Where the Score Lives: A Wavelet View of Diffusion

Emma Lucia Byrnes Finn
Harvard University
efinn@college.harvard.edu

Binxu Wang
Harvard University
binxu_wang@hms.harvard.edu

T. Anderson Keller
Harvard University
t.anderson.keller@gmail.com

Demba E. Ba
Harvard University
demba@seas.harvard.edu

Abstract

Diffusion models have had remarkable success over the last decade in generating a diverse set of visually plausible images. These models work by transforming the data to a centered Gaussian and then learning the reverse process by training a neural network to approximate the score of the underlying distribution. A variety of architectures from CNNs, to U-Nets, to transformers have been used as the score-approximation network in diffusion modeling. We propose an analytically solvable parameterization of the score function using an expansion in a wavelet basis. In particular, we derive interpretable optimal score functions in a 2D, orthogonal wavelet basis in terms of the moments of the data distribution. We use this parametrization to provide an an architecture-agnostic, moment-based analysis that reveals which attributes of the data distribution tend to matter most for denoising. Our score machine is flexible enough to partially mimic the relevant inductive biases of multiple architectures, including U-Nets, and CNNs, taking a step towards understanding why different score architectures can exhibit distinct generative behavior. Since our score is solvable in terms of the moments of the data, we can begin to understand how the data distribution interacts with the score network to produce the behavior we observe in diffusion models.

1 Introduction

Diffusion models have rapidly advanced image generation (and many other generative tasks) in recent years [5, 13]. However, it is still not clear what causes their remarkable ability to generalize beyond their training distribution and constructing visually coherent samples. This generalization property clearly depends both on the score network used and on the underlying property of the data distribution. It is clear that a diffusion model trained on one image would not generalize well.

Recent work has explored both architectural contributions to the kind of creativity displayed in diffusion [8] and data-determined contributions [14]. On the architectural side Kamb and Ganguli [8] demonstrate that a CNN encodes certain inductive biases into a diffusion network when used as a diffusion backbone, encouraging a certain 'patch mosaic' form of creativity; building on prior work [9] which demonstrates that under certain conditions, the ideal score simply memorizes the training data.

On the data distribution side, Niedoba et al. [11] demonstrate empirically that spatially local information is most relevant for denoising across most time scales. In (Wang and Vastola [14]) the authors demonstrate that the first order Gaussian approximation of the score function is sufficient to explain a great deal of the behavior of diffusion models. Meanwhile, Bonnaire et al. [1] have demonstrated that

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Structured Probabilistic Inference Generative Modeling: Probabilistic Inference in the Era of Large Foundation Models.

the capacity to generalize beyond the training set scales with the training set size. Kadkhodaie et al. [7] suggest that diffusion models learn a "shrinkage operation in an orthonormal basis consisting of harmonic functions that are adapted to the geometry of features in the underlying image." However, it still remains unclear exactly what information about the underlying data distribution is most important to learning the score function. This question is key to understanding and improving diffusion models.

In this work, we propose an interpretable parameterization of the score function using an expansion in a wavelet basis, which is remarkably flexible. We can optimize this score function over the standard L^2 loss and derive an ideal score function particular to our functional form and identify how different architectural changes effect the ideal score. In particular, we implement a wavelet based score machine and run extensive experiments across different families of denoiser to understand which components of the data distribution are most relevant in which settings.

1.1 Diffusion Background

We first provide a short overview of score-based generative modeling. Diffusion models operate by corrupting all of the input data over time according to a noising process, given by an Ornstein-Uhlenbeck stochastic differential equation.

$$d\mathbf{X}_t = \underbrace{f(\mathbf{X}_t, t)}_{\text{drift}}\, dt + \underbrace{g(t)}_{\text{diffusion}}\, dW \tag{1}$$

The reverse process is given by

$$d\mathbf{X}_t = \left[f(\mathbf{X}_t, t) - \frac{1}{2} g(t)^2 \underbrace{\nabla_{X_t} \log(p_t(X_t))}_{\text{score function}} \right] dt + g(t) d\bar{W} \tag{2}$$

We parametrize a neural network s_θ to approximate the score function, and in our setup, we use the score matching objective:

$$\mathcal{L}(\theta) = \int_0^T \lambda(t) \mathbb{E}_{p_t(\mathbf{X})}[||s(\mathbf{X}, t) - s_\theta(\mathbf{X}, t)||_2^2] \tag{3}$$

Where λ is a time dependent weight. For simplicity, in our case, we'll simply use $\lambda(t) = 1 \forall t$. To further simplify, we'll divide our time interval $[0, 1]$ into N discrete steps, and treat each step independently, so the loss decouples across time. Our loss at any particular timestep is

$$\mathcal{L}^{(t)}(\theta) = \mathbb{E}_{p_t(\mathbf{X})}[||s(\mathbf{X}, t) - s_\theta(\mathbf{X}, t)||_2^2] \tag{4}$$

and our overall loss is simply the sum of all such $\mathcal{L}^{(t)}$. In the subsequent work, we'll instead parametrize our score function in a wavelet basis and optimize our learnable parameters directly.

1.2 Wavelets and Denoising

Wavelets have long been used for image denoising and representation. Mallat introduced multiresolution analysis (MRA) and a fast algorithm for the discrete wavelet transform [10]. Classical wavelet denoising by soft thresholding was developed by Donoho and Johnstone [2]. More recently, Phung et al. [12] have attempted to use wavelets to speed up diffusion sampling and other score-based generative frameworks, as in (Guth et al. [4]). Like Fourier methods, wavelet decompositions represent square-integrable functions in a basis indexed by "frequencies," but unlike Fourier bases, wavelets are localized in both space and scale, providing joint spatial–frequency resolution.

We adopt compactly supported Daubechies wavelets. Let ϕ denote the scaling ("father") function and ψ the wavelet ("mother") function in one dimension. In two dimensions we form tensor products to obtain one scaling atom and three detail atoms at each location/scale. The translates and dilates of these functions form an orthogonal basis of the entirety of $L^2(\mathbb{R}^2)$.

For scale $j \in \mathbb{Z}$ and translation $k = (k_1, k_2) \in \mathbb{Z}^2$, define

$$\begin{aligned}
\phi_{j,k}(\mathbf{u}) &= 2^j\, \phi(2^j u_1 - k_1)\, \phi(2^j u_2 - k_2), \\
\psi^{(\ell)}_{j,k}(\mathbf{u}) &= 2^j\, g^{(\ell)}(2^j u_1 - k_1,\, 2^j u_2 - k_2), \qquad \ell \in \{0, 1, 2\},
\end{aligned}$$

where $g^{(0)}(a,b) = \psi(a)\phi(b)$, $g^{(1)}(a,b) = \phi(a)\psi(b)$, and $g^{(2)}(a,b) = \psi(a)\psi(b)$. While the functions ϕ and ψ are not available in closed form for the Daubechies wavelets, for integer $N \geq 1$, the Daubechies–N wavelet $\psi(x)$ and scaling function $\phi(x)$ are defined by the two-scale relations

$$\phi(x) = \sqrt{2} \sum_{n=0}^{2N-1} h_n \, \phi(2x-n), \qquad \psi(x) = \sqrt{2} \sum_{n=0}^{2N-1} g_n \, \phi(2x-n),$$

With periodic boundary handling on $[0,1]^2$ (our default), the set

$$\mathcal{B}_{J_0} \;=\; \Big\{\phi_{J_0,k}\Big\}_k \;\cup\; \Big\{\psi^{(\ell)}_{j,k}\Big\}_{j \geq J_0,\, k,\, \ell}$$

forms an orthonormal basis of $L^2([0,1]^2)$. (Here J_0 is the coarsest scale; only the scaling atoms at J_0 are kept, while all finer detail wavelets $\psi^{(\ell)}_{j,k}$ are included for $j \geq J_0$.) At each fixed (j,k), the triple $\big(\psi^{(1)}_{j,k}, \psi^{(2)}_{j,k}, \psi^{(3)}_{j,k}\big)$ are the three orientation components often called the "horizontal", "vertical", and "diagonal" details. We refer to this per-location orientation triplet as the detail band at (j,k). Empirically, many U-Nets implement wavelet-like multiresolution computations with Haar-like filters [3].

1.3 Our Contributions

We provide an interpretable, analytic parameterization of diffusion scores in an orthonormal wavelet basis. By exploiting orthogonality and a score–integration-by-parts (Stein) identity, each wavelet-coefficient score reduces to a closed-form least-squares system whose right-hand side is a vector of data moments. This yields a spectrum of models—from independent sub-bands to band-tied and local-coupled—that reveal which scales and orientations drive denoising across noise levels.

2 A Wavelet Expansion of the Score

2.1 Expanding the Score

Let $\langle f, g \rangle = \int_{[0,1]^2} f(\mathbf{u})\, g(\mathbf{u})\, d\mathbf{u}$ denote the L^2 inner product. For a grayscale image X_t at noise level t and score $s_{\text{true}}(\cdot, t)$, since wavelets form an orthonormal basis of L^2 the score at time t in a grayscale image can be expanded as

$$s^{(t)}(X_t) \;=\; \sum_{i \in \mathcal{I}} c_i(X_t)\, w_i, \qquad c_i(X_t) = \big\langle s_{\text{true}}(X_t, t),\, w_i \big\rangle, \tag{5}$$

where $\{w_i\}_{i \in \mathcal{I}} = \mathcal{B}_{J_0}$ indexes (J_0, k) for scaling atoms and (j,k,ℓ) for detail atoms. The way in which we choose to model $\langle s_{\text{true}}(X_t), w_i \rangle$ determines the properties of the score function estimator. We model each coefficient by features $\varphi_i(X_t) \in \mathbb{R}^{d_i}$ and parameters $\alpha_i^{(t)} \in \mathbb{R}^{d_i}$:

$$\widehat{c}_i(X_t) \;=\; \alpha_i^{(t)\top} \varphi_i(X_t), \qquad s_\theta^{(t)}(X_t) \;=\; \sum_{i \in \mathcal{I}} \widehat{c}_i(X_t)\, w_i. \tag{6}$$

Because $\{w_i\}$ is orthonormal, the population squared loss decouples across i and so taking the gradient w.r.t. $\alpha_i^{(t)}$ yields the simultaneous equations

$$\mathcal{L}^{(t)}(\theta) = \mathbb{E}_{X_t \sim p_t} \Big\| s(X_t) - \sum_i \alpha_i^{(t)\top} \varphi_i(X_t) w_i \Big\|_{L^2}^2 \tag{7}$$

$$\implies \frac{\partial \mathcal{L}^{(t)}}{\partial \alpha_i^{(t)}} = -2\mathbb{E}\Big[\varphi_i(X_t)\big(\langle s(X_t), w_i \rangle - \alpha_i^{(t)\top} \varphi_i(X_t)\big)\Big] = 0 \tag{8}$$

Solving for the optimal coefficients $\alpha_i^{(t)*}$, we find, if $\Sigma_i = \mathbb{E}[\varphi_i \varphi_i^\top]$ is invertible:

$$\underbrace{\mathbb{E}[\varphi_i \varphi_i^\top]}_{\Sigma_i}\, \alpha_i^{(t)} \;=\; \mathbb{E}\big[\varphi_i(X_t)\, \langle s(X_t), w_i \rangle\big] \implies \alpha_i^{(t)\star} \;=\; \Sigma_i^{-1}\, \mathbb{E}\big[\varphi_i(X_t)\, \langle s(X_t), w_i \rangle\big] \tag{9}$$

We can further simplify, by applying a more general form of Stein's Identity as in [6] to the expectation of the inner product. In particular, for some function f, Stein's Score Identity says

$$\mathbb{E}_{X\sim p_t}\big[\, s_t(X)\cdot f(X)\,\big] \;=\; -\,\mathbb{E}_{X\sim p_t}\big[\,\nabla\cdot f(X)\,\big] \tag{10}$$

for p_t smooth and under vanishing boundary flux (e.g., periodic boundaries). In our case, we write $\varphi_i(X_t) = [\varphi_{i,1}(X_t), \ldots, \varphi_{i,d_i}(X_t)]^\top$. For each component j, apply Stein's identity with the vector field

$$f_j(x) \;=\; \varphi_{i,j}(x)\, w_i, \qquad \|w_i\|_2 = 1.$$

Then

$$\mathbb{E}\big[\varphi_{i,j}(X_t)\,\langle s_t(X_t), w_i\rangle\big] \;=\; -\,\mathbb{E}\big[\nabla\cdot\big(\varphi_{i,j}(X_t)\, w_i\big)\big] \;=\; -\,\mathbb{E}\big[\langle\nabla\varphi_{i,j}(X_t),\, w_i\rangle\big].$$

Stacking over $j = 1, \ldots, d_i$ gives the vector form

$$\mathbb{E}\big[\varphi_i(X_t)\,\langle s_t(X_t), w_i\rangle\big] \;=\; -\,\mathbb{E}\big[(\nabla\varphi_i(X_t))^\top w_i\big] \tag{11}$$

Thus, we find that our solution is

$$\alpha_i^{(t)\star} \;=\; -\mathbb{E}[\varphi_i\varphi_i^\top]^{-1}\;\mathbb{E}\big[(\nabla\varphi_i(X_t))^\top w_i\big] \tag{12}$$

In practice, expectations are replaced by sample averages over n training images at time t, $\mathbb{E}[f(X_t)] \approx \frac{1}{n}\sum_{r=1}^{n} f\big(X_t^{(r)}\big)$. We also add a ridge regularization term, because $\hat{\Sigma}_i$ can be singular or ill-conditioned with high-degree features or correlated wavelet coefficients, we stabilize and control variance by adding a ridge regularization, as follows

$$\hat{\alpha}_i^{(t)} \;=\; \big(\hat{\Sigma}_i + \gamma I\big)^{-1}\hat{b}_i, \quad \gamma > 0,$$

which guarantees an invertible system and mitigates overfitting.

Thus, for samples $\{X_t^{(n)}\}_{n=1}^N$ and ridge $\gamma \geq 0$,

$$\hat{\alpha}_i^{(t)}(\gamma) = \left(\frac{1}{N}\sum_{n=1}^{N}\varphi_i^{(n)}\varphi_i^{(n)\top} + \gamma I\right)^{-1}\left(-\,\frac{1}{N}\sum_{n=1}^{N}(\nabla\varphi_i(X_t^{(n)}))^\top w_i\right). \tag{13}$$

We estimate $\nabla\varphi_i(X_t^{(n)})$ analytically from the chosen features using the method of moments.

Diagonalizing $\Sigma_i = U\Lambda U^\top$ shows that $(\Sigma_i + \gamma I)^{-1}$ weights eigen-directions by $1/(\lambda+\gamma)$. Thus, given that the features are well suited to the data such that Σ_i in (9) is well conditioned and its columns are not highly correlated, the estimator weights eigen directions by $\frac{1}{\lambda+\gamma}$. Small-λ directions receive higher coefficients but are more noise sensitive, which the ridge regression helps to limit. Intuitively, what this says is that this score approximator emphasizes lower-variance feature directions (small λ) and down weights higher-variance directions (large λ). This aligns with the observation that natural images exhibit approximately power-law spectral decay and sparse wavelet coefficients, whereas white noise spreads energy more uniformly; consequently, informative structure often concentrates in a subset of scales and orientations. The model learns to correct more strongly along low-variance modes of the data distribution and ignore the high-variance ones.

2.2 Correlation Structures in the Data

Natural images exhibit structured dependencies in the wavelet domain: heavy–tailed marginals per coefficient, co-activation across orientations at a fixed location, and spatial persistence along edges and textures. As the noise level decreases, these dependencies become more pronounced. We therefore study three families that isolate, then incrementally reintroduce, these correlations.

(i) *Independent (diagonal).* We model each coefficient $y_t^{(i)} = \langle X_t, w_i\rangle$ in isolation using degree–D monomials (or probabilists' Hermite polynomials for numerical stability). This "mean-field" score approximator is the right baseline when p_t is close to a product distribution (early time steps, near-Gaussian), and it makes failures highly interpretable: any gap to stronger models directly measures the predictive value of cross-coefficient structure that a diagonal model cannot use.

(ii) *Band-tied.* At a fixed scale and location (j,k), we couple the three detail orientations $\ell \in \{1,2,3\}$ with degree–D interactions (each monomial includes the target coordinate). This targets cross-orientation co-activation caused by edges and corners and mirrors the channel-mixing inductive bias of CNN/U-Net score networks at a single spatial site. Improvements here isolate the contribution of within-pixel, within-scale structure—testing whether local orientation interactions alone explain denoising gains as t decreases.

(iii) *Local-coupled.* At fixed scale and orientation, we allow the coefficient at location k to depend on neighbors $k+\delta$ within a Chebyshev ball $\|\delta\|_\infty \le r$. This realizes a small neighborhood in wavelet space and emulates the increasing receptive field of convolutional (or locally attentive) score networks. Gains that grow then saturate with r quantify how much spatial context the score actually needs for accurate denoising and globally consistent structure as noise recedes.

We use (11) to build families of models that probe how scale/orientation structure in the data influences the score. We propose three families of models which can provide insight into what properties of the data distribution are relevant for score-based diffusion. In the Daubechies wavelet setting, we have

$$s_\theta(X_t,t)^{(t)}(\mathbf{u}) = \sum_{k=(0,0)}^{(2^{J_0}-1,2^{J_0}-1)} \theta_k(X_t,t)\,\phi_{J_0,k}(\mathbf{u}) + \sum_{j=J_0}^{J_{\max}} \sum_{k=(0,0)}^{(2^j-1,2^j-1)} \sum_{l=1}^{3} \zeta_{j,k,l}(X_t,t)\,\psi_{j,k}^{(l)}(\mathbf{u}) \tag{14}$$

where J_0 is the coarsest scale of the approximation, analogous to the spatial resolution at the network's bottleneck and for an image of dimension $H \times W$ $J_{\max} = \lfloor \log_2(\min(H,W)) \rfloor$, which corresponds to the highest level of wavelet detail encoded.

2.2.1 Independent Baseline

Assume (unrealistically) that coefficients decouple across scale j, location k, and detail band l. Let $y_t^{(i)} = \langle X_t, w_i \rangle$ with moments $\mu_r^{(i)}(t) = \mathbb{E}\big[(y_t^{(i)})^r\big]$ and model our coefficients as

$$\theta_k(X_t,t) = \sum_{m=0}^{D} b_m^{(k)}(t)\,\langle X_t, \phi_{J_0,k}\rangle^m, \qquad \zeta_{j,k,\ell}(X_t,t) = \sum_{m=0}^{D} d_m^{(j,k,\ell)}(t)\,\langle X_t, \psi_{j,k}^{(\ell)}\rangle^m.$$

Using (11) with monomial features gives the normal equations analogous to (7)for each index i and $r = 0,\dots,D$,

$$\sum_{m=0}^{D} a_m^{(i)}(t)\,\mu_{m+r}^{(i)}(t) = -\,r\,\mu_{r-1}^{(i)}(t), \qquad a_\bullet^{(i)} \in \{b_\bullet^{(k)},\, d_\bullet^{(j,k,\ell)}\}, \tag{15}$$

This is a Hankel system $H^{(i)}(t)\,a^{(i)}(t) = -\,h^{(i)}(t)$, where

$$H_{r,m}^{(i)}(t) = \mu_{r+m}^{(i)}(t), \qquad h^{(i)}(t) = \big[\,0,\ \mu_0^{(i)}(t),\ 2\mu_1^{(i)}(t),\dots,D\mu_{D-1}^{(i)}(t)\,\big]^\top.$$

Observe that all entries of $H^{(j,k,l)}(t)$ and $h^{(j,k,l)}(t)$ are computable from clean-data raw moments of $Y_0 = \langle X_0, \psi_{j,k}^{(l)}\rangle$.

$$\mu_r^{(i)}(t) = \sum_{m=0}^{r} \binom{r}{m} \bar\alpha_t^{m/2}\,(1-\bar\alpha_t)^{(r-m)/2}\,\mathbb{E}\big[\langle X_0, w_i\rangle^m\big]\,\mathbb{E}[Z^{r-m}]. \tag{16}$$

For small D, we can easily compute $a^{(i)}(t) = -(H^{(i)}(t))^{-1}h^{(i)}(t)$ (or $(H^{(i)}+\gamma I)^{-1}$ with ridge $\gamma \ge 0$). The co-factor formula for the coefficients reads

$$\hat\alpha_i(t) = -\,\frac{1}{\det H^{(i)}(t)} \sum_{r=1}^{D} r\,\mu_{r-1}^{(i)}(t)\,C_{r,m}^{(i)}(t) \tag{17}$$

with $C_{r,m}^{(j,k,\ell)}(t)$ the (r,m)-cofactor of $H^{(i)}(t)$. This closed form can be used to investigate how higher order moments of the data distribution impact the score. Though clearly independence is an unrealistic assumption, this model is fast and interpretable, and serves as (i) an initial baseline, and (ii) a diagnostic lower bound. We can investigate the value of different kinds of correlation by measuring

the difference in performance between models with different kinds of limited dependence. Allowing each coefficient to depend arbitrarily on *all* wavelet coordinates $y_t = (\langle X_t, w_1\rangle, \ldots, \langle X_t, w_n\rangle)$—i.e., learning n functions $f_i : \mathbb{R}^n \to \mathbb{R}$ or even degree–D multivariate polynomials—leads to combinatorial parameter growth and brittle estimation. We therefore restrict to *structured*, computationally tractable, and interpretable dependencies (diagonal/independent, band-tied, and local-coupled), which retain closed-form or efficient normal-equation solvers while capturing the dominant correlations.

2.2.2 Wavelet Band Coupling

One very natural form of correlation is to allow wavelet coefficients at the same scale j and location k to depend on one another. Let $y_0 = \langle X_t, \psi^{(0)}_{j,k}\rangle, \quad y_1 = \langle X_t, \psi^{(1)}_{j,k}\rangle, \quad y_2 = \langle X_t, \psi^{(2)}_{j,k}\rangle$, At each scale/location in the detail bands (j, k, l), we assume a polynomial coefficient form of degree D:

$$\begin{aligned}
\zeta^{(t)}_{j,k,0}(X_t) &= C^{(t)}_0 + \sum_{a=1}^{D} \sum_{b=0}^{D-a} \sum_{c=0}^{D-a-b} \beta^{(0)}_{a,b,c}\, y_0^a y_1^b y_2^c, \\
\zeta^{(t)}_{j,k,1}(X_t) &= C^{(t)}_1 + \sum_{b=1}^{D} \sum_{a=0}^{D-b} \sum_{c=0}^{D-b-a} \beta^{(1)}_{a,b,c}\, y_0^a y_1^b y_2^c, \\
\zeta^{(t)}_{j,k,2}(X_t) &= C^{(t)}_2 + \sum_{c=1}^{D} \sum_{a=0}^{D-c} \sum_{b=0}^{D-c-a} \beta^{(2)}_{a,b,c}\, y_0^a y_1^b y_2^c.
\end{aligned} \tag{18}$$

The constraint ensures each monomial contains the target coordinate y_ℓ at least once; cross terms are allowed but pure "other-orientation" terms are excluded.) We define our θ coefficients as polynomials with the same optimal values as in the independent case. Estimation proceeds via the same normal-equation machinery as in (11), but with mixed moments at (j, k):

$$\mu^{(j,k)}_{pqr}(t) := \mathbb{E}\big[\, y_0^p y_1^q y_2^r \,\big], \qquad p + q + r \le D + 1,$$

which, under the forward process with orthonormal $\{\psi^{(\ell)}_{j,k}\}$, expand into clean-data mixed moments and factorized Gaussian noise moments.

2.2.3 Local Coupling

Another natural choice of coupling is to allow wavelets in the same local neighborhood. For a radius $r \in \mathbb{N}$, define the neighborhood $\Delta_r = \{\delta \in \mathbb{Z}^2 : \|\delta\|_\infty \le r,\ \delta \ne (0,0)\}$ We allow wavelets in the same neighborhood to interact. Fixing a scale j and orientation $\ell \in \{0, 1, 2\}$, let the oriented wavelet/detail coefficient at spatial location $k \in \mathcal{K}$ be $y_k = \langle X_t,\ \psi^{(\ell)}_{j,k}\rangle$. Let $D \ge 1$ be the total degree. Define $S_D = \{(d, e) \in \mathbb{N}^2 :\ d, e \ge 1,\ d + e \le D\}$. We define the functional forms of θ as follows:

$$\theta_{J_0,k}(X_t, t) \;=\; \sum_{i=0}^{D} \alpha_i \left\langle X_t,\ \phi_{J_0,k}\right\rangle^i \;+\; \sum_{\delta\in\Delta_r} \sum_{(d,e)\in S_D} \beta_{\delta,d,e} \left\langle X_t,\ \phi_{J_0,k}\right\rangle^d \left\langle X_t,\ \phi_{J_0,k+\delta}\right\rangle^e \tag{19}$$

where α_i and $\beta_{\delta,d,e}$ are the parameters over which we optimize. Similarly for ζ we define

$$\zeta_{j,k,l}(X_t, t) = \sum_{i=0}^{D} \xi_i \left\langle X_t, \psi_{j,k,l}\right\rangle^i \;+\; \sum_{\delta\in\Delta_r} \sum_{(d,e)\in S_D} \omega_{\delta,d,e} \left\langle X_t,\ \psi_{j,k,l}\right\rangle^d \left\langle X_t,\ \psi_{j,k+\delta,l}\right\rangle^e \tag{20}$$

where ξ_i and $\omega_{\delta,d,e}$ are the parameters over which we optimize. As in the wavelet band coupling, estimation proceeds via the same machinery as in (11), but with mixed moments.

3 Experiments

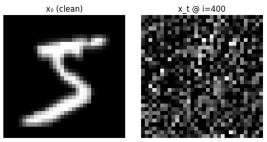


Figure 1: Original and noised MNIST image.

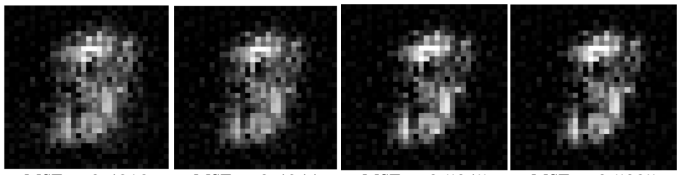
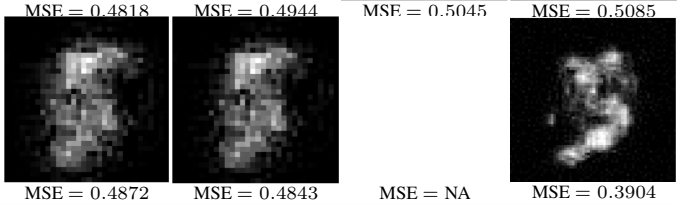
Model	Linear	Quadratic	Cubic	Quartic
Diagonal, Mono	 MSE = 0.4818	MSE = 0.4944	MSE = 0.5045	MSE = 0.5085
Diagonal, Hermite	 MSE = 0.4872	MSE = 0.4843	MSE = NA	MSE = 0.3904

Table 1: MSE with thumbnail per independent model; reconstruction MSE shown below each thumbnail.

With these closed-form equations and model families in hand, we now test which scales/orientations matter across noise levels and how much band/local coupling improves over the independent baseline. We first test the quality of the denoising.

We resize MNIST images so that they have dimension $(32, 32)$ and linearly scaled to $[-1, 1]$. We use compactly supported Daubechies (db2) wavelets with periodized boundaries. The 2-D tensor-product basis $\mathcal{B}_{J_0}$ includes scaling atoms at the coarsest kept scale J_0 and detail atoms for $j \geq J_0$ up to $J_{\max} = \lfloor \log_2 \min(H, W) \rfloor$ (here $J_{\max} = 3$ at 32×32 in order to avoid boundary effects) with three orientations $\ell \in \{0, 1, 2\}$.

We adopt the variance preserving noise schedule which gives $X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t}\, Z$, and discretize the time interval $[0, 1]$ into $N = 500$ points. All models are fit independently for each time point. We add a ridge term to improve numerical stability, and we also experiment with the probabilists' Hermite polynomials to expand our θ and ζ with increased stability. The probabilists' Hermite polynomials $\{\mathrm{He}_n(x)\}_{n \geq 0}$ are defined by $\mathrm{He}_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}$ and are orthogonal for $Z \sim \mathcal{N}(0, 1)$ with $\mathbb{E}[\mathrm{He}_m(Z)\mathrm{He}_n(Z)] = n!\, \delta_{mn}$. Using He_n in place of raw monomials makes feature vectors orthogonal in expectation under near-Gaussian coordinates $y_t^{(i)}$, reducing covariance and bringing $\Sigma_i = \mathbb{E}[\varphi_i \varphi_i^\top]$ closer to diagonal (hence a smaller condition number). In practice this improves the conditioning of the matrix equations (9) and yields coefficients less sensitive to scaling and polynomial degree, especially when combined with a small ridge penalty.

3.1 Denoising

We begin by denoising one MNIST digit with all six models, across four degrees. We use the same hyper parameters for each denoising and the same set of $8,000$ MNIST images for our method of moments estimates. We use the $db2$ family of wavelets with $J_0 = 1$ and $J_{\max} = 3$, with ridge coefficient $\gamma = 0.02$. The original image and noised version are shown in Figure 3.

Changing the Degree of the Approximation As evidenced in Table 1 increasing the degree of the expansion in the independent case only helps marginally for independent monomials. This suggests that higher moments of the data are only valuable in the context of correlation information. In contrast, The Hermite expansion decouples features in expectation and delivers clearer improvements as D grows; however, the cubic system was ill-conditioned for this sample at the chosen γ and failed to produce an image (reported as "NA").

Detail Band Correlation As shown in Table 2 (right), coupling the three orientations at a location improves visual quality across bases and degrees and often reduces MSE relative to the independent counterpart. Qualitatively, edges and corners become sharper, consistent with the hypothesis that cross-orientation co-activation carries most of the local predictive signal and is especially sensitive

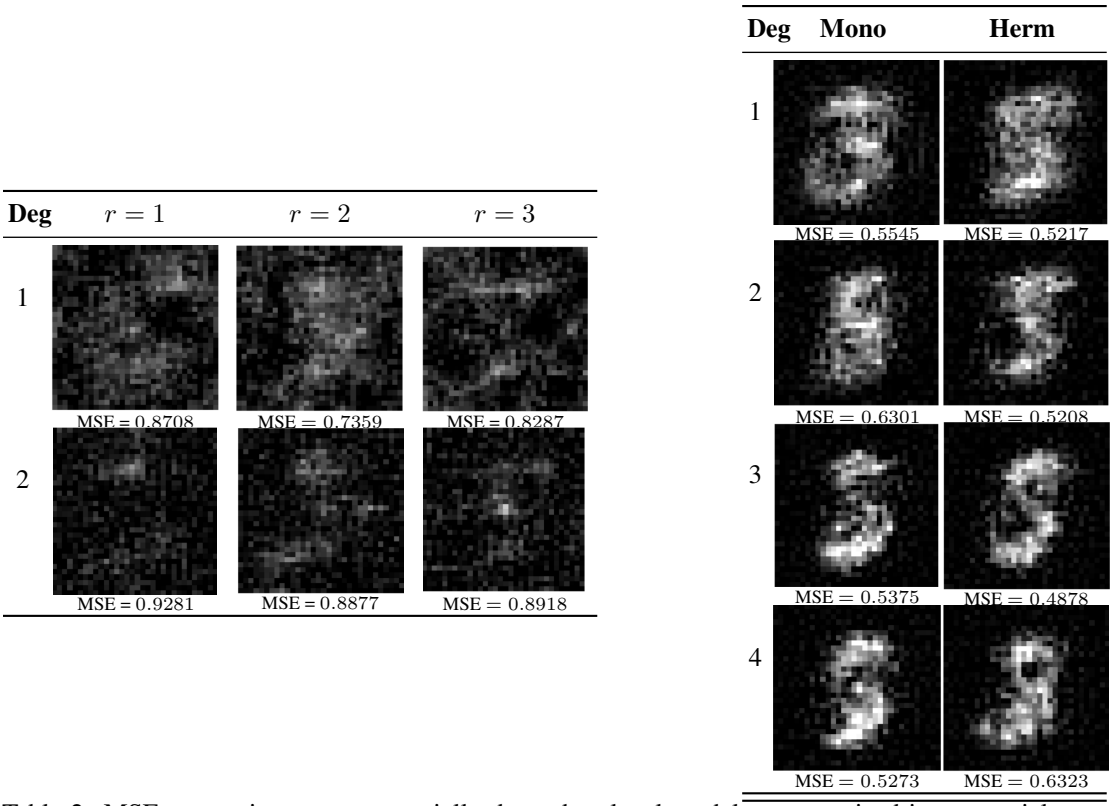


Table 2: MSE comparisons across spatially dependent local models parametrized in monomials (varying receptive field size r, left) and band-tied models (right).

to sharp changes like corners. We also observe a degree sweet spot: moving from $D{=}1 \to 3$ generally helps, while $D{=}4$ sometimes degrades (e.g., Hermite $D{=}4$), indicating overfitting or residual conditioning issues.

Local Correlation Table 2 (left) shows that, with *monomial* features and fixed γ, adding spatial neighbors can hurt MSE at this sample/setting. This is not entirely unexpected, since the feature dimension d_i grows with r and D, inflating the condition number of the matrix without enough data to support the extra parameters. We therefore anticipate local coupling to pay off only after (i) switching to Hermite features or (ii) increasing γ (or using truncated-SVD) and tuning r; in digits, small r should suffice, whereas texture-rich datasets may benefit from larger r.

Limitations and Further Experiments While these results are preliminary, they serve a good example of the kind of hypotheses this theory can help to empirically validate or rule out. The results above are single-image illustrations. In the camera-ready version we hope to report test-set aggregates (mean $\pm$ 95% CI over 1k images, multiple seeds, etc) and include a ridge sweep to locate the stability/accuracy frontier for band-tied and local models, plus per-(t, j) heatmaps of ΔMSE that tie improvements to scale and timestep. We hypothesize that correlation contributions vary over the noise schedule, with different modes becoming salient at different times. Further denoising and a few generation results are available in the Appendix A.

4 Discussion

The above theory offers a promising first step towards a more grounded understanding of what dataset distribution features are most important in score learning across noise scales. We introduced an analytically tractable, wavelet–based parameterization of the score with closed-form moment equations and three structured dependency families (independent, band-tied, local). The theory

clarifies which distributional features (marginals, cross-orientation co-activation, and short-range spatial correlations) can matter at different noise levels. Preliminary denoising results are consistent with these predictions. While our experiments are limited to MNIST, the framework is general and may provide insight for more data-efficient engineering decisions. Extending to diverse datasets, reporting aggregates with uncertainty, and testing generation metrics are natural next steps that our closed-form diagnostics can directly inform.

Acknowledgments

This work was supported by the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. We thank the Kempner for access to compute resources. We are also grateful to the CRISP group at Harvard SEAS for many thoughtful conversations. EF thanks the Calvin Coolidge Presidential Foundation for support during her undergraduate studies.

References

[1] Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training, 2025. URL https://arxiv.org/abs/2505.17638.

[2] D.L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41 (3):613–627, 1995. doi: 10.1109/18.382009.

[3] Fabian Falck, Christopher Williams, Dominic Danks, George Deligiannidis, Christopher Yau, Chris Holmes, Arnaud Doucet, and Matthew Willetts. A multi-resolution framework for u-nets with applications to hierarchical vaes, 2023. URL https://arxiv.org/abs/2301.08187.

[4] Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling, 2022. URL https://arxiv.org/abs/2208.05003.

[5] Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

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

[7] Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations, 2024. URL https://arxiv.org/abs/2310.02557.

[8] Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models, 2024. URL https://arxiv.org/abs/2412.20292.

[9] Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model, 2024. URL https://arxiv.org/abs/2401.04856.

[10] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

[11] Matthew Niedoba, Berend Zwartsenberg, Kevin Murphy, and Frank Wood. Towards a mechanistic explanation of diffusion model generalization, 2025. URL https://arxiv.org/abs/2411.19339.

[12] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators, 2023. URL https://arxiv.org/abs/2211.16152.

[13] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015. URL http://proceedings.mlr.press/v37/sohl-dickstein15.html.

[14] Binxu Wang and John J. Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications, 2024. URL https://arxiv.org/abs/2412.09726.

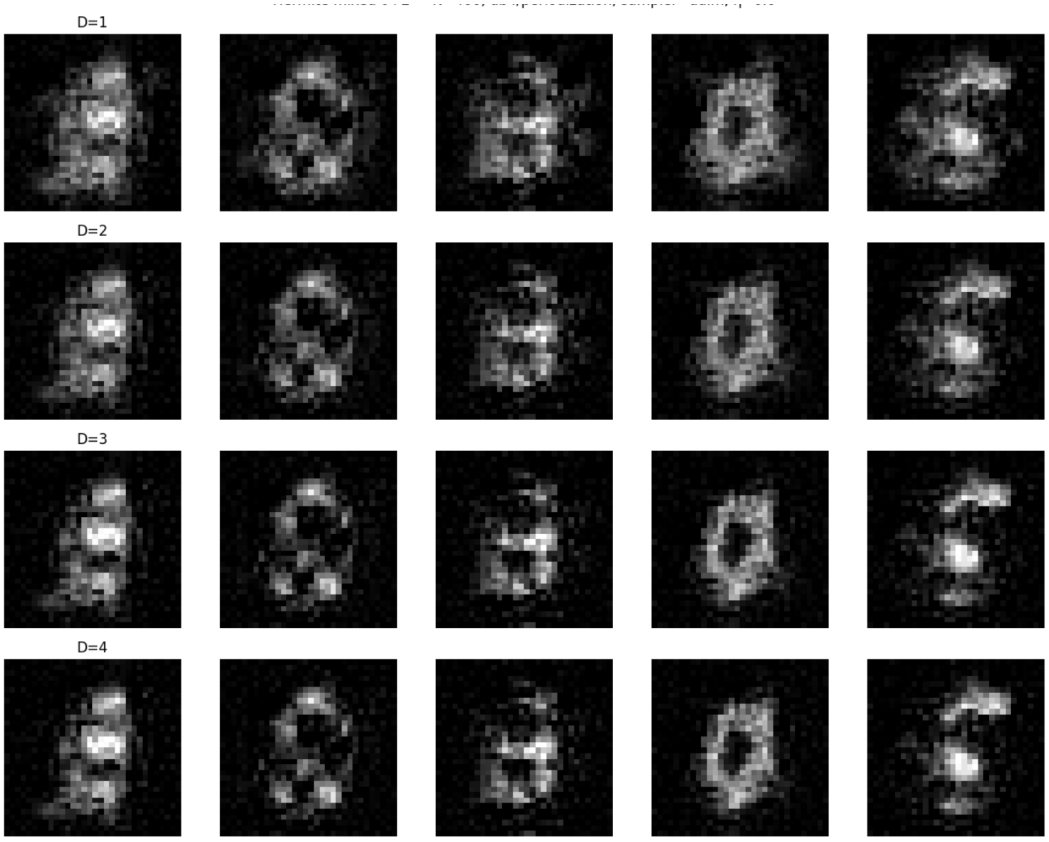


Figure 2: Original and noised MNIST image.

Appendix

A Additional Results

We also include generation results using the same setup as detailed in Section 3.1. In particular, we use band-coupling with 500 time steps, but the $db4$ wavelets, $J_0 = 1$, $J_{max} = 2$, and ridge 0.02. We use a ddim sampler. In the figures below, each column corresponds to a different seed value and reach row corresponds to the degree of the band-coupling Hermite model.

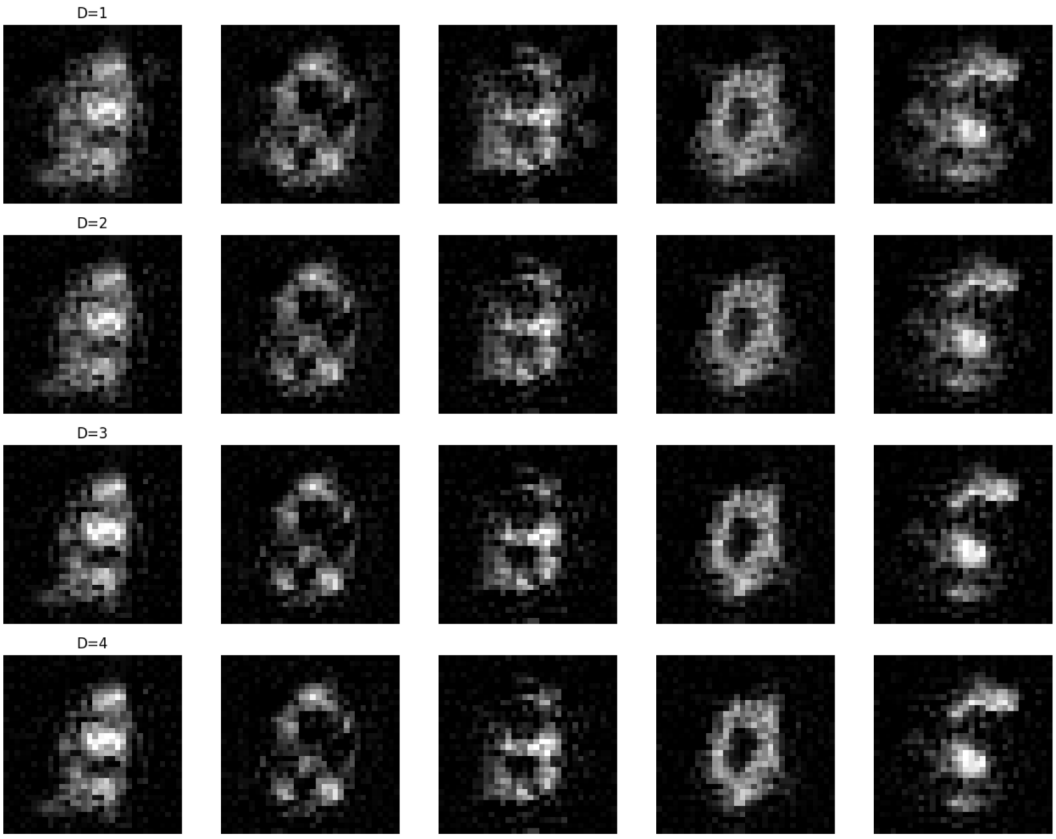


Figure 3: Original and noised MNIST image.