.41** | **1.01**	**21.16** | **1.03**	**17.42** | **1.01**	**21.06** | **1.07**
	Noise2Score	20.38 | 1.05	20.91 | 1.05	21.35 | 1.04	21.10 | 1.05	17.37 | 1.05	21.02 | 1.11
	UNSURE	19.95 | 1.12	20.48 | 1.12	21.02 | 1.09	20.71 | 1.11	16.39 | 1.25	19.74 | 1.28
	Genie	*20.47* | *1.00*	*20.99* | *1.00*	*21.42* | *0.99*	*21.19* | *1.01*	*17.43* | *1.00*	*21.06* | *1.08*

Away from the special case the centre of the noise cluster is displaced, and curvature is the one displacement we can bound: Proposition 1 places the true noise eigenvalues above the boundary, in $[\lambda_0, \lambda_0 + \eta]$, so it moves the centre to the right by an amount second order in σ/ρ. Other factors act on the centre as well — the interpolation residual; the coupling between the noise block and the signal directions, which restricting the Jacobian to the block discards; and the asymmetry of the learned Jacobian. We do not model these separately or assign a share of the displacement to each: they are not separable from one another in data, and quantifying any of them would require the true Jacobian, which is unavailable in most practical scenarios. What we report instead is that they do not destroy the noise estimation performance using the noise cluster mode, and the mode is insensitive to the *size* of the perturbation while the extremum is not. Section 6 reports the resulting accuracy and compares the mode- and extremum-based estimates.

Proposition 2(ii) locates the centre by the leftmost local maximum, which restricts the impact of signal-space eigenvalues to the noise estimation. In the image spectra of Figure 2 the signal and noise eigenvalues are mixed significantly. Conventional methods that try to use the whole spectrum statistics will unavoidably be affected by the signal components and the noise estimation will be contaminated. Also this method does not rely on clear separation of signal and noise clusters as we do not need the knowledge of space dimensions to isolate noise components.

Independent initializations give independent draws of $\mathbf{\Delta}$. Averaging the Jacobians of several networks trained on the same data therefore reduces $\mathbf{\Delta}$'s contribution while leaving $\nabla \mathbf{r}$ unchanged, since only $\mathbf{\Delta}$ differs between them. The colour configurations use an ensemble for this reason (Section 6).

5 METHODS: JANE

JANE is a complete self-supervised procedure: it takes the noisy observations $\{\mathbf{y}^{(i)}\}_{i=1}^{N}$ as its only input and returns both a noise estimate and the denoised signals; it is stated as pseudocode in Algorithm 1 of Appendix D. Given the noisy observations, a score network $\hat{\mathbf{s}}_\theta$ is first trained in a self-supervised manner to approximate the ground-truth score function, and the score Jacobian $\hat{\mathbf{H}}_{\mathbf{y}^{(i)}} = \nabla_{\mathbf{y}} \hat{\mathbf{s}}_\theta(\mathbf{y}^{(i)})$ at each sample is obtained by back propagation. To balance computational cost against the reliability of the eigenspectrum, a sufficiently large subset $\mathcal{S}$ of observations is randomly

selected for eigendecomposition, and the $D|\mathcal{S}|$ resulting eigenvalues are pooled into a single density $\hat{p}(\lambda)$. The noise variance estimate is then

$$\hat{\sigma}^2 = -1/\hat{\mu}, \text{ where } \hat{\mu} = \min\{\mu : \hat{p}\,'(\mu) = 0,\ \hat{p}\,''(\mu) < 0\}. \tag{14}$$

Proposition 2 motivates reading the mode. It places the centre of the pooled noise law exactly at $-1/\sigma^2$, whereas the extremum is displaced by the width of the perturbation; and of the statistics that locate that centre, the mode is the only one computable without knowing d, which both the mean and the median require. Remark 3 gives the reason to expect a single peak, which is then the centre. The density $\hat{p}$ is estimated from the pooled eigenvalues; the numerical recipe for locating its leftmost peak is given in Appendix D.2. The estimate is then applied to every observation through Tweedie's formula, $\hat{\mathbf{x}}^{(i)} = \mathbf{y}^{(i)} + \hat{\sigma}^2\,\hat{\mathbf{s}}_\theta(\mathbf{y}^{(i)})$, so that a single trained network supplies both the noise level and the denoised signal.

6 Experiments

We evaluate on 6 configurations from 3 image datasets and 1 synthetic wireless channel dataset, spanning grayscale and color. (1) CIFAR-10 (Krizhevsky, 2009): 32×32, grayscale, $D = 1024$, 50k samples; (2) IN-32: 32×32 patches from 13 edge-rich ImageNet (Deng et al., 2009) classes, grayscale, $D = 1024$, 67k samples; (3) IN-64: the same classes at 64×64, grayscale, $D = 4096$, 67k samples; (4) CUB-64: 64×64 bird images from CUB-200-2011 (Wah et al., 2011) with bounding-box crops, grayscale, $D = 4096$, 12k samples; (5) Channel: 32×32 simulated channel phase images, 10k samples, generated as in Appendix D.3; (6) CUB-32C: 32×32 RGB bird crops, $D = 3072$, 12k samples. CUB-32C is the only color configuration, the only one using an ensemble, and the one with the smallest $|\mathcal{S}|$; colour, ensembling and subset size are therefore confounded there, and we do not draw a colour-specific conclusion from it. Pixel values are normalized to $[0, 1]$ and we test at $\sigma \in \{12, 25, 51, 77\}$ in pixel scale. We compare against *UNSURE* (Tachella et al., 2025), *Noise2Score* (Kim et al., 2022), and *Genie MMSE*, which searches the step size α minimizing the error of $\hat{\mathbf{x}} = \mathbf{y} + \alpha\,\hat{\mathbf{s}}_\theta(\mathbf{y})$ against the clean signal with the same trained network, and is therefore the best error achievable with that score rather than the true MMSE. All methods share the same U-Nets; the color configuration uses an ensemble of 32 networks. We restrict evaluation to estimators that explicitly estimate σ^2 from data; other denoising methods are discussed in Section 2.1. We diagonalize the $D \times D$ Jacobian on a random data subset and retain real parts. For each simulation setup we run 10 different seeds and show mean performance, with the seed-to-seed dispersion of every cell in Table 4. Remaining settings are in Appendix D.

Figure 2 displays the eigenvalue histogram of the learned score Jacobian on IN-64 at $\sigma = 12$, as well as the negative inverse of noise estimation $-1/\hat{\sigma}^2$ with different methods. JANE (magenta) identifies the leftmost mode, closely matching the Genie estimate. The same panel shows why the randomness analysis is needed: for the ideal Jacobian the smallest eigenvalue would be the optimal estimator by (7), yet under score matching it falls far below the bound. Noise2Score and UNSURE also deviate, since both aggregate across all dimensions without separating signal from noise, whereas the mode reads the noise cluster directly.

Table 2 shows the relative error when we use different score Jacobian eigenspectrum statistics for noise estimation. For the 6 configurations considered in this work, using minimum eigenvalues results in 29.6% underestimation on average, and in every one of these runs, as the minimum eigenvalue is displaced by the perturbation shown in Section 4.2. The spectrum mean $-1/\bar{\lambda}$ results in 27.1% overestimation, given that the d signal directions enter the average. Restricting the average to the noise block would require knowing the signal dimension, which classical noise-subspace methods estimate by a separate detection step (Wax & Kailath, 1985; Kritchman & Nadler, 2009). In contrast, JANE, which reads the noise mode, shows an average of 3.3% error, significantly outperforming other spectrum statistics without relying on the subspace dimension knowledge.

Table 1 reports the ratio $\hat{\sigma}^2/\sigma^2$ for all 24 configurations. Averaged over all settings, JANE attains a mean absolute relative error of 3.3%, against 9.4% for Noise2Score and 28.1% for UNSURE, which are $2.9\times$ and $8.6\times$ reduction respectively. These aggregates take $|\hat{\sigma}^2/\sigma^2 - 1|$ per seed and average over seeds and settings from unrounded values, so they do not equal the same function of the seed-averaged cells in Table 1. JANE also approaches genie performance (1.9%) that relies on the ground truth, while the gap is usually wider at low noise, where the signal directions weigh most

Table 2: Mean relative error $|\hat{\sigma}^2/\sigma^2 - 1|$ for three statistics of the *same* eigenvalue spectrum, over all 6 configurations and the same four noise levels. $\bar{\lambda}$ is the whole-spectrum mean and $\hat{\mu}$ the leftmost mode of the pooled density, which could isolate the noise cluster without knowing d.

Statistic	CIFAR-10	IN-32	IN-64	CUB-64	Channel	CUB-32C	**All**
$-1/\lambda_{\min}$	31.3%	32.8%	26.5%	26.8%	32.3%	27.7%	**29.6%**
$-1/\bar{\lambda}$	28.2%	29.4%	26.0%	24.7%	28.2%	26.2%	**27.1%**
$-1/\hat{\mu}$ (JANE)	3.2%	4.5%	2.8%	3.4%	1.2%	4.3%	**3.3%**

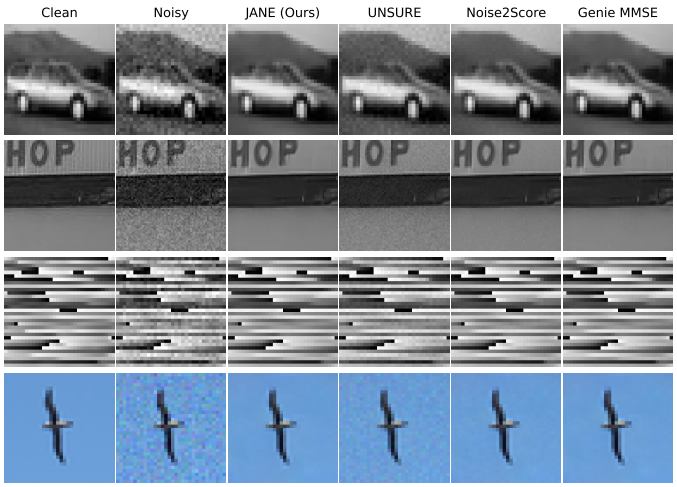

Figure 3: Visual denoising comparison at $\sigma = 12$ across four datasets: CIFAR-10 (row 1), ImageNet patches (row 2), wireless channel phase (row 3), and CUB-200 in color (row 4).

heavily against the noise. Among all 6 configurations we show the estimation error trend of IN-64 and CUB-32C in Figure 2. At low noise regime, since the signal manifold is not exactly low-dimensional, the posterior covariance retains small but nonzero eigenvalues $\nu_i > 0$ in most directions, which by (6) lifts the cluster slightly above $-1/\sigma^2$ and biases the estimate upward.

JANE attains the highest PSNR of the blind methods in every setting and lies within 0.05 dB of oracle Genie MMSE in all 24 seed-averaged settings. Averaged over the 24 settings the gain is 1.21 dB over UNSURE. The margin over Noise2Score is largest at low noise and narrows as σ grows, from 0.26 dB at $\sigma = 12$ to 0.05 dB at $\sigma = 77$, while JANE's distance to Genie stays essentially constant: the denoising task becomes insensitive to $\hat{\sigma}^2$ at high noise, where every method approaches the oracle, rather than the estimate losing accuracy. Figure 3 shows representative results at $\sigma = 12$ across the 4 datasets: JANE is visually indistinguishable from Genie MMSE, and leaves fewer artifacts than Noise2Score or UNSURE, most visibly in background regions where high-frequency noise should not appear. An over-large step overshoots the posterior mean and leaves a scaled, sign-inverted copy of the noise, so an over-estimated σ^2 appears as residual graininess rather than over-smoothing. Note that denoising performance need not track noise estimation error, since with an imperfect score it is proximity to *Genie* $\hat{\sigma}^2$ that is optimal for denoising by construction.

7 Conclusion

We analyzed the dependence of score Jacobian eigenspectrum on noise variance in both analytical and score matching setup, locating the unavoidable perturbation introduced by the neural network. We proposed the blind denoising algorithm JANE to mitigate the impact of perturbation during blind noise estimation and denoising, which estimates the noise variance from the leftmost mode of the learned score Jacobian's eigenspectrum, attaining 3.3% mean relative error in σ^2 and matching an oracle step-size search to within 0.05 dB. Two directions remain open: the $\mathcal{O}(D^3)$ eigendecomposition per observation, and comparison with estimators that do not report an explicit σ^2 (Appendix D.2).

AI USE STATEMENT

In this work, the research problem, the proposed estimator, the algorithm, and the experimental design are the authors' own, as is the interpretation of the results. We used generative AI tools in support of this work: to help develop the theoretical analysis and the proofs, to implement the methods and experiments in code, to prepare figures, to draft and edit parts of the manuscript, and to survey related literature. We did not use them to directly generate synthetic data sets. The synthetic channel data follows the 3GPP CDL specification (3GPP, 2020) and was produced by code. We have reviewed all AI-assisted work: every theoretical statement was checked by the authors, and every reported number was produced by code the authors executed and verified. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies noise estimation for image and wireless-channel denoising and raises no concerns involving human subjects, personally identifying data, or dual use. The 13 ImageNet classes used (Appendix D.2) are object, animal and texture categories, not person categories. The image datasets are standard public benchmarks used under their intended terms and the channel data is synthetic.

REPRODUCIBILITY STATEMENT

All datasets are public except the synthetic channel data, which is generated from the 3GPP CDL specification (3GPP, 2020) with parameters given in Appendix D.3. Score networks are U-Nets (Ronneberger et al., 2015) trained with DSM at $\sigma_{\mathrm{dsm}} = 0.01$ for 1000 epochs (300 for IN-64), learning rate 5×10^{-4}, batch size 256–1024 depending on the dataset (Appendix D.2), Adam, with early stopping. Noise estimation pools the eigenvalues over a random subset of observations and locates the leftmost peak of their density by the two-stage procedure of Appendix D.2. Eigenvalues are obtained from `numpy.linalg.eigvals` with real parts retained. Color configurations average scores and Jacobians over an ensemble of 32 independently initialized networks. Every reported number is a mean over the same 10 training seeds, whose seed-to-seed dispersion is given in Table 4.

REFERENCES

3GPP. Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report TR 38.901, 3rd Generation Partnership Project, 2020.

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pp. 524–533. PMLR, 2019.

Guangyong Chen, Fengyuan Zhu, and Pheng Ann Heng. An efficient statistical method for image noise level estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 477–485, 2015.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, et al. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

Jia Deng, Wei Dong, Richard Socher, et al. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

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

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Shi Guo, Zifei Yan, Kai Zhang, et al. Toward convolutional blind denoising of real photographs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1712–1722. IEEE, 2019.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgndT4KwB.

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

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

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

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, et al. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024.

Daniel N. Kaslovsky and François G. Meyer. Non-asymptotic analysis of tangent space perturbation. *Information and Inference: A Journal of the IMA*, 3(2):134–187, 2014.

Kwanyoung Kim, Taesung Kwon, and Jong Chul Ye. Noise distribution adaptive self-supervised image denoising using Tweedie distribution and score matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1998–2006. IEEE, 2022.

Shira Kritchman and Boaz Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void-learning denoising from single noisy images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2124–2132. IEEE, 2019.

Jaehoon Lee, Yasaman Bahri, Roman Novak, et al. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, et al. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, et al. Noise2Noise: Learning image restoration without clean data. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lehtinen18a.html.

Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Single-image noise level estimation for blind denoising. *IEEE Transactions on Image Processing*, 22(12):5226–5237, 2013.

Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. In *International Conference on Learning Representations*, 2024.

Chenlin Meng, Yang Song, Wenzhe Li, et al. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34:25359–25369, 2021.

Christopher A. Metzler, Ali Mousavi, Reinhard Heckel, et al. Unsupervised learning with Stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

Sreyas Mohan, Zahra Kadkhodaie, Eero P. Simoncelli, et al. Robust and interpretable blind image denoising via bias-free convolutional neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlSmC4FPS.

Radford M. Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

Roman Novak, Lechao Xiao, Jaehoon Lee, et al. Bayesian deep convolutional networks with many channels are Gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.

Martin Raphan and Eero P. Simoncelli. Least squares estimation without priors or supervision. *Neural Computation*, 23(2):374–420, 2011.

Jan Rataj and Luděk Zajíček. On the structure of sets with positive reach. *Mathematische Nachrichten*, 290(11-12):1806–1829, 2017.

Herbert Robbins. An empirical Bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 388–394. Springer, 1992.

Daniel A. Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory*. Cambridge University Press, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Shirin Shoushtari, Yi Wang, Xiao Shi, et al. Eigenscore: OOD detection using posterior covariance in diffusion models. In *International Conference on Learning Representations*, 2026.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Shakarim Soltanayev and Se Young Chun. Training deep learning based denoisers without ground truth data. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Yang Song, Sahaj Garg, Jiaxin Shi, et al. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020.

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

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

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Julián Tachella, Junqi Tang, and Michael E. Davies. The neural tangent link between CNN denoisers and non-local filters. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8614–8623. IEEE, 2021.

Julián Tachella, Mike Davies, and Laurent Jacques. UNSURE: self-supervised learning with unknown noise level and Stein's unbiased risk estimate. In *International Conference on Learning Representations*, 2025.

Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of tr(f(A)) via stochastic Lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, 2017.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, et al. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *International Conference on Learning Representations*, 2025.

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

Catherine Wah, Steve Branson, Peter Welinder, et al. The Caltech-UCSD Birds-200-2011 dataset. 2011.

Mati Wax and Thomas Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392, 1985.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006.

Yihong Wu and Sergio Verdú. MMSE dimension. *IEEE Transactions on Information Theory*, 57(8): 4857–4879, 2011.

Sho Yaida. Non-Gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pp. 165–192. PMLR, 2020.

Greg Yang. Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

Greg Yang and Etai Littwin. Tensor programs IIb: Architectural universality of neural tangent kernel training dynamics. In *International Conference on Machine Learning*, pp. 11762–11772. PMLR, 2021.

Zongsheng Yue, Hongwei Yong, Qian Zhao, et al. Variational denoising network: Toward blind noise modeling and removal. *Advances in Neural Information Processing Systems*, 32, 2019.

A The Anisotropic Case

This appendix treats the general anisotropic case, which Section 3 sets aside. It is included for completeness: no experiment in this paper uses anisotropic noise, and the estimator returns only the smallest noise variance in that setting. When the noise level is different in each dimension, the noise bound below only applies to the minimum noise variance.

Proposition 4 (Eigenvalue lower bound for $\mathbf{H_y}$)**.** *Let $\mathbf{\Sigma_{x|y}} = \mathbf{C} + \mathbf{C}\mathbf{H_y}\mathbf{C}$ (Miyasawa identity) with $\mathbf{H_y}$ symmetric, $\mathbf{C} \succ 0$, and $\mathbf{\Sigma_{x|y}} \succeq 0$. If there exists at least one eigenvalue of $\mathbf{H_y}$ that is negative (which is common for Jacobian of score function), then*

$$\lambda_i(\mathbf{H_y}) \geq \lambda_{\min}(\mathbf{H_y}) \geq \frac{\lambda_{\min}(\mathbf{\Sigma_{x|y}}) - \lambda_{\max}(\mathbf{C})}{\lambda_{\min}(\mathbf{C})\,\lambda_{\max}(\mathbf{C})}. \tag{15}$$

Proof of Proposition 4. Define $\tilde{\mathbf{H}} = \mathbf{C}^{1/2}\mathbf{H_y}\mathbf{C}^{1/2}$, so that $\tilde{\mathbf{H}} = \mathbf{C}^{-1/2}\mathbf{\Sigma_{x|y}}\mathbf{C}^{-1/2} - \mathbf{I}$. By the Courant–Fischer theorem applied to the congruence $\mathbf{C}^{-1/2}\mathbf{\Sigma_{x|y}}\mathbf{C}^{-1/2}$:

$$\lambda_i(\tilde{\mathbf{H}}) \geq \frac{\lambda_{\min}(\mathbf{\Sigma_{x|y}})}{\lambda_{\max}(\mathbf{C})} - 1 = \frac{\lambda_{\min}(\mathbf{\Sigma_{x|y}}) - \lambda_{\max}(\mathbf{C})}{\lambda_{\max}(\mathbf{C})}. \tag{16}$$

Since $\mathbf{H_y} = \mathbf{C}^{-1/2}\tilde{\mathbf{H}}\mathbf{C}^{-1/2}$, for any unit $\mathbf{v}$ with $\mathbf{w} = \mathbf{C}^{-1/2}\mathbf{v}$:

$$\mathbf{v}^\top \mathbf{H_y}\mathbf{v} = \|\mathbf{w}\|^2 \cdot \frac{\mathbf{w}^\top \tilde{\mathbf{H}}\mathbf{w}}{\|\mathbf{w}\|^2} \geq \lambda_{\min}(\tilde{\mathbf{H}}) \cdot \|\mathbf{w}\|^2. \tag{17}$$

The hypothesis $\lambda_{\min}(\mathbf{H_y}) < 0$ transfers to $\tilde{\mathbf{H}}$ because $\tilde{\mathbf{H}} = \mathbf{C}^{1/2}\mathbf{H_y}\mathbf{C}^{1/2}$ is a congruence and congruences preserve signature by Sylvester's law of inertia. Since $\lambda_{\min}(\tilde{\mathbf{H}}) < 0$ and $\|\mathbf{w}\|^2 = \mathbf{v}^\top \mathbf{C}^{-1}\mathbf{v} \le 1/\lambda_{\min}(\mathbf{C})$, the product achieves minimum at the largest $\|\mathbf{w}\|^2$:

$$\lambda_{\min}(\mathbf{H_y}) \ge \frac{\lambda_{\min}(\tilde{\mathbf{H}})}{\lambda_{\min}(\mathbf{C})} \ge \frac{\lambda_{\min}(\mathbf{\Sigma_{x|y}}) - \lambda_{\max}(\mathbf{C})}{\lambda_{\min}(\mathbf{C})\,\lambda_{\max}(\mathbf{C})}. \tag{18}$$

□

General case. The right-hand side of (15) increases with $\lambda_{\min}(\mathbf{\Sigma_{x|y}})$, so replacing it by the least value that $\mathbf{\Sigma_{x|y}} \succeq 0$ permits, namely 0, can only weaken the bound:

$$\lambda_{\min}(\mathbf{H_y}) \ge \frac{0 - \lambda_{\max}(\mathbf{C})}{\lambda_{\min}(\mathbf{C})\,\lambda_{\max}(\mathbf{C})} = \frac{-1}{\lambda_{\min}(\mathbf{C})} = \frac{-1}{\sigma^2_{\min}}, \tag{19}$$

where $\sigma^2_{\min} \coloneqq \lambda_{\min}(\mathbf{C})$ is the smallest noise variance. Since $\lambda_{\min}(\mathbf{H_y}) < 0$, rearranging gives $\sigma^2_{\min} \le -1/\lambda_{\min}(\mathbf{H_y})$: the quantity $-1/\lambda_{\min}(\mathbf{H_y})$ provides an upper bound on the smallest noise variance.

B The Curvature Gap for the True Score Jacobian

This appendix proves Proposition 7, which supplies the bound $\varepsilon = \mathcal{O}(\sigma^4/\rho^2)$ quoted in Proposition 1. It concerns the *true* score Jacobian alone, with no reference to neural networks. The proposition shows that for a d-dimensional signal manifold of reach ρ, the $D - d$ trailing eigenvalues of $\mathbf{H_y}$ all lie in $[-1/\sigma^2,\ -1/\sigma^2 + \eta]$ with $\eta = c_4/\rho^2$, while the d leading ones lie at least c_2/σ^2 above the left endpoint. The following analysis studies signals and observations satisfying:

- **(B1)** *Signal geometry.* The clean signal $\mathbf{x}$ is supported on a compact C^2 submanifold $\mathcal{M} \subset \mathbb{R}^D$ of dimension d with reach $\rho > 0$ (Federer, 1959); see also Niyogi et al. (2008), and its distribution has a density q with respect to the volume measure on $\mathcal{M}$ satisfying $0 < q_{\min} \le q \le q_{\max} < \infty$.
- **(B2)** *Noise.* $\mathbf{y} = \mathbf{x} + \mathbf{n}$ with $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_D)$ independent of $\mathbf{x}$.

Here (B1) and (B2) fix the signal geometry and the noise model, and hold for any smoothly-embedded signal manifold observed under isotropic Gaussian noise. Two further conditions regarding noise level and manifold curvature will be discussed in the following paragraphs to support the proof.

B.1 The curvature-set floor

A curved manifold does not satisfy the degeneracy (9): the posterior spreads *along* the manifold, curvature converts part of that spread into variance along the normal directions, and $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$ is generically full rank. The task is therefore to bound ν_i rather than to set it to zero. Related analyses of the sample covariance of noisy points near a curved manifold appear in local principal component analysis (Kaslovsky & Meyer, 2014).

For each $p \in \mathcal{M}$, the *tangent space* $T_p\mathcal{M}$ is the d-dimensional linear subspace of $\mathbb{R}^D$ spanned by the velocities $\gamma'(0)$ of smooth curves γ lying in $\mathcal{M}$ with $\gamma(0) = p$, translated to the origin so that it is a linear rather than an affine subspace. And $B(\mathbf{z}, r) \coloneqq \{\mathbf{z}' \in \mathbb{R}^D : \|\mathbf{z}' - \mathbf{z}\| < r\}$ denotes the ambient Euclidean ball, so $\mathcal{M} \cap B(\mathbf{z}, r)$ is the part of the manifold lying within Euclidean distance r of $\mathbf{z}$. Two further quantities are used below: $\mathrm{diam}(\mathcal{M}) \coloneqq \sup_{\mathbf{x},\mathbf{x}' \in \mathcal{M}} \|\mathbf{x} - \mathbf{x}'\|$ and $\mathrm{vol}(\mathcal{M})$, the total mass of the volume measure on $\mathcal{M}$; both are finite because $\mathcal{M}$ is compact by (B1). Besides, everything in this subsection rests on four facts about sets of positive reach, all available in Rataj & Zajíček (2017); we quote rather than reprove them.

(i) There is a unique nearest point. Let $\mathbf{x}^*$ denote a point of $\mathcal{M}$ at minimum distance from $\mathbf{y}$, so that $\|\mathbf{y} - \mathbf{x}^*\| = \tau = \mathrm{dist}(\mathbf{y}, \mathcal{M})$. At least one such point exists because $\mathcal{M}$ is compact. Reach ρ says exactly that any point lying within distance ρ of $\mathcal{M}$ has *only one* nearest point on $\mathcal{M}$. We therefore work throughout with $\tau < \rho$, so that $\mathbf{x}^*$ is unique and is a well-defined function of $\mathbf{y}$.

(ii) The offset to that point is normal. Write $\mathbf{T} \coloneqq T_{\mathbf{x}^*}\mathcal{M}$ and $\mathbf{N} \coloneqq \mathbf{T}^{\perp}$. Because $\mathbf{x}^*$ minimises $\|\mathbf{y}-\mathbf{x}\|^2$ over $\mathcal{M}$, differentiating along any curve in $\mathcal{M}$ through $\mathbf{x}^*$ gives $\mathbf{y}-\mathbf{x}^* \perp \mathbf{T}$, so $\mathbf{y}-\mathbf{x}^*$ lies in the normal space $\mathbf{N}$, of dimension $D-d$ (Rataj & Zajíček, 2017, Prop. 3.1(vi)). It is worth being explicit that $\mathbf{y}-\mathbf{x}^*$ is the offset to the *nearest point on the manifold*, not the noise realisation $\mathbf{y}-\mathbf{x}$. The latter is a Gaussian draw and has a tangential component in general; $\mathbf{x}^* \neq \mathbf{x}$, and only $\mathbf{y}-\mathbf{x}^*$ is normal.

(iii) A graph chart exists. By Rataj & Zajíček (2017, Prop. 7.4) there are a radius $r_0 > 0$ and a $C^{1,1}$ map ψ on $\{\mathbf{v} \in \mathbf{T} : \|\mathbf{v}\| < r_0\}$ with values in $\mathbf{N}$, $\psi(\mathbf{0}) = \mathbf{0}$ and $D\psi(\mathbf{0}) = \mathbf{0}$, such that the *chart map*

$$F(\mathbf{v}) \coloneqq \mathbf{x}^* + \mathbf{v} + \psi(\mathbf{v}), \qquad \mathbf{v} \in \mathbf{T},\ \|\mathbf{v}\| < r_0, \tag{20}$$

has image $\mathcal{G} \coloneqq F\big(\{\|\mathbf{v}\| < r_0\}\big)$, and whose derivative $D\psi$ is Lipschitz with constant c_ψ/ρ, where c_ψ is an *absolute* constant: it does not depend on $\mathcal{M}$, on d or D, or on σ and ρ. Shrinking the radius if necessary we take $r_0 \le \rho/2$. Thus ψ is fixed by $\mathcal{M}$, not chosen: it records how far $\mathcal{M}$ departs from its own tangent space. As sets,

$$\mathcal{G} \ \subseteq\ \mathcal{M}, \qquad\qquad \mathcal{M} \cap B(\mathbf{x}^*, r_0) \ \subseteq\ \mathcal{G}, \tag{21}$$

the second because $\|\mathbf{v}\| \le \|\mathbf{x}-\mathbf{x}^*\|$ puts the tangential offset of any such $\mathbf{x}$ in the domain of ψ, while reach ρ excludes a second sheet of $\mathcal{M}$ crossing $B(\mathbf{x}^*, r_0)$. That same inequality is why equality fails: the domain of ψ constrains $\|\mathbf{v}\|$ alone, whereas $B(\mathbf{x}^*, r_0)$ constrains the full displacement $\sqrt{\|\mathbf{v}\|^2 + \|\psi(\mathbf{v})\|^2}$.

(iv) Two curvature bounds. Rataj & Zajíček (2017, Prop. 3.3), which is Federer's Theorem 4.18, states $\operatorname{dist}(\mathbf{x}-\mathbf{x}^*, \mathbf{T}) \le \|\mathbf{x}-\mathbf{x}^*\|^2/(2\rho)$. With $\|\mathbf{x}-\mathbf{x}^*\|^2 = \|\mathbf{v}\|^2 + \|\psi(\mathbf{v})\|^2$ this reads $\|\psi(\mathbf{v})\| \le (\|\mathbf{v}\|^2 + \|\psi(\mathbf{v})\|^2)/(2\rho)$; solving the quadratic on the branch through $\psi(\mathbf{0}) = \mathbf{0}$, and using $D\psi(\mathbf{0}) = \mathbf{0}$ with the Lipschitz bound from (iii),

$$\|\psi(\mathbf{v})\| \ \le\ \rho - \sqrt{\rho^2 - \|\mathbf{v}\|^2} \ \le\ \frac{\|\mathbf{v}\|^2}{\rho}, \qquad \|D\psi(\mathbf{v})\| \ \le\ \frac{c_\psi \|\mathbf{v}\|}{\rho}. \tag{22}$$

The second inequality on the left uses only $\rho + \sqrt{\rho^2 - \|\mathbf{v}\|^2} \ge \rho$, and is attained by a sphere of radius ρ at both endpoints $\|\mathbf{v}\| = 0$ and $\|\mathbf{v}\| = \rho$. The source gives no numerical value for c_ψ, so it is carried symbolically and absorbed into c_2 and c_4 below. Let $\mathbf{v}$ denote the tangential coordinate of $\mathbf{x}$ under the posterior $\pi(\cdot \mid \mathbf{y})$.

With r_0 available we can state the remaining two conditions, both used from Lemma 5 onwards.

- **(B3)** *Noise below the chart radius.* $\sigma \le r_0$.
- **(T)** *Observation close to the manifold.* $\tau \coloneqq \operatorname{dist}(\mathbf{y}, \mathcal{M}) \le r_0/4$, which since $r_0 \le \rho/2$ implies $\tau \le \rho/8 < \rho$ and so supplies the condition used in (i).

Here (T) is an *event* rather than an assumption we may impose: Gaussian noise is unbounded, so no condition on σ can guarantee it, and every statement below is conditional on it. Since τ concentrates at $\sigma\sqrt{D-d}$, most observations satisfy (T) when $\sigma\sqrt{D-d} \lesssim r_0$, and it is therefore (T) rather than (B3) that binds in high ambient dimension, (B3) asking only $\sigma \le r_0$. Quantitatively, at $D = 4096$ and $\sigma = 77/255$, $\sigma\sqrt{D-d} \approx 19$ against a data-cube diameter of 64, so (T) fails at the image resolutions of Section 6 and Proposition 1 is claimed only for the low-dimensional setting of Section 3. We use (T) for a structural rather than a quantitative conclusion. A comparable condition appears in analyses of tangent-space recovery (Kaslovsky & Meyer, 2014).

Lemma 5 (Posterior tangential moments)**.** *Under (B1)–(B3) and (T) there are constants $c_2, c_4 > 0$ depending only on d, $q_{\max}/q_{\min}$, $\operatorname{diam}(\mathcal{M})/\rho$, $\operatorname{vol}(\mathcal{M})/\rho^d$, r_0/ρ and c_ψ such that $\mathbb{E}[\|\mathbf{v}\|^4 \mid \mathbf{y}] \le c_4\,\sigma^4$ and $\operatorname{Var}[\mathbf{u}^\top\mathbf{v} \mid \mathbf{y}] \ge c_2\,\sigma^2$ for every unit $\mathbf{u} \in \mathbf{T}$.*

Proof. Write $\mathbf{b} \coloneqq \mathbf{y}-\mathbf{x}^*$, so $\|\mathbf{b}\| = \tau$ and, by fact (ii), $\mathbf{b} \in \mathbf{N}$. Parameterising $\mathbf{x} = F(\mathbf{v})$ through (21) and using orthogonality of $\mathbf{T}$ and $\mathbf{N}$,

$$\mathbf{y}-\mathbf{x} = \underbrace{\big(\mathbf{b}-\psi(\mathbf{v})\big)}_{\in\,\mathbf{N}} + \underbrace{(-\mathbf{v})}_{\in\,\mathbf{T}}, \quad \|\mathbf{y}-\mathbf{x}\|^2 = \tau^2 + \Phi(\mathbf{v}),\ \Phi(\mathbf{v}) \coloneqq \|\mathbf{v}\|^2 - 2\mathbf{b}^\top\psi(\mathbf{v}) + \|\psi(\mathbf{v})\|^2. \tag{23}$$

We first bound Φ, which is purely a statement about the geometry. It has a leading term $\|\mathbf{v}\|^2$ and two corrections: $-2\mathbf{b}^\top\psi(\mathbf{v})$, whose sign is not determined, and $\|\psi(\mathbf{v})\|^2 \geq 0$. Each is a fraction of $\|\mathbf{v}\|^2$. For the cross term, Cauchy–Schwarz followed by the curvature bound (22) and then (T) with $r_0 \leq \rho/2$ gives

$$|\mathbf{b}^\top\psi(\mathbf{v})| \;\leq\; \tau\,\|\psi(\mathbf{v})\| \;\leq\; \tau\,\frac{\|\mathbf{v}\|^2}{\rho} \;\leq\; \frac{\rho}{8}\cdot\frac{\|\mathbf{v}\|^2}{\rho} \;=\; \frac{\|\mathbf{v}\|^2}{8}, \tag{24}$$

the reach cancelling in the last step. For the quadratic term, squaring (22),

$$\|\psi(\mathbf{v})\|^2 \;\leq\; \frac{\|\mathbf{v}\|^4}{\rho^2} \;=\; \|\mathbf{v}\|^2\Big(\frac{\|\mathbf{v}\|}{\rho}\Big)^2 \;\leq\; \frac{\|\mathbf{v}\|^2}{4} \qquad (\|\mathbf{v}\| \leq r_0 \leq \rho/2), \tag{25}$$

which is the only step that uses the restriction to the chart. The two corrections are now bounded *separately*, each in the direction unfavourable to the inequality being proved; since (24) controls $|\mathbf{b}^\top\psi(\mathbf{v})|$, it does so whatever the sign of $\mathbf{b}^\top\psi(\mathbf{v})$. For the lower bound the non-negative term is discarded, and for the upper bound it is kept and bounded by (25):

$$\tfrac{3}{4}\|\mathbf{v}\|^2 \;\leq\; \Phi(\mathbf{v}) \;\leq\; \tfrac{3}{2}\|\mathbf{v}\|^2 \qquad (\|\mathbf{v}\| \leq r_0), \tag{26}$$

the lower bound discarding $\|\psi(\mathbf{v})\|^2 \geq 0$ and using $-2\mathbf{b}^\top\psi(\mathbf{v}) \geq -2|\mathbf{b}^\top\psi(\mathbf{v})| \geq -\tfrac{1}{4}\|\mathbf{v}\|^2$, the upper bound using $-2\mathbf{b}^\top\psi(\mathbf{v}) \leq \tfrac{1}{4}\|\mathbf{v}\|^2$ together with (25). This completes the treatment of Φ.

We now turn this into a statement about the posterior, $\pi(\mathbf{x} \mid \mathbf{y}) \propto q(\mathbf{x})\, e^{-\|\mathbf{y}-\mathbf{x}\|^2/(2\sigma^2)}$ with respect to the volume measure on $\mathcal{M}$. Two preliminaries. First, $D\psi(\mathbf{v})$ maps into $\mathbf{N}$: writing $\mathbf{P_T}$ for the orthogonal projection onto $\mathbf{T}$, the identity $\mathbf{P_T}\psi(\mathbf{v}) = 0$ holds for every $\mathbf{v}$ because ψ takes values in $\mathbf{N}$, and differentiating it—$\mathbf{P_T}$ being a constant linear map—gives

$$\mathbf{P_T}\, D\psi(\mathbf{v}) = 0. \tag{27}$$

Second, fix an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_d$ of $\mathbf{T}$ and coordinatise $\mathbf{v}$ by it. Differentiating F from (20) in the ith coordinate gives $\mathbf{e}_i + D\psi(\mathbf{v})\mathbf{e}_i$, the term $\mathbf{v}$ contributing $\mathbf{e}_i$ and ψ contributing $D\psi(\mathbf{v})\mathbf{e}_i$. The volume factor is the square root of the determinant of the $d \times d$ matrix $G(\mathbf{v})$ of pairwise inner products of these d vectors, and

$$G_{ij} = \big(\mathbf{e}_i + D\psi\,\mathbf{e}_i\big)^\top\big(\mathbf{e}_j + D\psi\,\mathbf{e}_j\big) = \underbrace{\mathbf{e}_i^\top\mathbf{e}_j}_{=\,\delta_{ij}} + \underbrace{\mathbf{e}_i^\top D\psi\,\mathbf{e}_j + \big(D\psi\,\mathbf{e}_i\big)^\top\mathbf{e}_j}_{=\,0} + \big(D\psi\,\mathbf{e}_i\big)^\top\big(D\psi\,\mathbf{e}_j\big), \tag{28}$$

where $D\psi$ is evaluated at $\mathbf{v}$ throughout. Each of the two middle terms pairs a vector of $\mathbf{T}$ with a vector of $\mathbf{N}$ and so vanishes by (27). The final term does *not* vanish: both of its factors lie in $\mathbf{N}$, and $\mathbf{T} \perp \mathbf{N}$ says nothing about two vectors of $\mathbf{N}$. Hence $G(\mathbf{v}) = \mathbf{I} + D\psi^\top D\psi$ with $\mathbf{I}$ the $d \times d$ identity, and by the area formula

$$\int_{\mathcal{G}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) \;=\; \int_{\|\mathbf{v}\|<r_0} f\big(F(\mathbf{v})\big)\,J(\mathbf{v})\,\mathrm{d}\mathbf{v},\; J(\mathbf{v}) \;\coloneqq\; \det G(\mathbf{v})^{1/2} \;=\; \det\big(\mathbf{I} + D\psi^\top D\psi\big)^{1/2}, \tag{29}$$

holds for every μ-integrable f on $\mathcal{G}$, where μ is the d-dimensional volume measure on $\mathcal{M}$ with respect to which (B1) defines q. The formula requires the parameterisation to be injective, which it is—distinct $\mathbf{v}$ have distinct tangential offsets—and $\mathcal{G}$ is by definition its image on $\{\|\mathbf{v}\| < r_0\}$.

That J is computable from G alone is the Gram determinant identity. The differential carries the unit cube spanned by $\mathbf{e}_1, \ldots, \mathbf{e}_d$ to the parallelepiped spanned by the d vectors $\mathbf{e}_i + D\psi\,\mathbf{e}_i$, and the d-dimensional volume of a parallelepiped depends only on the lengths and angles of its edges, which is exactly the information in G. Concretely, if A collects the edges as columns and $A = QR$ with $Q^\top Q = \mathbf{I}$ and R upper triangular, then Q is an isometry onto the span, so the volume equals $|\det R|$, while $G = A^\top A = R^\top R$ gives $\det G = (\det R)^2$; hence the volume is $\det G^{1/2}$. For $d = 1$ this reduces to the length $\|\mathbf{e}_1 + D\psi\,\mathbf{e}_1\|$, and for $d = 2$ to the area of the parallelogram spanned by the two edges, as it should. Finally $G(\mathbf{v}) = \mathbf{I} + D\psi^\top D\psi \succeq \mathbf{I}$ gives $J(\mathbf{v}) \geq 1$: the graph is stretched relative to its tangent plane, never compressed. This is the lower bound on J used below.

Transporting the posterior. Let $g : \mathcal{M} \to \mathbb{R}$ be bounded and measurable. By definition of the posterior,

$$\mathbb{E}\big[g(\mathbf{x}) \mid \mathbf{y}\big] = \frac{\displaystyle\int_{\mathcal{M}} g(\mathbf{x})\, q(\mathbf{x})\, e^{-\|\mathbf{y}-\mathbf{x}\|^2/(2\sigma^2)}\,\mathrm{d}\mu(\mathbf{x})}{\displaystyle\int_{\mathcal{M}} q(\mathbf{x})\, e^{-\|\mathbf{y}-\mathbf{x}\|^2/(2\sigma^2)}\,\mathrm{d}\mu(\mathbf{x})}. \tag{30}$$

Split both integrals over $\mathcal{G}$ and $\mathcal{M} \setminus \mathcal{G}$. On $\mathcal{G}$, insert $\|\mathbf{y}-\mathbf{x}\|^2 = \tau^2 + \Phi(\mathbf{v})$ and apply (29) to $f(\mathbf{x}) = g(\mathbf{x})\, q(\mathbf{x})\, e^{-\|\mathbf{y}-\mathbf{x}\|^2/(2\sigma^2)}$; the numerator's $\mathcal{G}$ part becomes $e^{-\tau^2/(2\sigma^2)} \int_{\|\mathbf{v}\|<r_0} g(F(\mathbf{v}))\, w(\mathbf{v})\, \mathrm{d}\mathbf{v}$, where

$$w(\mathbf{v}) \coloneqq q\big(F(\mathbf{v})\big)\, e^{-\Phi(\mathbf{v})/(2\sigma^2)}\, J(\mathbf{v}). \tag{31}$$

The factor $e^{-\tau^2/(2\sigma^2)}$ involves neither $\mathbf{x}$ nor g, so it comes out of the remaining piece as well. With

$$R_g \coloneqq e^{\tau^2/(2\sigma^2)} \int_{\mathcal{M}\setminus\mathcal{G}} g(\mathbf{x})\, q(\mathbf{x})\, e^{-\|\mathbf{y}-\mathbf{x}\|^2/(2\sigma^2)}\, \mathrm{d}\mu(\mathbf{x}), \tag{32}$$

the numerator of (30) equals $e^{-\tau^2/(2\sigma^2)}\big(\int_{\|\mathbf{v}\|<r_0} g(F(\mathbf{v}))\, w(\mathbf{v})\, \mathrm{d}\mathbf{v} + R_g\big)$, and the denominator is the same expression with $g \equiv 1$. The common factor cancels:

$$\mathbb{E}\big[g(\mathbf{x}) \mid \mathbf{y}\big] = \frac{1}{Z}\left(\int_{\|\mathbf{v}\|<r_0} g\big(F(\mathbf{v})\big)\, w(\mathbf{v})\, \mathrm{d}\mathbf{v} + R_g\right), \qquad Z \coloneqq \int_{\|\mathbf{v}\|<r_0} w(\mathbf{v})\, \mathrm{d}\mathbf{v} + R_1. \tag{33}$$

Here R_g is the posterior mass on $\mathcal{M} \setminus \mathcal{G}$ rescaled by the same factor and R_1 its contribution to the normalizer; both are bounded in (40).

Bounding w. Of the three factors in (31), the first lies in $[q_{\min}, q_{\max}]$ by (B1) and the second obeys (26). For the third, note that $D\psi(\mathbf{v})$ is a linear map $\mathbf{T} \to \mathbf{N}$, that is a $(D-d) \times d$ matrix, so it need not be square and need not have eigenvalues; the relevant eigenvalues are those of the $d \times d$ symmetric matrix $D\psi^\top D\psi$, which we write as $s_1^2 \geq \cdots \geq s_d^2 \geq 0$. Since $D\psi^\top D\psi \succeq 0$ we have $\mathbf{I} + D\psi^\top D\psi \succeq \mathbf{I}$, so each $1 + s_i^2 \geq 1$ and $J(\mathbf{v}) = \det(\mathbf{I} + D\psi^\top D\psi)^{1/2} \geq 1$. For an upper bound, the operator norm $\|D\psi(\mathbf{v})\|_{\mathrm{op}} \coloneqq \max_{\|\mathbf{h}\|=1} \|D\psi(\mathbf{v})\mathbf{h}\|$—the largest factor by which $D\psi(\mathbf{v})$ stretches a tangent vector—equals s_1, so $s_i \leq \|D\psi(\mathbf{v})\|_{\mathrm{op}}$ for every i. Since $D\psi(\mathbf{0}) = \mathbf{0}$ and $D\psi$ is Lipschitz with constant c_ψ/ρ by fact (iii),

$$\|D\psi(\mathbf{v})\|_{\mathrm{op}} = \big\|D\psi(\mathbf{v}) - D\psi(\mathbf{0})\big\|_{\mathrm{op}} \;\leq\; \frac{c_\psi}{\rho}\, \|\mathbf{v}\|, \tag{34}$$

Lipschitz continuity together with vanishing at the origin giving linear growth; this is the second bound in (22). On the chart $\|\mathbf{v}\| \leq r_0 \leq \rho/2$, so this already caps the stretching:

$$\|D\psi(\mathbf{v})\|_{\mathrm{op}} \;\leq\; \frac{c_\psi \|\mathbf{v}\|}{\rho} \;\leq\; \frac{c_\psi}{2}, \tag{35}$$

whence every $s_i \leq c_\psi/2$ by (35), and

$$J(\mathbf{v})^2 \;=\; \prod_{i=1}^{d} \big(1 + s_i^2\big) \;\leq\; \Big(1 + \frac{c_\psi^2}{4}\Big)^d, \qquad \text{so} \qquad 1 \;\leq\; J(\mathbf{v}) \;\leq\; K \;\coloneqq\; \Big(1 + \frac{c_\psi^2}{4}\Big)^{d/2}. \tag{36}$$

The point is that K is a *constant*: because the chart already caps $\|\mathbf{v}\|$ at $r_0 \leq \rho/2$, no condition relating σ to ρ is needed to control J. The exponent d is the number of factors in the product, and K is absorbed into c_2 and c_4 along with c_ψ. Multiplying the three factors,

$$q_{\min} e^{-3\|\mathbf{v}\|^2/(4\sigma^2)} \;\leq\; w(\mathbf{v}) \;\leq\; q_{\max} K\, e^{-3\|\mathbf{v}\|^2/(8\sigma^2)} \qquad (\|\mathbf{v}\| \leq r_0), \tag{37}$$

the lower bound using $J \geq 1$ with the *right* half of (26), and the upper bound $J \leq K$ from (36) with the *left* half. The two halves swap roles because e^{-t} is decreasing: the largest admissible Φ, namely $\frac{3}{2}\|\mathbf{v}\|^2$, produces the smallest exponential and hence the *lower* bound on w, while the smallest, $\frac{3}{4}\|\mathbf{v}\|^2$, produces the upper one. So w lies between two isotropic Gaussians of width of order σ, differing only in constants: this is the precise sense in which the posterior concentrates at that scale.

Bounding Z. By (B3), $\sigma \leq r_0$, so the ball $\|\mathbf{v}\| \leq \sigma$ lies inside the chart. For the lower bound, $w \geq 0$ everywhere and $R_1 \geq 0$, so both may be discarded outside that ball; on it $\|\mathbf{v}\|^2/\sigma^2 \leq 1$, so the left inequality of (37) gives $w(\mathbf{v}) \geq q_{\min} e^{-3\|\mathbf{v}\|^2/(4\sigma^2)} \geq q_{\min} e^{-3/4}$, and the remaining integral is the volume of a ball of radius σ, namely $\omega_d \sigma^d$ with ω_d the volume of the unit d-ball:

$$Z \;\geq\; \int_{\|\mathbf{v}\|\leq\sigma} w(\mathbf{v})\, \mathrm{d}\mathbf{v} \;\geq\; q_{\min} e^{-3/4} \int_{\|\mathbf{v}\|\leq\sigma} \mathrm{d}\mathbf{v} \;=\; q_{\min} e^{-3/4}\, \omega_d\, \sigma^d. \tag{38}$$

For the upper bound the same non-negativity lets us *enlarge* the domain from $\|\mathbf{v}\| < r_0$ to all of $\mathbf{T}$, after which the right inequality of (37) leaves a Gaussian integral; with $\int_{\mathbb{R}^d} e^{-a\|\mathbf{v}\|^2}\mathrm{d}\mathbf{v} = (\pi/a)^{d/2}$ at $a = 3/(8\sigma^2)$ this is $(8\pi\sigma^2/3)^{d/2}$, so

$$Z \;\le\; q_{\max} K \int_{\mathbf{T}} e^{-3\|\mathbf{v}\|^2/(8\sigma^2)}\,\mathrm{d}\mathbf{v} + R_1 \;=\; q_{\max} K\,(8\pi\sigma^2/3)^{d/2} + R_1. \tag{39}$$

Both directions are needed because the expectations in (33) are ratios: the lower bound on Z is what bounds $\mathbb{E}\|\mathbf{v}\|^4$ from above, and the upper bound on Z is what bounds the variance from below. Note $Z \asymp \sigma^d$ either way, which is the volume of the region the posterior occupies.

Bounding the remainders. Recall R_g from (32). Since $\mathcal{M} \cap B(\mathbf{x}^*, r_0) \subseteq \mathcal{G}$ by (21), every $\mathbf{x} \in \mathcal{M} \setminus \mathcal{G}$ has $\|\mathbf{x} - \mathbf{x}^*\| \ge r_0$, so $\|\mathbf{y} - \mathbf{x}\| \ge \|\mathbf{x} - \mathbf{x}^*\| - \|\mathbf{y} - \mathbf{x}^*\| \ge r_0 - \tau \ge 3r_0/4$ by (T), giving $e^{-\|\mathbf{y}-\mathbf{x}\|^2/(2\sigma^2)} \le e^{-9r_0^2/(32\sigma^2)}$. The prefactor in (32) obeys $e^{\tau^2/(2\sigma^2)} \le e^{r_0^2/(32\sigma^2)}$, again by (T), and bounding q by $q_{\max}$ and the remaining integral by $\mathrm{vol}(\mathcal{M})$,

$$|R_g| \;\le\; \sup|g| \cdot q_{\max}\,\mathrm{vol}(\mathcal{M})\, e^{r_0^2/(32\sigma^2)}\, e^{-9r_0^2/(32\sigma^2)} \;=\; \sup|g| \cdot q_{\max}\,\mathrm{vol}(\mathcal{M})\, e^{-r_0^2/(4\sigma^2)}. \tag{40}$$

The rescaling therefore costs a factor $e^{r_0^2/(32\sigma^2)}$, leaving $8/9$ of the exponent.

The fourth moment. Apply (33) with $g(\mathbf{x}) = \|\mathbf{P}_{\mathbf{T}}(\mathbf{x} - \mathbf{x}^*)\|^4$, which equals $\|\mathbf{v}\|^4$ on the chart, is defined on all of $\mathcal{M}$, and satisfies $\sup|g| \le \mathrm{diam}(\mathcal{M})^4$. To evaluate the numerator of (33) on the chart, use the right inequality of (37), enlarge the domain to $\mathbf{T}$, and integrate with $\int_{\mathbb{R}^d} \|\mathbf{v}\|^4 e^{-a\|\mathbf{v}\|^2}\mathrm{d}\mathbf{v} = (\pi/a)^{d/2} d(d+2)/(4a^2)$ at $a = 3/(8\sigma^2)$, where $4a^2 = 9/(16\sigma^4)$ and so $d(d+2)/(4a^2) = \frac{16}{9} d(d+2)\sigma^4$:

$$\int_{\|\mathbf{v}\|<r_0} \|\mathbf{v}\|^4\, w(\mathbf{v})\,\mathrm{d}\mathbf{v} \;\le\; q_{\max} K\, \tfrac{16}{9} d(d+2)\,\sigma^4\,(8\pi\sigma^2/3)^{d/2}. \tag{41}$$

Dividing by (38) and writing $(8\pi\sigma^2/3)^{d/2} = (8\pi/3)^{d/2}\sigma^d$, so that the σ^d cancels, the chart part of $\mathbb{E}[\|\mathbf{v}\|^4 \mid \mathbf{y}]$ obeys

$$\frac{1}{Z}\int_{\|\mathbf{v}\|<r_0} \|\mathbf{v}\|^4\, w(\mathbf{v})\,\mathrm{d}\mathbf{v} \;\le\; \frac{q_{\max} K}{q_{\min}}\, e^{3/4}\, \frac{(8\pi/3)^{d/2}}{\omega_d} \cdot \tfrac{16}{9} d(d+2)\,\sigma^4, \tag{42}$$

a constant multiple of σ^4 depending only on d, $q_{\max}/q_{\min}$ and c_ψ through K. What remains of (33) is R_g/Z, the part of the posterior carried by $\mathcal{M} \setminus \mathcal{G}$, where $\mathbf{v}$ is not even defined. By (40) and (38) it is at most $c_R\,\sigma^{-d} e^{-r_0^2/(4\sigma^2)}$, where

$$c_R \;:=\; \frac{\mathrm{diam}(\mathcal{M})^4\, q_{\max}\,\mathrm{vol}(\mathcal{M})}{q_{\min}\, e^{-3/4}\,\omega_d}$$

is set by d, $q_{\max}/q_{\min}$, $\mathrm{diam}(\mathcal{M})/\rho$ and $\mathrm{vol}(\mathcal{M})/\rho^d$. This term is also bounded by a constant multiple of σ^4, and the bound holds at *every* noise level, not merely for small σ. Substituting $u = r_0^2/(4\sigma^2)$, so that $\sigma = r_0/(2\sqrt{u})$,

$$\sigma^{-(d+4)} e^{-r_0^2/(4\sigma^2)} \;=\; \Big(\frac{2}{r_0}\Big)^{d+4} u^{(d+4)/2} e^{-u}. \tag{43}$$

The u-dependent factor is bounded by its largest value: for any $\alpha > 0$ the function $t \mapsto t^\alpha e^{-t}$ is maximised on $(0,\infty)$ at $t = \alpha$, because $\alpha \log t - t$ is, so

$$t^\alpha e^{-t} \;\le\; \alpha^\alpha e^{-\alpha} \qquad (t > 0). \tag{44}$$

Applying (44) with $\alpha = (d+4)/2$ to (43),

$$\frac{R_g}{Z} \;\le\; c_R \Big(\frac{2}{r_0}\Big)^{d+4} \Big(\frac{d+4}{2}\Big)^{(d+4)/2} e^{-(d+4)/2}\,\sigma^4 \qquad \text{for every } \sigma > 0. \tag{45}$$

The bound (45) is unconditional in σ: the inequality applied to (43) holds at every u, so (B3) is not used here—it is needed only for (38). The constant in (45) may be very large. Each step producing it is crude: in (40) the integrand is replaced by its largest value and q by $q_{\max}$, while (38) retains

only a ball of radius σ. Only the σ^4 scaling is claimed, and that is all (46) requires. Adding the two contributions,

$$\mathbb{E}\big[\|\mathbf{v}\|^4 \mid \mathbf{y}\big] \;\le\; c_4\,\sigma^4, \tag{46}$$

with c_4 the sum of the constants in (42) and (45), which is the first claim of the lemma.

The tangential variance. Fix a unit $\mathbf{u} \in \mathbf{T}$ and $m \in \mathbb{R}$, and apply (33) with $g(\mathbf{x}) = (\mathbf{u}^\top\mathbf{v} - m)^2$. Here $R_g \ge 0$, so the numerator is at least its chart part; restricting to $\|\mathbf{v}\| \le \sigma$ and using the left inequality of (37) as in (38),

$$\int_{\|\mathbf{v}\|<r_0} (\mathbf{u}^\top\mathbf{v} - m)^2 w(\mathbf{v})\,\mathrm{d}\mathbf{v} \;\ge\; q_{\min} e^{-3/4} \int_{\|\mathbf{v}\|\le\sigma} (\mathbf{u}^\top\mathbf{v} - m)^2\,\mathrm{d}\mathbf{v}. \tag{47}$$

The last integral is smallest at $m = 0$, since $\mathbf{u}^\top\mathbf{v}$ has mean zero over the ball, and there it equals $\omega_d\sigma^{d+2}/(d+2)$, because $\int_{\|\mathbf{v}\|\le\sigma}(\mathbf{u}^\top\mathbf{v})^2\mathrm{d}\mathbf{v} = \frac{1}{d}\int_{\|\mathbf{v}\|\le\sigma}\|\mathbf{v}\|^2\mathrm{d}\mathbf{v}$ by symmetry and the latter is $d\,\omega_d\sigma^{d+2}/(d+2)$.

It remains to divide by Z, and this is the only step at which R_1 matters: a lower bound on a ratio needs the denominator bounded *above*, and R_1 enters that upper bound, (39), with a positive sign, so it cannot be dropped. Neither is it negligible. Write $Z_0 := q_{\max}K(8\pi/3)^{d/2}\sigma^d$ for the Gaussian part of (39). Taking $\sup|g| = 1$ in (40) bounds R_1, and $q_{\max}$ cancels, leaving

$$\frac{R_1}{Z_0} \;\le\; \frac{\mathrm{vol}(\mathcal{M})}{K(8\pi/3)^{d/2}}\,\sigma^{-d}e^{-r_0^2/(4\sigma^2)} \;\le\; \frac{\mathrm{vol}(\mathcal{M})}{K(8\pi/3)^{d/2}}\Big(\frac{2}{r_0}\Big)^d\Big(\frac{d}{2}\Big)^{d/2}e^{-d/2} \;=:\; C_1, \tag{48}$$

the second step being (44) with $\alpha = d/2$, after the same substitution $u = r_0^2/(4\sigma^2)$ used in (43). As there, the inequality holds at every u, so the bound (48) is valid at every σ. But C_1 is merely finite, not small: the quantity $\sigma^{-d}e^{-r_0^2/(4\sigma^2)}$ attains its maximum at $\sigma = r_0/\sqrt{2d}$, an interior point, where R_1 may exceed Z_0. The remainder is therefore *retained*, not discarded, as the factor $Z \le Z_0(1 + C_1)$. Dividing (47) by it,

$$\mathbb{E}\big[(\mathbf{u}^\top\mathbf{v} - m)^2 \mid \mathbf{y}\big] \;\ge\; \frac{1}{1+C_1}\,\frac{q_{\min}}{q_{\max}K}\,\frac{e^{-3/4}\,\omega_d}{(d+2)\,(8\pi/3)^{d/2}}\,\sigma^2, \tag{49}$$

uniformly in m. Taking the minimising m, which is what the variance is, gives $\mathrm{Var}[\mathbf{u}^\top\mathbf{v} \mid \mathbf{y}] \ge c_2\sigma^2$. Both c_2 and c_4 therefore depend only on d, $q_{\max}/q_{\min}$, $\mathrm{diam}(\mathcal{M})/\rho$, $\mathrm{vol}(\mathcal{M})/\rho^d$, r_0/ρ and the absolute constant c_ψ, as the statement claims. □

Lemma 6 (Normal-direction posterior variance)**.** *Under (B1)–(B3) and (T), for every unit* $\mathbf{u} \in \mathbf{N}$, $\mathrm{Var}[\mathbf{u}^\top\mathbf{x} \mid \mathbf{y}] \le c_4\sigma^4/\rho^2$.

Proof. Reduction to a fixed centre. For a scalar random variable X and any $a \in \mathbb{R}$, $\mathbb{E}[(X-a)^2 \mid \mathbf{y}] = \mathrm{Var}[X \mid \mathbf{y}] + (\mathbb{E}[X \mid \mathbf{y}] - a)^2$, so every choice of a gives an upper bound on the variance, with equality at the posterior mean. Taking $X = \mathbf{u}^\top\mathbf{x}$ and $a = \mathbf{u}^\top\mathbf{x}^*$,

$$\mathrm{Var}[\mathbf{u}^\top\mathbf{x} \mid \mathbf{y}] \;\le\; \mathbb{E}\big[(\mathbf{u}^\top(\mathbf{x} - \mathbf{x}^*))^2 \mid \mathbf{y}\big]. \tag{50}$$

We know nothing about the posterior mean, whereas the chart is centred at $\mathbf{x}^*$; this is why the comparison point is moved there.

Splitting. Apply (33) to the right-hand side of (50), with $g(\mathbf{x}) = (\mathbf{u}^\top(\mathbf{x} - \mathbf{x}^*))^2$. This g is non-negative, and $|\mathbf{u}^\top(\mathbf{x} - \mathbf{x}^*)| \le \|\mathbf{x} - \mathbf{x}^*\| \le \mathrm{diam}(\mathcal{M})$ by Cauchy–Schwarz and $\|\mathbf{u}\| = 1$, so $\sup|g| \le \mathrm{diam}(\mathcal{M})^2$, which is the quantity that (40) needs. The expectation is then a chart part plus R_g/Z, and each is bounded in turn.

The chart part. On the chart $\mathbf{x} = F(\mathbf{v})$ by (20), so $\mathbf{x} - \mathbf{x}^* = \mathbf{v} + \psi(\mathbf{v})$. Since $\mathbf{v} \in \mathbf{T}$ and $\mathbf{u} \in \mathbf{N}$ we have $\mathbf{u}^\top\mathbf{v} = 0$, so only the normal component survives:

$$\mathbf{u}^\top(\mathbf{x} - \mathbf{x}^*) \;=\; \mathbf{u}^\top\psi(\mathbf{v}), \qquad \text{whence} \qquad \big|\mathbf{u}^\top(\mathbf{x} - \mathbf{x}^*)\big| \;\le\; \|\psi(\mathbf{v})\| \;\le\; \frac{\|\mathbf{v}\|^2}{\rho}, \tag{51}$$

the first inequality by Cauchy–Schwarz with $\|\mathbf{u}\| = 1$ and the second by (22). This is the one place curvature enters: the normal displacement is *quadratic* in the tangential one, so squaring (51) gives $g(F(\mathbf{v})) \le \|\mathbf{v}\|^4/\rho^2$ and the chart part is at most

$$\frac{1}{\rho^2}\cdot\frac{1}{Z}\int_{\|\mathbf{v}\|<r_0}\|\mathbf{v}\|^4\,w(\mathbf{v})\,\mathrm{d}\mathbf{v} \;\le\; \frac{1}{\rho^2}\,\mathbb{E}\big[\|\mathbf{v}\|^4 \mid \mathbf{y}\big] \;\le\; \frac{c_4\,\sigma^4}{\rho^2}, \tag{52}$$

the first inequality because by (33) the two sides differ by the non-negative remainder attached to $\|\mathbf{P}_{\mathbf{T}}(\mathbf{x}-\mathbf{x}^*)\|^4$, and the second by Lemma 5. So the fourth moment of Lemma 5, divided by ρ^2, is exactly what controls this lemma—the two powers of ρ being the price of the quadratic in (51).

The remainder. Off the chart the parameterisation is unavailable and $\mathbf{v}$ is undefined, so we use only $|\mathbf{u}^\top(\mathbf{x}-\mathbf{x}^*)| \le \operatorname{diam}(\mathcal{M})$. By (40) the remainder then contributes at most $\operatorname{diam}(\mathcal{M})^2 q_{\max}\operatorname{vol}(\mathcal{M})e^{-r_0^2/(4\sigma^2)}/Z$. This is the quantity bounded in (45) with $\operatorname{diam}(\mathcal{M})^4$ replaced by $\operatorname{diam}(\mathcal{M})^2$, so dividing by (38) and applying (44) with $\alpha=(d+4)/2$ exactly as there bounds it by

$$\begin{aligned}&\frac{c_R}{\operatorname{diam}(\mathcal{M})^2}\Big(\frac{2}{r_0}\Big)^{d+4}\Big(\frac{d+4}{2}\Big)^{(d+4)/2}e^{-(d+4)/2}\,\sigma^4\\ &\quad=\left[\rho^2\,\frac{c_R}{\operatorname{diam}(\mathcal{M})^2}\Big(\frac{2}{r_0}\Big)^{d+4}\Big(\frac{d+4}{2}\Big)^{(d+4)/2}e^{-(d+4)/2}\right]\frac{\sigma^4}{\rho^2}\qquad\text{for every }\sigma>0.\end{aligned}\tag{53}$$

The bracket is dimensionless and is fixed by d, $q_{\max}/q_{\min}$ and the ratios $\operatorname{diam}(\mathcal{M})/\rho$, $\operatorname{vol}(\mathcal{M})/\rho^d$ and r_0/ρ, so it is absorbed into c_4. As after (48), this term is *not* exponentially small—$\sigma^{-d}e^{-r_0^2/(4\sigma^2)}$ peaks at $\sigma=r_0/\sqrt{2d}$—but (53) bounds it by a constant multiple of σ^4/ρ^2 at every noise level, which is all the lemma claims.

Combining. By (52) and (53) both contributions are constant multiples of σ^4/ρ^2 at every σ, so their sum is one too, which is the claim. Throughout, c_4 denotes a single constant taken large enough for every bound in which it appears—here the sum of the constant of Lemma 5 and the bracket in (53). This is legitimate because finitely many such bounds occur and each constant depends only on d, $q_{\max}/q_{\min}$, $\operatorname{diam}(\mathcal{M})/\rho$, $\operatorname{vol}(\mathcal{M})/\rho^d$, r_0/ρ and c_ψ. □

Proposition 7 (Near-degeneracy on a manifold)**.** *Under (B1)–(B3) and (T), the eigenvalues $\nu_1\ge\cdots\ge\nu_D$ of $\operatorname{Cov}[\mathbf{x}\mid\mathbf{y}]$ satisfy $\nu_{d+1}\le\varepsilon:=c_4\sigma^4/\rho^2$ and $\nu_d\ge c_2\sigma^2$. Consequently the eigenvalues of $\mathbf{H}_{\mathbf{y}}$ obey*

$$\lambda_i(\mathbf{H}_{\mathbf{y}})\in\Big[-\tfrac{1}{\sigma^2},\,-\tfrac{1}{\sigma^2}+\eta\Big]\;\;(i>d),\qquad\lambda_j(\mathbf{H}_{\mathbf{y}})\ge-\tfrac{1}{\sigma^2}+\tfrac{c_2}{\sigma^2}\;\;(j\le d),\qquad\eta:=\frac{c_4}{\rho^2}.\tag{54}$$

These two ranges are disjoint—so that the spectrum genuinely separates into a noise group and a signal group—exactly when $\eta<c_2/\sigma^2$, that is when

$$\sigma^2\;<\;\frac{c_2}{c_4}\,\rho^2.\tag{55}$$

Proof. Quadratic forms are variances. Writing $\boldsymbol{\mu}:=\mathbb{E}[\mathbf{x}\mid\mathbf{y}]$, for any unit $\mathbf{u}$,

$$\mathbf{u}^\top\operatorname{Cov}[\mathbf{x}\mid\mathbf{y}]\,\mathbf{u}\;=\;\mathbb{E}\big[\mathbf{u}^\top(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{u}\mid\mathbf{y}\big]\;=\;\operatorname{Var}[\mathbf{u}^\top\mathbf{x}\mid\mathbf{y}],\tag{56}$$

so Lemmas 5 and 6 bound exactly the quantities that Courant–Fischer reads. We use both of its forms: for a symmetric matrix with eigenvalues $\nu_1\ge\cdots\ge\nu_D$,

$$\nu_k=\min_{\dim V=k-1}\;\max_{\|\mathbf{u}\|=1,\;\mathbf{u}\perp V}\mathbf{u}^\top\operatorname{Cov}[\mathbf{x}\mid\mathbf{y}]\,\mathbf{u}\;\text{ and }\;\nu_k=\max_{\dim W=k}\;\min_{\|\mathbf{u}\|=1,\;\mathbf{u}\in W}\mathbf{u}^\top\operatorname{Cov}[\mathbf{x}\mid\mathbf{y}]\,\mathbf{u}.\tag{57}$$

Each form is used once, with the *same* subspace $\mathbf{T}$ in both.

The trailing eigenvalues. Take the first form of (57) at $k=d+1$ and choose $V=\mathbf{T}$, which is admissible since $\dim\mathbf{T}=d$. Because $\mathbb{R}^D=\mathbf{T}\oplus\mathbf{N}$ orthogonally, the constraint $\mathbf{u}\perp\mathbf{T}$ means precisely $\mathbf{u}\in\mathbf{N}$, so the inner maximum runs over the unit vectors of $\mathbf{N}$ and

$$\nu_{d+1}\;\le\;\max_{\|\mathbf{u}\|=1,\;\mathbf{u}\in\mathbf{N}}\operatorname{Var}[\mathbf{u}^\top\mathbf{x}\mid\mathbf{y}]\;\le\;\frac{c_4\sigma^4}{\rho^2}\;=\;\varepsilon$$

by (56) and Lemma 6. Since the eigenvalues are ordered, $\nu_i\le\nu_{d+1}\le\varepsilon$ for every $i>d$. Note the direction of the choice: fixing V can only *increase* a minimum, so this is an upper bound, which is what a bound on the trailing eigenvalues requires.

The leading eigenvalues. Lemma 5 concerns $\mathbf{v}$, so it must first be transferred to $\mathbf{x}$. Fix a unit $\mathbf{u}\in\mathbf{T}$ and $m\in\mathbb{R}$, and apply (33) with $g(\mathbf{x})=(\mathbf{u}^\top\mathbf{x}-m)^2$. On the chart $\mathbf{x}=F(\mathbf{v})$, and $\psi(\mathbf{v})\in\mathbf{N}$ with

$\mathbf{u} \in \mathbf{T}$ gives $\mathbf{u}^\top \psi(\mathbf{v}) = 0$, so $\mathbf{u}^\top \mathbf{x} = \mathbf{u}^\top \mathbf{x}^* + \mathbf{u}^\top \mathbf{v}$: the two differ by the constant $\mathbf{u}^\top \mathbf{x}^*$. Hence $g(F(\mathbf{v})) = (\mathbf{u}^\top \mathbf{v} - m')^2$ with $m' \coloneqq m - \mathbf{u}^\top \mathbf{x}^*$. Since $g \geq 0$ we have $R_g \geq 0$, so dropping it—safe here, because it is being removed from a quantity we bound from *below*—gives

$$\mathbb{E}\big[(\mathbf{u}^\top \mathbf{x} - m)^2 \mid \mathbf{y}\big] \;\geq\; \frac{1}{Z} \int_{\|\mathbf{v}\| < r_0} (\mathbf{u}^\top \mathbf{v} - m')^2 \, w(\mathbf{v}) \, \mathrm{d}\mathbf{v} \;\geq\; c_2 \sigma^2,$$

the last inequality being what the tangential-variance step of Lemma 5 established, uniformly in m'. As m ranges over $\mathbb{R}$ so does m', so the bound holds for every m; minimising over m and using (50) with equality at the mean gives $\mathrm{Var}[\mathbf{u}^\top \mathbf{x} \mid \mathbf{y}] \geq c_2 \sigma^2$ for every unit $\mathbf{u} \in \mathbf{T}$.

Now take the second form of (57) at $k = d$ and choose $W = \mathbf{T}$, again admissible because $\dim \mathbf{T} = d$:

$$\nu_d \;\geq\; \min_{\|\mathbf{u}\|=1,\ \mathbf{u} \in \mathbf{T}} \mathrm{Var}[\mathbf{u}^\top \mathbf{x} \mid \mathbf{y}] \;\geq\; c_2 \sigma^2.$$

Fixing W can only *decrease* a maximum, so this is a lower bound, as needed. Both halves of the spectrum are thus obtained from the single theorem (57), in its two forms and in opposite directions, with $\mathbf{T}$ as the trial subspace each time.

Substituting into the exact correspondence (6), $\lambda_i = (\nu_i - \sigma^2)/\sigma^4$, yields (54). Since this is an increasing function of ν_i, bounds on ν carry over to bounds on λ in the same direction. For $i > d$ we have $\nu_i \in [0, \varepsilon]$, the lower end because $\mathrm{Cov}[\mathbf{x}|\mathbf{y}] \succeq 0$, giving $\lambda_i \in [-1/\sigma^2,\ -1/\sigma^2 + \varepsilon/\sigma^4]$ with $\varepsilon/\sigma^4 = c_4/\rho^2 = \eta$ —the cancellation of σ^4 that makes the displacement independent of σ; and for $j \leq d$, $\nu_j \geq \nu_d \geq c_2\sigma^2$ gives $\lambda_j \geq (c_2\sigma^2 - \sigma^2)/\sigma^4 = -1/\sigma^2 + c_2/\sigma^2$. The right end of the first range is $-1/\sigma^2 + \eta$ and the left end of the second is $-1/\sigma^2 + c_2/\sigma^2$, so the two fail to overlap precisely when $\eta < c_2/\sigma^2$, that is $c_4/\rho^2 < c_2/\sigma^2$, which rearranges to (55). □

At the level of the trace this recovers a known fact: the minimum mean-square error of a d-dimensional signal in $\mathbb{R}^D$ is $d\sigma^2 + o(\sigma^2)$ (Wu & Verdú, 2011), so the trailing eigenvalues must sum to $o(\sigma^2)$. Proposition 7 refines this to a bound on each of them separately, with an explicit rate in σ/ρ, which is what the estimator needs. So the cluster is a faithful record of σ^2 precisely when the noise scale is small relative to the manifold's radius of curvature, with $\eta \to 0$ for a linear subspace. The quantitative form of "small" is $\eta\sigma^2 = c_4(\sigma/\rho)^2 < 1$, which is what places the whole interval (54) to the left of the origin; only then is $-1/\lambda_{\min}$ a positive quantity at all, and the relative excess $\eta\sigma^2/(1 - \eta\sigma^2)$ of Proposition 1 finite. Because c_4 is generous, this is a stronger requirement than $\sigma \ll \rho$. The separation gap c_2/σ^2 meanwhile *grows* as σ decreases, so signal and noise groups become easier to distinguish at low noise even as the curvature bias improves.

C Neural Tangent Kernel Model of the Jacobian Perturbation

Notation. This appendix keeps the component notation in which the derivation was carried out, which maps to the main text as follows: $\nabla \mathbf{f}_0(\mathbf{x}_j)$ is the true score Jacobian $\mathbf{H}_\mathbf{y}$, $\mathbf{J}(\mathbf{x}_j; \theta_\infty)$ the learned Jacobian $\hat{\mathbf{J}}$, $\partial r_m / \partial x_k$ the entries of $\nabla \mathbf{r}$, and $\Delta_{mk} = \partial g_m / \partial x_k|_{\mathbf{x}_j}$ the entries of $\mathbf{\Delta}$.

C.1 Neural Tangent Kernel Theory

Consider a general neural network $\mathbf{f}(\mathbf{x}; \theta) : \mathbb{R}^D \to \mathbb{R}^D$ satisfying Setup 5.2 in (Yang & Littwin, 2021), trained on N input–output pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ with inputs $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and targets $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$. In the infinite-width limit, the network output at initialization converges to a Gaussian process $\mathbf{f}(\cdot; \theta_0) \xrightarrow{d} \mathcal{GP}(\mathbf{0}, K)$ with NNGP kernel K (Neal, 2012; Lee et al., 2018), and training proceeds via kernel gradient descent with a fixed neural tangent kernel (NTK) $\mathbf{\Theta}^*$ (Jacot et al., 2018; Lee et al., 2019; Yang & Littwin, 2021). Both K and $\mathbf{\Theta}^*$ are determined by the architecture; these results hold for MLPs, CNNs, and U-Nets (Novak et al., 2019; Tachella et al., 2021). Under MSE loss, the trained output at convergence is (Lee et al., 2019):

$$\hat{\mathbf{f}}_\infty(\mathbf{x}) = \hat{\mathbf{f}}_0(\mathbf{x}) + \mathbf{\Theta}^*(\mathbf{x}, \mathbf{X})\, \mathbf{\Theta}^*(\mathbf{X}, \mathbf{X})^{-1} (\mathbf{Y} - \hat{\mathbf{f}}_0(\mathbf{X})), \tag{58}$$

where $\hat{\mathbf{f}}_0$ is the random output at initialization. Since (58) is linear in the GP initialization, $\hat{\mathbf{f}}_\infty(\mathbf{x})$ is Gaussian with mean and covariance:

$$\mathbb{E}[\hat{\mathbf{f}}_\infty(\mathbf{x})] = \mathbf{\Theta}^*(\mathbf{x}, \mathbf{X})\, \mathbf{\Theta}^*(\mathbf{X}, \mathbf{X})^{-1}\, \mathbf{Y}, \qquad \mathrm{Cov}[\hat{\mathbf{f}}_\infty(\mathbf{x})] = \tilde{K}(\mathbf{x}, \mathbf{x}), \tag{59}$$

where the residual kernel $\tilde{K}$ is a fixed quadratic expression in K and $\mathbf{\Theta}^*$ (given explicitly in Appendix C.4) and captures the initialization randomness that persists after training. Existence and smoothness of $\tilde{K}$ are all that the derivation below uses; its explicit form is needed once, in Appendix C.4, to show that $\tilde{K}$ vanishes at the training points while its mixed second derivative does not. In our self-supervised setting, the score network is trained and tested on the same noisy observations, so $\mathbf{x}$ coincides with one of the training points $\mathbf{x}_j$.

Applicability to score-matching losses. Equation (58) is stated for squared loss with *fixed* targets, whereas denoising score matching (Vincent, 2011) regresses on the random target $-\boldsymbol{\varepsilon}/\sigma_{\mathrm{dsm}}$, redrawn at every step. The two agree because for any random target $\mathbf{Y}$ the squared loss decomposes as

$$\mathbb{E}\big\|\mathbf{f}(\mathbf{x}) - \mathbf{Y}\big\|^2 = \big\|\mathbf{f}(\mathbf{x}) - \mathbb{E}[\mathbf{Y} \mid \mathbf{x}]\big\|^2 + \mathrm{trCov}[\mathbf{Y} \mid \mathbf{x}], \tag{60}$$

whose second term does not depend on θ. The decomposition therefore identifies the two objectives' *minimiser*: Minimizing the DSM objective is therefore minimizing a squared loss against the *conditional mean* of the target, and (58) applies with $\mathbf{Y}$ replaced by $\mathbb{E}[\mathbf{Y} \mid \mathbf{x}]$ at the optimum. For DSM applied to already-noisy observations that conditional mean is $\nabla \log p_{\sigma_{\mathrm{eff}}}$ with $\sigma^2_{\mathrm{eff}} = \sigma^2 + \sigma^2_{\mathrm{dsm}}$, which is why the noise level recovered from the spectrum is σ^2_{eff} and the known offset σ^2_{dsm} must be subtracted; since σ_{dsm} is a hyperparameter we choose, this correction requires no knowledge of σ and is compatible with blind operation. Algorithm 1 omits the correction: at the σ_{dsm} used throughout, the offset is at most 0.25% of the recovered variance (Appendix D.2), an order of magnitude below the estimation errors we report, so the reported estimates are $-1/\hat{\mu}$ uncorrected. What the decomposition does not transfer is the law of the trained parameters. Targets redrawn at each step vary θ_∞, and hence $\mathbf{\Delta}$, beyond the initialization randomness that (58) describes, and we do not model that contribution. The controlled experiments of Appendix D.4 train against the analytical score, where it is absent, and the cluster is present there, so it adds to the phenomenon rather than causing it.

C.2 Jacobian as Gaussian Process

From Lee et al. (2019) (eq. (59)), the trained network output decomposes into a deterministic part and a zero-mean GP residual:

$$\mathbf{f}(\mathbf{x}; \theta_\infty) = \underbrace{\mathbf{\Theta}^*(\mathbf{x}, \mathbf{X})\mathbf{\Theta}^*(\mathbf{X}, \mathbf{X})^{-1}\mathbf{f}_0(\mathbf{X})}_{\boldsymbol{\mu}(\mathbf{x})\text{ (deterministic)}} + \underbrace{\mathbf{f}(\mathbf{x}; \theta_0) - \mathbf{\Theta}^*(\mathbf{x}, \mathbf{X})\mathbf{\Theta}^*(\mathbf{X}, \mathbf{X})^{-1}\mathbf{f}_{\mathrm{init}}}_{\mathbf{g}(\mathbf{x})\text{ (zero-mean GP)}}, \tag{61}$$

where $\mathbf{f}_0(\mathbf{X}) = \mathbf{Y}$ are the training targets (i.e., the true score values at the training points, not to be confused with the network output at initialization $\mathbf{f}(\mathbf{x}; \theta_0)$) and $\mathbf{f}_{\mathrm{init}} = [\mathbf{f}(\mathbf{x}_1; \theta_0), \ldots, \mathbf{f}(\mathbf{x}_N; \theta_0)]$ are the initial outputs. Defining $\mathbf{a}(\mathbf{x})^\top = \mathbf{\Theta}^*(\mathbf{x}, \mathbf{X})\mathbf{\Theta}^*(\mathbf{X}, \mathbf{X})^{-1}$, the GP residual $\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}; \theta_0) - \mathbf{a}(\mathbf{x})^\top \mathbf{f}_{\mathrm{init}}$ has covariance $\tilde{K}$ given by (59). Taking the input gradient of (61) and writing $\boldsymbol{\mu} = \mathbf{f}_0 + \mathbf{r}$ where $\mathbf{r}$ is the interpolation residual, we obtain for each output dimension $m = 1, \ldots, D$:

$$\left.\frac{\partial f_m}{\partial x_k}\right|_{\mathbf{x}} = \underbrace{\frac{\partial [f_0]_m}{\partial x_k}}_{\text{true Jacobian}} + \underbrace{\frac{\partial r_m}{\partial x_k}}_{\text{deterministic bias}} + \underbrace{\frac{\partial g_m}{\partial x_k}}_{\text{GP noise}}. \tag{62}$$

Since $\mathbf{g}$ is a zero-mean GP (from the NNGP theory; $\mathbb{E}[\mathbf{g}(\mathbf{x})] = \mathbf{0}$ because $\mathbb{E}[\mathbf{f}(\mathbf{x}; \theta_0)] = \mathbf{0}$ and $\mathbf{a}(\mathbf{x})^\top \mathbb{E}[\mathbf{f}_{\mathrm{init}}] = \mathbf{0}$), its derivative $\partial g_m / \partial x_k$ is also a GP provided $\tilde{K}$ is *mean-square differentiable* (Williams & Rasmussen, 2006). This condition is satisfied when the activation function ϕ is smooth (e.g., erf, tanh, softplus, SiLU), making $\tilde{K}$ be C^∞. The derivative GP has zero mean, since $\mathbb{E}[\partial g_m / \partial x_k] = \partial \mathbb{E}[g_m] / \partial x_k = \partial 0 / \partial x_k = 0$. Its covariance is (Williams & Rasmussen, 2006):

$$\mathrm{Cov}\left[\frac{\partial g_m(\mathbf{x})}{\partial x_k}, \frac{\partial g_{m'}(\mathbf{x}')}{\partial x'_l}\right] = \frac{\partial^2 \tilde{K}_{mm'}(\mathbf{x}, \mathbf{x}')}{\partial x_k \partial x'_l}, \tag{63}$$

where $\tilde{K}_{mm'}(\mathbf{x}, \mathbf{x}') \coloneqq \mathrm{Cov}[g_m(\mathbf{x}), g_{m'}(\mathbf{x}')]$ is the (m, m')-entry of the matrix-valued residual kernel. For architectures with independent readout weights per output dimension (e.g., fully-connected networks), $\tilde{K}_{mm'} = \delta_{mm'} \tilde{K}$ and (63) simplifies accordingly (Neal, 2012; Lee et al., 2018). For architectures with weight sharing across output positions (e.g., convolutional networks, U-Nets), $\tilde{K}_{mm'} \neq 0$ for $m \neq m'$ in general.

Contracting this covariance against a unit vector gives the variance of the scalar perturbation that Section 4.2 reads off each eigenvalue; that contraction is written out in (64) below. Every architectural detail reaches the spectrum through this covariance and nothing else survives to first order.

C.3 General Eigenvalue Perturbation

The cross-output derivative covariance $\mathbf{\Sigma}_\partial^{(mm')} \in \mathbb{R}^{D\times D}$, one block per pair of output dimensions (m, m'), is defined in (63). This matrix captures the covariance between the m-th row and the m'-th row of the Jacobian perturbation: $[\mathbf{\Sigma}_\partial^{(mm')}]_{kl} = \mathrm{Cov}[\Delta_{mk}, \Delta_{m'l}]$, where $\Delta_{mk} = \partial g_m/\partial x_k|_{\mathbf{x}_j}$ with row index m (output dimension) and column index k (input coordinate). The collection $\{\mathbf{\Sigma}_\partial^{(mm')}\}_{m,m'=1}^{D}$ forms the $D \times D$ block structure of the full $D^2 \times D^2$ covariance of $\mathrm{vec}(\mathbf{\Delta})$.

Under the NTK regime with smooth activation, the Jacobian of the trained network at a training point $\mathbf{x}_j$ is the three-term decomposition (11) of Section 4.1, where $\nabla\mathbf{r}$ is the derivative of the deterministic 0-th order approximation error and $\mathbf{\Delta} \in \mathbb{R}^{D\times D}$ is a zero-mean Gaussian random matrix whose D^2 entries are jointly Gaussian with the block covariance above. In any practical setting $\nabla\mathbf{r}$ does not vanish, and it contributes to the perturbation of the Jacobian. The argument below does not rest on its size. Because $\mathbf{\Delta}$ originates in the randomness of the initialization rather than in any approximation error, it is present even in the idealized case $\nabla\mathbf{r} = \mathbf{0}$: a perturbation of the eigenvalues survives an exactly interpolating network, and nothing that follows requires knowing which of the two terms dominates. Let λ_i^0 and $\mathbf{u}_i \in \mathbb{R}^D$ ($\|\mathbf{u}_i\| = 1$) denote the i-th eigenvalue and eigenvector of the true score Jacobian $\nabla\mathbf{f}_0(\mathbf{x}_j)$, respectively. First-order perturbation theory gives, with $r_i \coloneqq \mathbf{u}_i^\top \nabla\mathbf{r}\, \mathbf{u}_i$,

$$\begin{aligned} \hat{\lambda}_i &= \lambda_i^0 + r_i + \mathbf{u}_i^\top \mathbf{\Delta} \mathbf{u}_i + \mathcal{O}(\|\mathbf{\Delta}\|^2), \\ \mathbf{u}_i^\top \mathbf{\Delta} \mathbf{u}_i &\sim \mathcal{N}\left(0,\ \sum_{m,m'=1}^{D} [u_i]_m\, [u_i]_{m'} \cdot \mathbf{u}_i^\top \mathbf{\Sigma}_\partial^{(mm')} \mathbf{u}_i\right). \end{aligned} \tag{64}$$

The perturbation $\mathbf{u}_i^\top \mathbf{\Delta} \mathbf{u}_i = \sum_{m,k} [u_i]_m [u_i]_k \Delta_{mk}$ linearly combines jointly Gaussian entries, so its Gaussian law is exact for a fixed $\mathbf{u}_i$; the first-order truncation is the only approximation, and the deterministic r_i shifts the centre without affecting that law. The **zero-mean Gaussian** property of the eigenvalue perturbation holds for any architecture satisfying the NTK conditions, only with differences in the perturbation variance caused by $\mathbf{\Sigma}_\partial^{(mm')}$. It is therefore possible to extract the true score Jacobian eigenvalues from the observed eigenspectrum using similar methods regardless of the network architecture.

With $\mathbf{u}_i$ the eigenvectors of $\mathbf{H}_\mathbf{y}$, the quantities $\lambda_0 + b_i + r_i + \mathbf{u}_i^\top \mathbf{\Delta} \mathbf{u}_i$, where $b_i \in [0, \eta]$ is the curvature offset of the i-th true noise eigenvalue above λ_0 supplied by Proposition 1, are the diagonal entries of the learned Jacobian in that basis rather than its exact eigenvalues, which is why Remark 3 is offered as a heuristic and Proposition 2(i) is proved on the block itself rather than through this basis.

C.4 Explicit NNGP Residual Covariance

With $\mathbf{\Theta}^{*-1} \coloneqq \mathbf{\Theta}^*(\mathbf{X}, \mathbf{X})^{-1}$, the residual kernel referenced in Appendix C.1 is

$$\begin{aligned} \tilde{K}(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}') &- K(\mathbf{x}, \mathbf{X}) \mathbf{\Theta}^{*-1} \mathbf{\Theta}^*(\mathbf{X}, \mathbf{x}') \\ &- \mathbf{\Theta}^*(\mathbf{x}, \mathbf{X}) \mathbf{\Theta}^{*-1} K(\mathbf{X}, \mathbf{x}') \\ &+ \mathbf{\Theta}^*(\mathbf{x}, \mathbf{X}) \mathbf{\Theta}^{*-1} K(\mathbf{X}, \mathbf{X})\, \mathbf{\Theta}^{*-1} \mathbf{\Theta}^*(\mathbf{X}, \mathbf{x}'). \end{aligned} \tag{65}$$

It mixes the prior kernel K with the training-dynamics kernel $\mathbf{\Theta}^*$, and vanishes at training points in the exact-interpolation limit. Setting $\mathbf{x}' = \mathbf{x}$ recovers the covariance in (59). That vanishing is a statement about values, not derivatives: $\tilde{K}(\mathbf{x}_j, \mathbf{x}_j) = \mathbf{0}$ while the mixed second derivative $\partial^2 \tilde{K}/\partial\mathbf{x}\partial\mathbf{x}'$ of (63), evaluated at $\mathbf{x} = \mathbf{x}' = \mathbf{x}_j$, does not vanish there, so $\mathbf{\Delta}$ is nonzero at the training points even under exact interpolation. Interpolating a function does not interpolate its derivative, which is why the noise cluster retains a width that better fitting does not remove.

C.5 Proof of Proposition 2

In the special case of Section 4.2 the blocks pooled over $\mathcal{S}$ share the centre λ_0 and differ only in width, the scale of $\mathbf{\Delta}$ varying from observation to observation. Pooling in that case reduces both properties the estimator uses to a statement about a single block: a non-negative mixture $\sum_i w_i p_i$ of

densities that are symmetric about λ_0 and decreasing in $|\lambda - \lambda_0|$ has both properties pointwise, so its only local maximum is at λ_0. Away from that case Proposition 1 places the per-observation centres anywhere in $[\lambda_0, \lambda_0 + \eta]$, so the centres no longer coincide and the reduction does not apply.

We turn to Proposition 2 itself.

Proof. (i) Let $\mathbf{V}_N \in \mathbb{R}^{D\times q}$, with $q = D - d$, be an isometry whose columns form an orthonormal basis of the noise eigenspace, so that $\mathbf{P} = \mathbf{V}_N\mathbf{V}_N^\top$, and write $\mathbf{M} = \mathbf{V}_N^\top\mathbf{\Delta}\mathbf{V}_N$ for the compression of the perturbation to the noise eigenspace, so that by (9) the noise block of the learned Jacobian is $\lambda_0\mathbf{I}_q + \mathbf{M}$, a $q \times q$ random matrix representing $\lambda_0\mathbf{I} + \mathbf{P}\mathbf{\Delta}\mathbf{P}$ restricted to that eigenspace. We first fix notation for the objects involved.

The argument is short and the formal apparatus below only makes it precise, so we give it first. $\mathbf{\Delta}$ is centred Gaussian, so its density is unchanged by $\mathbf{\Delta} \mapsto -\mathbf{\Delta}$: whatever draw occurs, the negated draw is equally likely. Negating the matrix negates its eigenvalues. So every spectrum the block can produce and its mirror image about λ_0 occur with the same probability, and averaging over the randomness leaves a distribution symmetric about λ_0. Only one property of $\mathbf{\Delta}$ is used, that its law is invariant under negation, and nothing about the *value* of $\mathbf{\Sigma}_\partial$. Note also what the argument does *not* say: a single realization is not symmetric — its mirror image is merely as likely — so the symmetry is a property of the ensemble, and this is why the estimator reads a density pooled over $\mathcal{S}$ rather than one observation's spectrum.

Let $\mu_1(\mathbf{A}), \dots, \mu_q(\mathbf{A}) \in \mathbb{C}$ denote the eigenvalues of a $q \times q$ matrix $\mathbf{A}$, listed with *algebraic* multiplicity: the characteristic polynomial $\det(\mathbf{A} - z\mathbf{I})$ factors as $(-1)^q \prod_i (z - \mu_i(\mathbf{A}))$, and each eigenvalue is repeated as often as its factor occurs. Counting this way, rather than by the dimension of the eigenspace, guarantees exactly q entries in the list even when $\mathbf{A}$ is not diagonalisable — which matters here because $\mathbf{\Delta}$ is not symmetric. For a random $\mathbf{A}$, its *empirical spectral distribution* is the random probability measure on $\mathbb{C}$

$$\mathrm{ESD}_{\mathbf{A}} \;\coloneqq\; \frac{1}{q}\sum_{i=1}^{q} \delta_{\mu_i(\mathbf{A})}, \qquad \text{so} \qquad \mathrm{ESD}_{\mathbf{A}}(B) \;=\; \frac{1}{q}\,\#\{\, i : \mu_i(\mathbf{A}) \in B \,\}. \tag{66}$$

Here δ_x is the point mass at x, the measure assigning 1 to every set containing x and 0 to every set not containing it, and $\#S$ is the number of elements of S; B ranges over the Borel subsets of $\mathbb{C}$, that is over sets obtained from open sets by countable unions, intersections and complements, which is the family on which a measure is defined. In words, (66) is the normalised histogram of the eigenvalues of one realization, and $\mathrm{ESD}_{\mathbf{A}}(B)$ is the *fraction of the q eigenvalues that lie in B*. Each $\mathrm{ESD}_{\mathbf{A}}(B)$ is a random variable with values in $[0, 1]$. Define the *mean spectral measure* ν of $\mathbf{M}$ by

$$\nu(B) \;\coloneqq\; \mathbb{E}\big[\mathrm{ESD}_{\mathbf{M}}(B)\big], \tag{67}$$

so that $\nu(B)$ is the *probability that an eigenvalue drawn at random from the block lies in B*, averaged over the randomness of $\mathbf{\Delta}$. That ν is again a probability measure is a routine check: $\nu(B) \geq 0$ because $\mathrm{ESD}_{\mathbf{M}}(B) \geq 0$; $\nu(\mathbb{C}) = 1$ because every realization has all q of its eigenvalues somewhere in $\mathbb{C}$, so $\mathrm{ESD}_{\mathbf{M}}(\mathbb{C}) = 1$ identically; and for disjoint $B_1, B_2, \dots$ each realization satisfies $\mathrm{ESD}_{\mathbf{M}}(\bigcup_k B_k) = \sum_k \mathrm{ESD}_{\mathbf{M}}(B_k)$, whose partial sums increase to the total and are bounded by 1, so monotone convergence permits taking the expectation term by term and $\nu(\bigcup_k B_k) = \sum_k \nu(B_k)$. Equivalently, if an index ι is drawn uniformly from $\{1, \dots, q\}$ independently of $\mathbf{M}$, then ν is the law of $\mu_\iota(\mathbf{M})$: that is, ν is the distribution of *one eigenvalue picked at random from the block*. It is ν, and not the spectrum of any single realization, that the pooled density of Section 5 estimates. Three operations are applied to eigenvalues in what follows — negating the matrix negates them, passing from $\mathbf{M}$ to the block $\lambda_0\mathbf{I} + \mathbf{M}$ shifts them, and the estimator reads their real parts — and we need notation for the effect of each on a measure. For a map ϕ and a measure ρ, the *pushforward* $\phi_\#\rho$ is the measure defined by

$$(\phi_\#\rho)(B) \;\coloneqq\; \rho\big(\phi^{-1}(B)\big), \qquad \phi^{-1}(B) = \{\, z : \phi(z) \in B \,\}, \tag{68}$$

that is, the measure obtained by moving each point z to $\phi(z)$ and carrying its mass along. Note that ϕ itself acts on points, whereas $\phi_\#$ acts on measures, and that on a point mass $\phi_\#\delta_x = \delta_{\phi(x)}$. We use one such map, the reflection $R(z) = -z$. We also write $-B \coloneqq \{-z : z \in B\}$, and note that R is its own inverse, so $R^{-1}(B) = -B$.

Step 1: $\mathbf{M} \overset{d}{=} -\mathbf{M}$. By Section 4.1 the entries of $\mathbf{\Delta}$ are jointly Gaussian with mean zero. The matrix $-\mathbf{\Delta}$ has the same mean, namely $\mathbf{0}$, and the same covariance, since $\mathrm{Cov}[-\Delta_{mk}, -\Delta_{m'l}] = \mathrm{Cov}[\Delta_{mk}, \Delta_{m'l}]$. A centred Gaussian law is determined by its covariance, so $\mathbf{\Delta} \overset{d}{=} -\mathbf{\Delta}$. The projector $\mathbf{P}$ is the noise eigenprojector of the *true* Jacobian $\mathbf{H}_{\mathbf{y}}$, so it is fixed by the observation and the data distribution and involves no network quantity, whereas the randomness this proposition averages over is that of the initialization. The two are nonetheless read at the same random observation and so are not independent unconditionally. We therefore condition on the training set and the observation: given those, $\mathbf{P}$ and $\mathbf{V}_N$ are fixed, $\mathbf{\Delta}$ is still centred Gaussian, and $\mathbf{A} \mapsto \mathbf{V}_N^\top \mathbf{A} \mathbf{V}_N$ is a fixed linear map, so $\mathbf{M} \overset{d}{=} -\mathbf{M}$ conditionally. The conclusion holds for every conditioning value and therefore also after averaging over them, which is the form used below. Three things are worth noting. The *value* of $\mathbf{\Sigma}_\partial$ is never used: negation preserves it, so the conclusion holds for every covariance, and in particular no isotropy or independence among the entries of $\mathbf{\Delta}$ is required. Gaussianity is sufficient but not necessary — any law invariant under $\mathbf{\Delta} \mapsto -\mathbf{\Delta}$ would serve. And nothing here requires the blocks read at distinct observations to be independent of one another: what Step 3 needs is that the *joint* law of the perturbations over the pooled observations be invariant under negating all of them at once, which holds because $\mathbf{g}$ is a centred Gaussian *process*, so negating a single draw of it negates its value at every input simultaneously. This matters because the deployed estimator pools blocks read from one trained network, which carry a single draw of $\mathbf{g}$ and are therefore dependent; the symmetry of the pooled law is untouched by that dependence.

Step 2: $\mathrm{ESD}_{-\mathbf{M}} = R_{\#}\mathrm{ESD}_{\mathbf{M}}$*, pathwise.* In words: negating a matrix mirrors its eigenvalue histogram about the origin. If $\mathbf{M}\mathbf{v} = \mu\mathbf{v}$ with $\mathbf{v} \neq \mathbf{0}$ then $(-\mathbf{M})\mathbf{v} = -\mu\mathbf{v}$, and conversely, so the eigenvalues of $-\mathbf{M}$ are exactly $-\mu_1(\mathbf{M}), \ldots, -\mu_q(\mathbf{M})$ with the same multiplicities. Counting how many of them fall in a given set B is therefore the same as counting how many eigenvalues of $\mathbf{M}$ fall in $-B$, since $-\mu_i \in B$ exactly when $\mu_i \in -B$. Hence for every Borel B,

$$\begin{aligned} \mathrm{ESD}_{-\mathbf{M}}(B) &= \frac{1}{q}\#\{i : -\mu_i(\mathbf{M}) \in B\} = \frac{1}{q}\#\{i : \mu_i(\mathbf{M}) \in -B\} \\ &= \mathrm{ESD}_{\mathbf{M}}(-B) = \big(R_{\#}\mathrm{ESD}_{\mathbf{M}}\big)(B). \end{aligned} \tag{69}$$

Equivalently, and more directly, pushforward moves point masses to their images, so

$$R_{\#}\mathrm{ESD}_{\mathbf{M}} \;=\; \frac{1}{q}\sum_{i=1}^{q} R_{\#}\delta_{\mu_i(\mathbf{M})} \;=\; \frac{1}{q}\sum_{i=1}^{q} \delta_{-\mu_i(\mathbf{M})} \;=\; \mathrm{ESD}_{-\mathbf{M}}, \tag{70}$$

which uses only that the eigenvalues of $-\mathbf{M}$ are the $-\mu_i(\mathbf{M})$. In words: the eigenvalue histogram of $-\mathbf{M}$ is the histogram of $\mathbf{M}$ mirrored about the origin. This step is exact for each realization and uses no probabilistic assumption; it requires neither that $\mathbf{M}$ be symmetric nor that its eigenvalues be real.

Step 3: ν *is invariant under reflection.* Fix a Borel set B and let $f(\mathbf{A}) \coloneqq \mathrm{ESD}_{\mathbf{A}}(B)$, a deterministic and measurable function of a matrix: it returns the fraction of the eigenvalues of $\mathbf{A}$ that lie in B. Then $\mathrm{ESD}_{\mathbf{M}}(B) = f(\mathbf{M})$ and $\mathrm{ESD}_{-\mathbf{M}}(B) = f(-\mathbf{M})$. If two random objects have the same distribution then so does any fixed function of them, and hence they have the same expectation; Step 1 gives $\mathbf{M} \overset{d}{=} -\mathbf{M}$, so $f(\mathbf{M}) \overset{d}{=} f(-\mathbf{M})$ and in particular $\mathbb{E}[f(\mathbf{M})] = \mathbb{E}[f(-\mathbf{M})]$. This is a statement about laws and not about realizations: for a single draw of $\mathbf{M}$ the two numbers $\mathrm{ESD}_{\mathbf{M}}(B)$ and $\mathrm{ESD}_{-\mathbf{M}}(B)$ generally differ, since they describe two different matrices. Combining with the pathwise identity (69),

$$\nu(B) \underset{(67)}{=} \mathbb{E}\big[\mathrm{ESD}_{\mathbf{M}}(B)\big] \underset{\text{Step 1}}{=} \mathbb{E}\big[\mathrm{ESD}_{-\mathbf{M}}(B)\big] \underset{\text{Step 2}}{=} \mathbb{E}\big[\mathrm{ESD}_{\mathbf{M}}(-B)\big] \underset{(67)}{=} \nu(-B). \tag{71}$$

As B was arbitrary, $R_{\#}\nu = \nu$: the mean spectral measure is symmetric about the origin. We emphasise that this is a statement about ν and not about any single realization; the spectrum of one draw of $\mathbf{M}$ is in general not symmetric, which is why the estimator reads a density pooled over $\mathcal{S}$, the empirical counterpart of ν.

Consequences for the block. The eigenvalues of $\lambda_0\mathbf{I} + \mathbf{M}$ are $\lambda_0 + \mu_i(\mathbf{M})$, so their distribution is that of ν shifted by λ_0, and symmetry about the origin becomes symmetry about λ_0. Since $\mathbf{M}$ need not be symmetric these eigenvalues may be complex, and the estimator reads their real parts; because $\mathrm{Re}(-z) = -\mathrm{Re}(z)$ and λ_0 is real, the real-part marginal inherits the symmetry, so the law ϱ of $\mathrm{Re}\,\lambda$ on $\mathbb{R}$ is symmetric about λ_0. Equivalently $\lambda - \lambda_0 \overset{d}{=} \lambda_0 - \lambda$ for $\lambda \sim \varrho$, which is (12).

Algorithm 1 JANE: Blind Score Jacobian Denoising

Require: Noisy observations $\{\mathbf{y}^{(i)}\}_{i=1}^{N}$, score network architecture
Ensure: Denoised estimates $\{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{N}$, noise estimate $\hat{\sigma}^2$

1: **Train score network:** Learn $\hat{\mathbf{s}}_\theta(\mathbf{y}) \approx \nabla_{\mathbf{y}} \log p(\mathbf{y})$ via denoising score matching
2: **Compute Jacobians:** For a subset $\mathcal{S} \subset \{1, \dots, N\}$, compute $\hat{\mathbf{H}}_{\mathbf{y}^{(i)}} = \nabla_{\mathbf{y}} \hat{\mathbf{s}}_\theta(\mathbf{y}^{(i)})$ for all $i \in \mathcal{S}$
3: **Eigendecomposition:** For each $i \in \mathcal{S}$, compute eigenvalues $\{\hat{\lambda}_1^{(i)}, \dots, \hat{\lambda}_D^{(i)}\}$ of $\hat{\mathbf{H}}_{\mathbf{y}^{(i)}}$
4: **Density estimation:** Pool all eigenvalues $\{\hat{\lambda}_k^{(i)}\}_{i\in\mathcal{S},\, k=1}^{D}$ and estimate the density $\hat{p}(\lambda)$ (Appendix D.2)
5: **Find leftmost critical point:** $\hat{\mu} \leftarrow \min\{\lambda : \hat{p}'(\lambda) = 0,\ \hat{p}''(\lambda) < 0\}$
6: **Estimate noise:** $\hat{\sigma}^2 \leftarrow -1/\hat{\mu}$
7: **Denoise:** For all $i = 1, \dots, N$: $\hat{\mathbf{x}}^{(i)} \leftarrow \mathbf{y}^{(i)} + \hat{\sigma}^2\, \hat{\mathbf{s}}_\theta(\mathbf{y}^{(i)})$

A law symmetric about λ_0 assigns equal mass to $(-\infty, \lambda_0)$ and (λ_0, ∞), so λ_0 is a median of ϱ. Its first moment is finite: every eigenvalue obeys $|\mu_i(\mathbf{M})| \le \|\mathbf{M}\|_{\mathrm{op}} \le \|\mathbf{\Delta}\|_{\mathrm{op}}$ because $\mathbf{V}_N$ is an isometry, and a centred Gaussian matrix has moments of every order, so $\int |x|\, \mathrm{d}\varrho < \infty$. Hence $\int (x - \lambda_0)\, \mathrm{d}\varrho = 0$ by the same symmetry, so its mean is λ_0 as well.

The extremum. The minimum over the pooled block is an extreme order statistic of ϱ, whose spread is set by the scale of $\mathbf{\Delta}$. It is therefore displaced below λ_0 by an amount that grows with that scale, so $-1/\lambda_{\min}$ under-estimates σ^2. The contrast with the previous paragraph is the point: λ_0 is pinned by the reflection symmetry (71) alone, which involves no scale, whereas the extremum is displaced by precisely that scale.

(ii) Forming the mean or the median of the noise block requires restricting an average or an order statistic to the block, which presupposes knowing d. The mode does not. Suppose the noise density p has a unique local maximum, at m. Symmetry gives $p(2\lambda_0 - \lambda) = p(\lambda)$, so $2\lambda_0 - m$ is also a local maximum; uniqueness forces $2\lambda_0 - m = m$, that is $m = \lambda_0$. By Proposition 1 the d signal eigenvalues of $\mathbf{H}_{\mathbf{y}}$ lie above λ_0, and the corresponding eigenvalues of the learned Jacobian are supported to the right of the block, so their contribution leaves the leftmost maximum where it is when they are separated from the block. □

D Experimental Details

This appendix collects the configuration details for the experiments of Section 6.

D.1 Baselines

Baselines. (1) *UNSURE* (Tachella et al., 2025): blind noise estimation via Stein's identity applied to the score norm, $\hat{\sigma}^2 = D/\mathbb{E}\|\hat{\mathbf{s}}_\theta(\mathbf{y})\|^2$ (their Proposition 1), evaluated with the same score network as all other methods so that only the noise estimate varies. (2) *Noise2Score* (Kim et al., 2022): the noise level obtained from finite-difference score differences, with $\mathbf{y}_2 = \mathbf{y}_1 + \epsilon\mathbf{u}$ for $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ giving $\hat{\sigma}^2$ from $-\epsilon\mathbf{u}/(\hat{\mathbf{s}}_\theta(\mathbf{y}_2) - \hat{\mathbf{s}}_\theta(\mathbf{y}_1))$ averaged over random directions and samples, followed by Tweedie denoising with the same score network. This is the estimator of Kim et al. (2022) rather than the total-variation search of the earlier Noise2Score, which that paper replaces. (3) *Genie MMSE*: oracle step-size search for the *learned* score network—finds the step α minimizing MSE against the clean signal using Tweedie's formula $\hat{\mathbf{x}} = \mathbf{y} + \alpha\, \hat{\mathbf{s}}_\theta(\mathbf{y})$, representing the best achievable MSE for a given trained network (not the true MMSE). All baselines use the same trained score network and apply Tweedie's formula with different noise estimates, isolating the effect of noise estimation quality. Classical image-domain estimators (e.g., wavelet-based MAD (Donoho & Johnstone, 1994)) address a different setting where the denoiser is decoupled from the noise estimator, and architecture-based blind denoisers (e.g., Noise2Void (Krull et al., 2019)) do not produce explicit noise estimates for comparison.

D.2 Training and Evaluation

We use a U-Net (Ronneberger et al., 2015) trained with denoising score matching (DSM) for 1000 epochs, except 300 for IN-64, with learning rate 5×10^{-4}, batch size 512 for CIFAR-10, IN-32 and Channel, 256 for CUB-64 and CUB-32C, and 1024 for IN-64, and early stopping. The network is an encoder–decoder U-Net with two 3×3 convolutions per scale, SiLU activations, batch normalisation and concatenated skip connections; the convolutions carry additive bias terms and the network is not residual. CIFAR-10, IN-32 and Channel use four scales with 64, 128, 256 and 512 channels; IN-64 and CUB-64 use two scales with 64 and 128; CUB-32C uses two scales with 256 and 512 and no batch normalisation. The DSM perturbation scale is $\sigma_{\mathrm{dsm}} = 0.01$ in normalized space. Since DSM smoothing makes the learned score that of $p_{\mathbf{y}} * \mathcal{N}(0, \sigma^2_{\mathrm{dsm}})$, it inflates the recovered variance by $\sigma^2_{\mathrm{dsm}}/\sigma^2_{\mathrm{norm}}$, where both are expressed in the network's normalized input space, in which the data are divided by their standard deviation: on CUB-32C that standard deviation is 0.233, so a pixel-scale $\sigma = 12$ gives $\sigma^2_{\mathrm{norm}} = (12/255)^2/0.233^2 = 0.0407$, not $(12/255)^2$. This ratio is largest at the lowest noise level and equals 0.25% there ($10^{-4}/0.0407$), an order of magnitude below the estimation errors we report, and it decreases as σ grows. During eigenspectrum analysis, we compute the full $D \times D$ Jacobian for a random subset $\mathcal{S}$ of samples via automatic differentiation. We use $|\mathcal{S}| = 5000$ for CIFAR-10, IN-32 and Channel, 4000 for IN-64 and CUB-64, and 1000 for CUB-32C.

ImageNet configurations. IN-32 and IN-64 draw on 13 ImageNet (Deng et al., 2009) classes selected for high edge and texture content: `n02791270` (honeycomb), `n03000134` (chain-link fence), `n03598930` (jigsaw puzzle), `n03085013` (computer keyboard), `n02870880` (bookshop), `n03857828` (printed circuit board), `n01773157` (spider web), `n02391049` (zebra), `n01917289` (brain coral), `n04254680` (soccer ball), `n03482405` (hamper), `n02123045` (tabby cat) and `n02128757` (snow leopard). Images are converted to grayscale and cropped rather than resized: four 32×32 patches, respectively 64×64 for IN-64, are taken from a 2×2 grid at the centre of each image, or a single central patch when the image is smaller than twice the patch size.

Locating the peak. The mode is found in two stages. A 500-bin histogram of the pooled eigenvalues is smoothed with a Gaussian filter, and the leftmost bin-peak whose height reaches 5% of the tallest peak's height is taken as a coarse location; the threshold discards small ripples in the left tail so that the leftmost peak is not a sampling fluctuation. A kernel density estimate is then run on the eigenvalues within a fixed window of that coarse location, with the rule-of-thumb bandwidth $0.9\,\hat{s}\,N_{\mathrm{loc}}^{-1/5}$, and $\hat{\mu}$ is its argmax.

Limitations. JANE's smoothing bandwidth and its 5% peak-height threshold are fixed across all configurations. Each evaluated sample costs D backward passes and one $D \times D$ eigendecomposition; matrix-free spectral density estimation (Ubaru et al., 2017) is the natural route past the $\mathcal{O}(D^3)$ term, which we have not pursued. The geometric analysis of Proposition 1 is a conceptual account at image resolutions, where other effects also act on the spectrum. Classical patch-based noise estimators form a separate family we do not evaluate, and the 10 training seeds reported throughout vary network initialisation and batch order at one noise realisation per setting. Without symmetry constraint, in practice the learned Jacobian is generally asymmetric, and eigenvalues are approximately estimated via the real part results of `numpy.linalg.eigvals`. Measured asymmetry $\|\hat{\mathbf{J}} - \hat{\mathbf{J}}^\top\|_F / \|\hat{\mathbf{J}}\|_F$ on the controlled manifold is 0.28% under exact-score training and 3.15% under DSM; across a DSM noise sweep the asymmetry rises from 1.8% at $\sigma = 0.02$ to 5.5% at $\sigma = 0.2$, which is why retaining real parts is adequate in practice. The true score Jacobian is symmetric by construction, so any asymmetry is model error. We retain the real parts rather than symmetrizing so that this error remains visible in the spectrum rather than being averaged away. For color image datasets ($D = 3072$), we train an ensemble of 32 score networks and average their scores and Jacobians to improve estimation quality. Denoising quality is measured by PSNR (dB) and noise estimation quality by the ratio $\hat{\sigma}^2/\sigma^2$. PSNR clips the estimate to $[0, 1]$ before measuring the error, as is standard, and is averaged per image; the Noisy row therefore lies above $-20 \log_{10}(\sigma/255)$, increasingly so at large σ. The observations supplied to the score network are *not* clipped, so the model of (1) and the reference σ^2 used for $\hat{\sigma}^2/\sigma^2$ are unaffected.

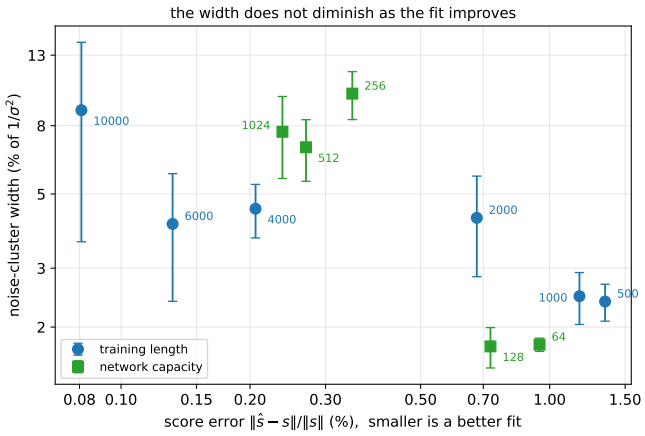

Figure 4: The spread is not a training deficiency. Same 2-torus setting as Figure 1. Noise-cluster width, as a percentage of $1/\sigma^2$, against score error as training length and network capacity vary, with error bars giving ± 1 s.d. over the 10 seeds of each point's three-network mean, and no fitted trend. The width does not fall as the score error improves; undertraining would require it to vanish with the score error.

D.3 Channel Dataset Details

Regarding the channel phase dataset, we generate synthetic wireless channel images using 3GPP CDL (Clustered Delay Line) channel models (3GPP, 2020). Each image represents the phase of the channel frequency response $\angle H(f, t)$ over a 32-subband $\times$ 32-time-slot grid. The simulation uses CDL-A, CDL-B, and CDL-C delay profiles (selected randomly per image), each with 20 rays per cluster. Time-domain fading uses a sum-of-sinusoids model with 16 sinusoidal components per ray, UE velocity of 33.33 m/s (120 km/h), carrier frequency $f_c = 3.5$ GHz, and 20 ms sampling intervals (0.64 s total observation). Frequency-domain channels are obtained via 1024-point FFT with subband downsampling and normalization. The phase is wrapped to $[-\pi, \pi]$ and rescaled to $[0, 1]$ to form the clean image, and the observation noise is added directly to that image as i.i.d. $\mathcal{N}(0, \sigma^2)$ in the same pixel scale as the image datasets, so this configuration satisfies the same additive isotropic Gaussian model.

D.4 Toy-Example Training Sweeps

The two sweeps behind Figure 4 are run on the closed-form manifold of Section 3 against the analytical score. Each holds the other factor fixed and reuses the same 200,000 training samples: epochs 500 to 10,000 at width 512, and width 64 to 1024 at 3000 epochs. Every point averages three independently seeded networks, and each network's spectrum is measured at all 200,000 training observations, so the measurement uses no data the network was not trained on. For each network we measure the width of its own noise cluster, the distance from the cluster peak down to its most negative eigenvalue relative to $1/\sigma^2$, and average that width over the three networks.

D.5 Tabulated Toy-Example Results

The numbers plotted in Figure 1(b), tabulated for reference. The learned columns are relative to $-1/\lambda_{\text{ref}}$, the value a perfect network would return on this manifold, rather than to σ^2; the labels *under* and *over* refer to that reference.

Table 3: Relative error $|\hat{\sigma}^2/\sigma^2 - 1|$ on the 2-torus. The analytical rows use the exact spectrum (10) and involve no network; the learned rows are read off networks trained by denoising score matching. λ_{med} is the median of the closed-form noise eigenvalues, not their minimum; both analytical rows over-estimate σ^2.

		noise level σ					
Jacobian	statistic	0.020	0.035	0.050	0.080	0.120	0.200
analytical	$-1/\lambda_{\mathrm{med}}$	0.02%	0.06%	0.12%	0.32%	0.72%	2.00%
	$\frac{1}{2}(\sigma/\rho)^2$	0.02%	0.06%	0.13%	0.32%	0.72%	2.00%
learned	$-1/\lambda_{\mathrm{min}}$ (under)	8.66%	14.48%	35.79%	51.25%	56.24%	58.67%
	ours (under)	0.76%	1.04%	0.78%	0.66%	2.37%	9.28%

Table 4: Seed-to-seed standard deviation of every entry of Table 1, over the 10 training seeds: **PSNR s.d. in dB | s.d. of** $\hat{\sigma}^2/\sigma^2$. Each cell is the dispersion of the quantity whose mean Table 1 reports, so it measures run-to-run variability at fixed data and noise level, not estimator error. The *Noisy* row is identically zero because every seed shares one noise realisation: the seed varies network initialisation and batch order only, so all dispersions below are training variability alone. No entry is highlighted, since ranking dispersions is not the comparison Table 1 makes.

σ	Method	CIFAR-10	IN-32	IN-64	CUB-64	Channel	CUB-32C
12	Noisy	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –
	JANE	0.08 | 0.020	0.20 | 0.026	0.12 | 0.011	0.07 | 0.022	0.16 | 0.011	0.01 | 0.003
	Noise2Score	0.09 | 0.023	0.22 | 0.023	0.13 | 0.018	0.08 | 0.019	0.17 | 0.010	0.01 | 0.004
	UNSURE	0.21 | 0.059	0.21 | 0.044	0.20 | 0.046	0.18 | 0.046	0.20 | 0.028	0.05 | 0.012
	Genie	0.08 | 0.017	0.18 | 0.013	0.12 | 0.013	0.07 | 0.015	0.16 | 0.009	0.01 | 0.005
25	Noisy	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –
	JANE	0.19 | 0.028	0.29 | 0.019	0.18 | 0.012	0.11 | 0.015	0.38 | 0.011	0.04 | 0.007
	Noise2Score	0.20 | 0.028	0.29 | 0.015	0.19 | 0.013	0.13 | 0.019	0.40 | 0.017	0.04 | 0.005
	UNSURE	0.34 | 0.063	0.29 | 0.034	0.25 | 0.027	0.22 | 0.037	0.49 | 0.043	0.06 | 0.012
	Genie	0.18 | 0.021	0.28 | 0.013	0.18 | 0.009	0.12 | 0.014	0.38 | 0.013	0.04 | 0.005
51	Noisy	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –
	JANE	0.34 | 0.033	0.38 | 0.027	0.39 | 0.016	0.20 | 0.020	0.56 | 0.016	0.04 | 0.010
	Noise2Score	0.35 | 0.025	0.36 | 0.016	0.40 | 0.016	0.21 | 0.019	0.57 | 0.013	0.04 | 0.010
	UNSURE	0.45 | 0.050	0.46 | 0.037	0.50 | 0.039	0.26 | 0.033	0.67 | 0.042	0.12 | 0.023
	Genie	0.34 | 0.019	0.35 | 0.012	0.39 | 0.014	0.19 | 0.013	0.56 | 0.015	0.04 | 0.009
77	Noisy	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –	0.00 | –
	JANE	0.46 | 0.037	0.55 | 0.028	0.44 | 0.014	0.18 | 0.025	0.45 | 0.016	0.06 | 0.022
	Noise2Score	0.45 | 0.027	0.52 | 0.013	0.46 | 0.012	0.18 | 0.025	0.48 | 0.018	0.06 | 0.024
	UNSURE	0.56 | 0.053	0.57 | 0.028	0.49 | 0.026	0.33 | 0.049	0.49 | 0.047	0.33 | 0.056
	Genie	0.44 | 0.022	0.52 | 0.010	0.45 | 0.011	0.18 | 0.021	0.44 | 0.023	0.06 | 0.023