Generalization in diffusion models arises from geometric-adaptive harmonic bases

Zahra Kadkhodaie; Florentin Guth; Eero Simoncelli; Stéphane Mallat

New York University, Flatiron Institute, Collège de France

NYU Center for Data Science

FLATIRON INSTITUTE Center for Computational Neuroscience

Do diffusion models memorize the training set [Carlini et al, 2023, Somepalli et al, 2023] or generalize to learn continuous density models of images, despite the curse of dimensionality?

What are inductive biases of the denoiser which give rise to generalization?

Diffusion models and denoising

$y = x + z$ where $z \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$

$$p_\sigma(y) = \int p(y|x)\, p(x)\, \mathrm{d}x = \int g_\sigma(y-x)\, p(x)\, \mathrm{d}x,$$

Optimal denoiser: $f^\star(y) = \mathop{\mathbb{E}}_x[x|y] = \int x p(x|y) \mathrm{d}x \boxed{= y + \sigma^2 \nabla \log p_\sigma(y)}$

[Tweedie, via Robbins, 1956; Miyasawa, 1961]

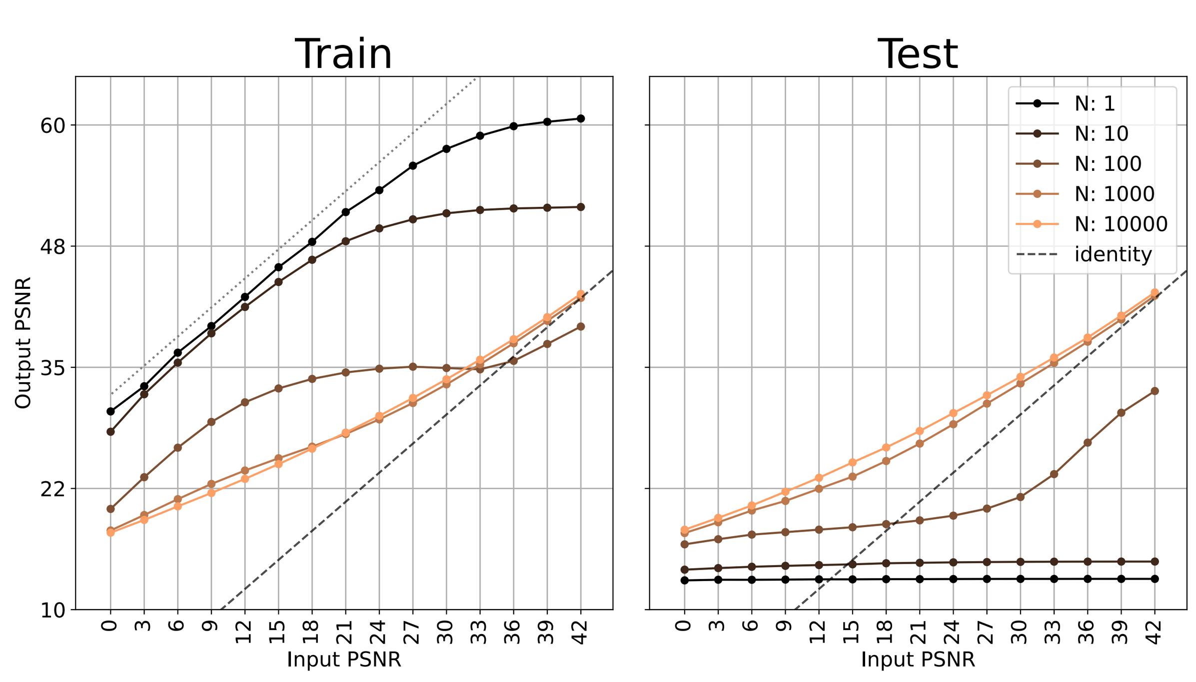
Empirical denoiser: $f_{\hat\theta}(y) \approx f^\star(y)$ where $\hat\theta = \arg\min_\theta \mathbb{E}\left[\|x - f_\theta(y)\|^2\right],$

$$\boxed{D_{\mathrm{KL}}(p(x)\,\|\,p_\theta(x)) \leq \int_0^\infty \left(\mathrm{MSE}(f_\theta, \sigma^2) - \mathrm{MSE}(f^\star, \sigma^2)\right) \sigma^{-3}\, \mathrm{d}\sigma,}$$

Model density error is bounded by the denoiser error, integrated over σ

Transition from memorization to generalization

in one shot denoising

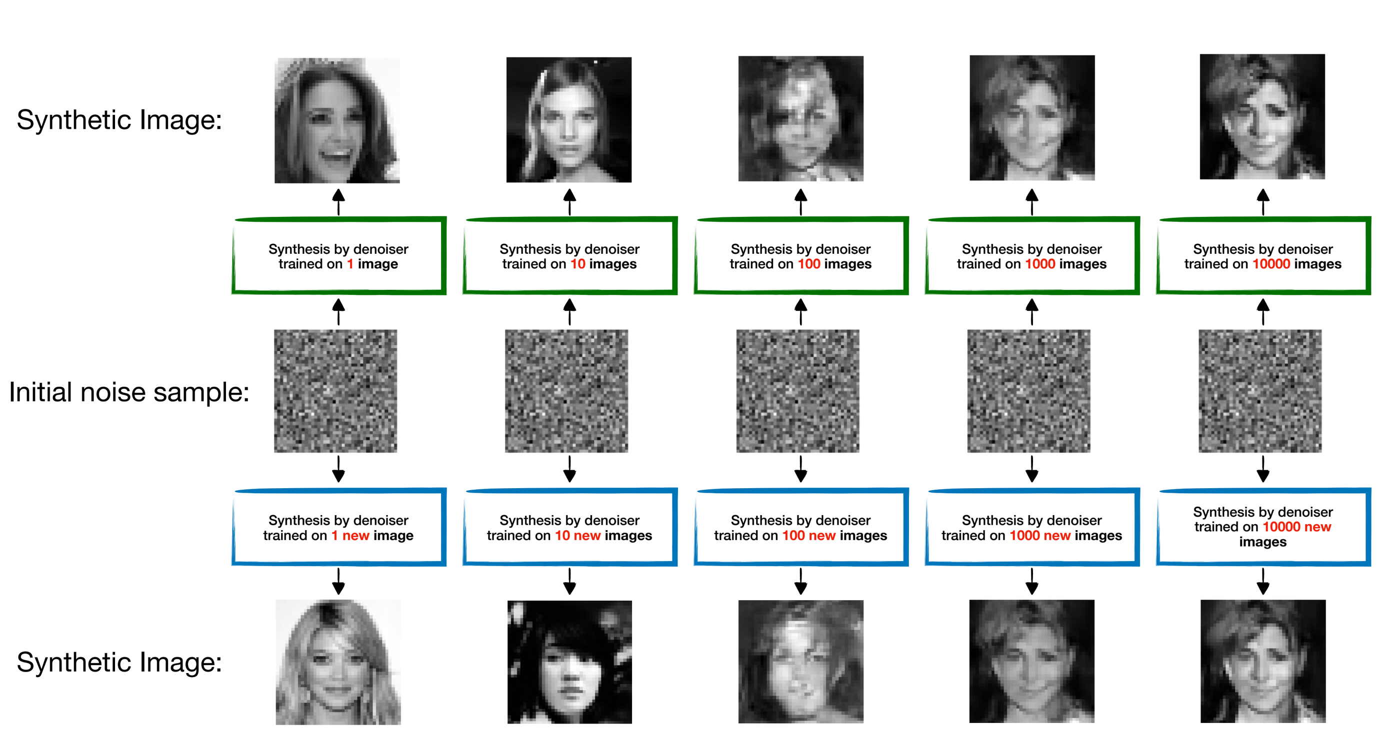


- Small training set size, N∈{1,10}: models memorize the training images.
- N = 100: peculiar performance behavior indicates a transition phase.
- N = 10,000: test performance matches train performance (classic test for overfitting).

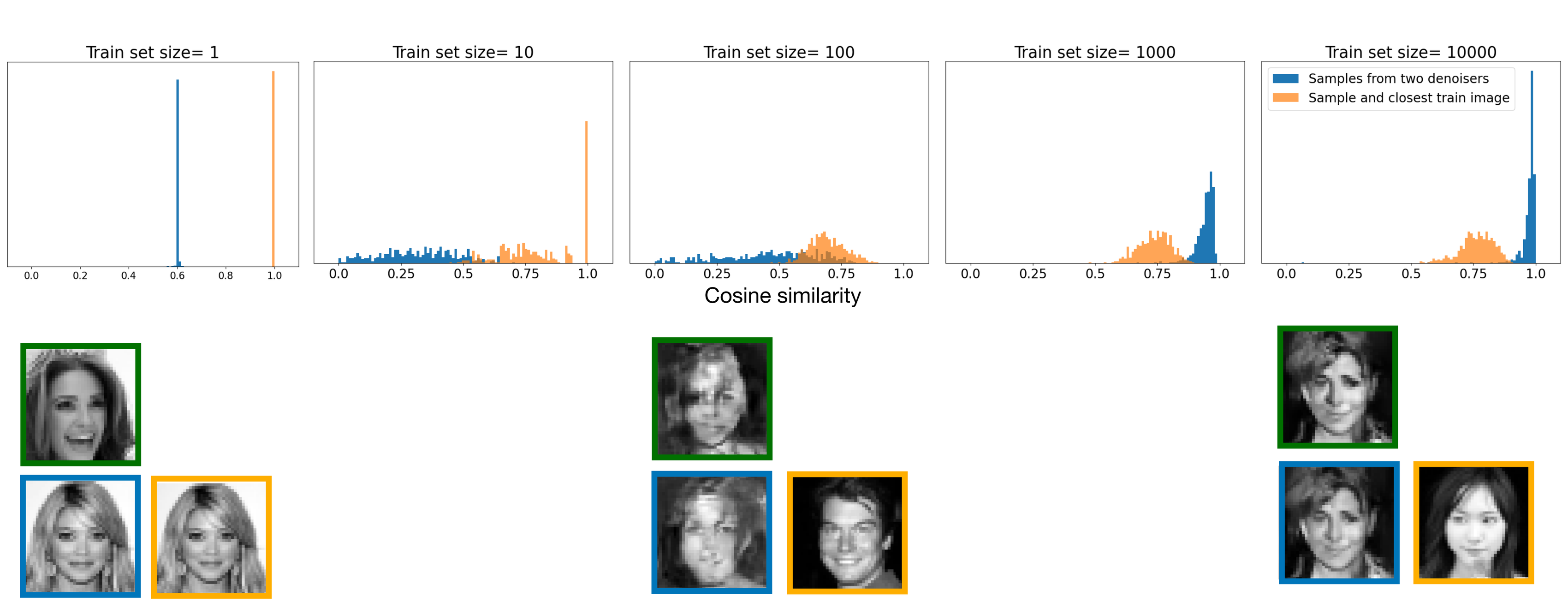
There is a transition phase from memorization to generalization and with enough data, the network enters the generalization regime

Strong generalization

in synthesis



Two denoisers trained on non-overlapping training sets converge to almost the same function
The model variance is tending to zero.



Denoising as shrinkage in a basis

Classical framework for denoising:

1. Transform the noisy image to a basis where noise and signal are separable
2. Suppress the noise (shrinkage)
3. Transform back to the pixel domain

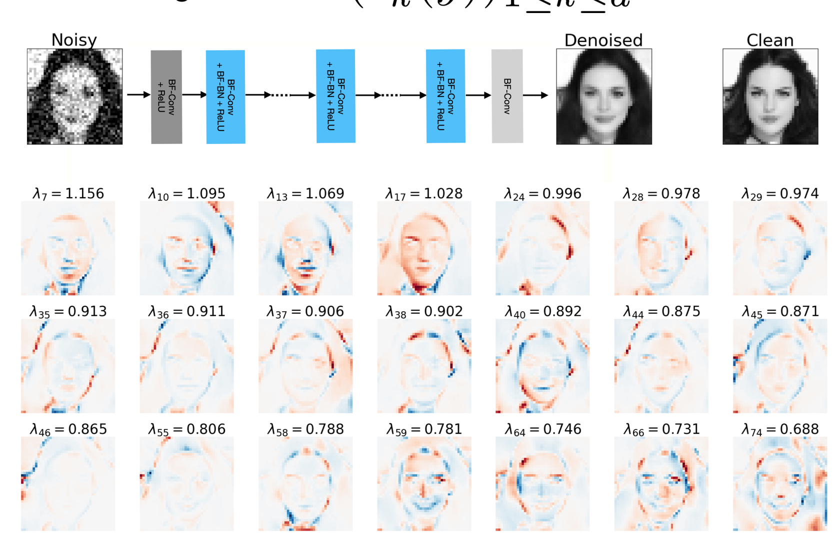
$f(y) = W_L R(W_{L-1} ... R(W_1 y)) = A_y y,$ Locally linear function (no additive constants in the network)

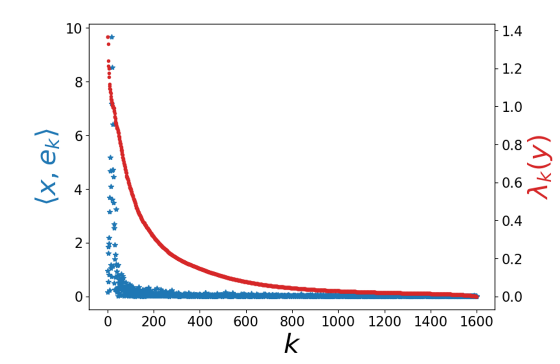
$f(y) = \nabla f(y)\, y$ Jacobian w.r.t. Input y (Nearly symmetric) [Mohan*, Kadkhodaie* et al 2020]

$$f(y) = \nabla f(y)\, y = \sum_k \lambda_k(y)\, \langle y, e_k(y)\rangle\, e_k(y)$$

Eigen basis: $(e_k(y))_{1 \leq k \leq d}$

Shrinkage factors: $(\lambda_k(y))_{1 \leq k \leq d}$



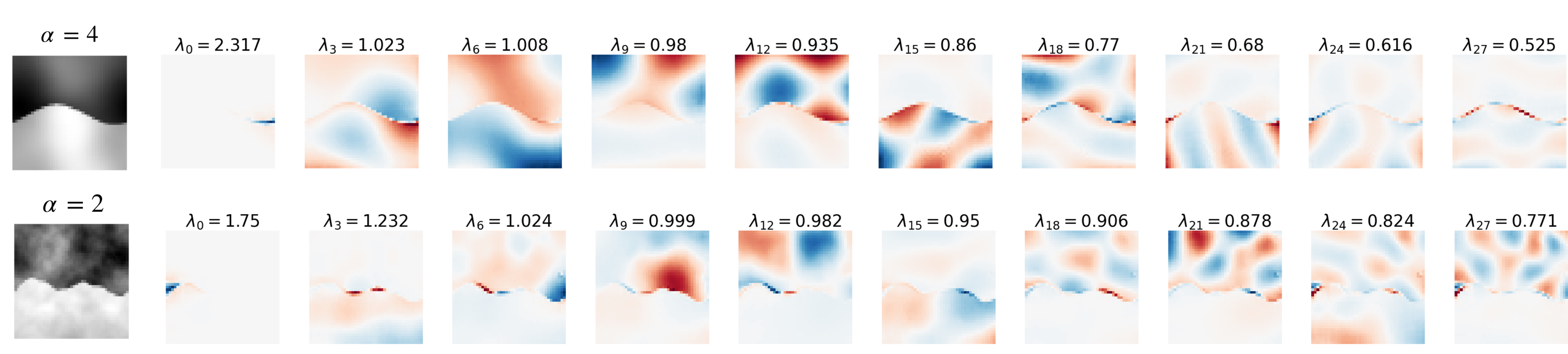


Oscillating patterns along the contours and in smooth regions
Eigen basis adaptive to the underlying image

Sparse coefficients → fast rate of decay of Eigenvalues → better denoising

Conjecture: DNN denoisers have inductive biases towards learning GAHBs

Aligned inductive biases and optimality



Geometric C^α images

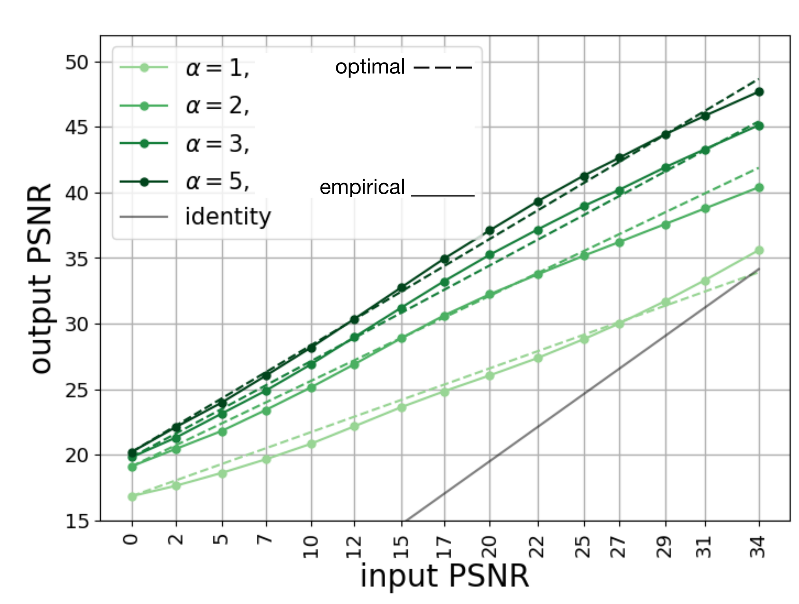
Optimal denoiser on C^α images has slope $\dfrac{\alpha}{\alpha+1}$.

[Korostelev & Tsybakov, 1993]

This optimal slope is also obtained by best "bandlet" basis denoising estimators.

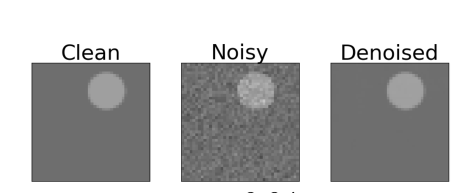
[Peyré & Mallat, 2008]

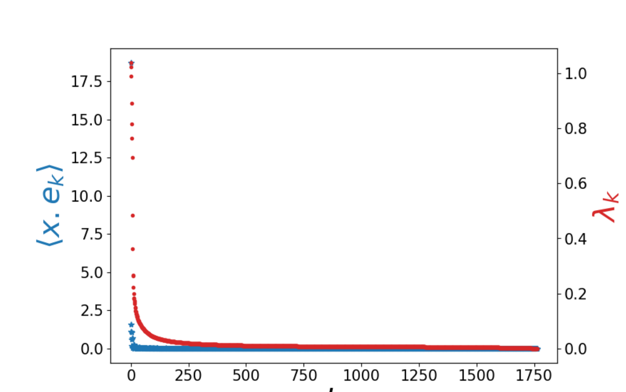
When the optimal basis is GAHB, DNN achieves optimal denoising: alignment



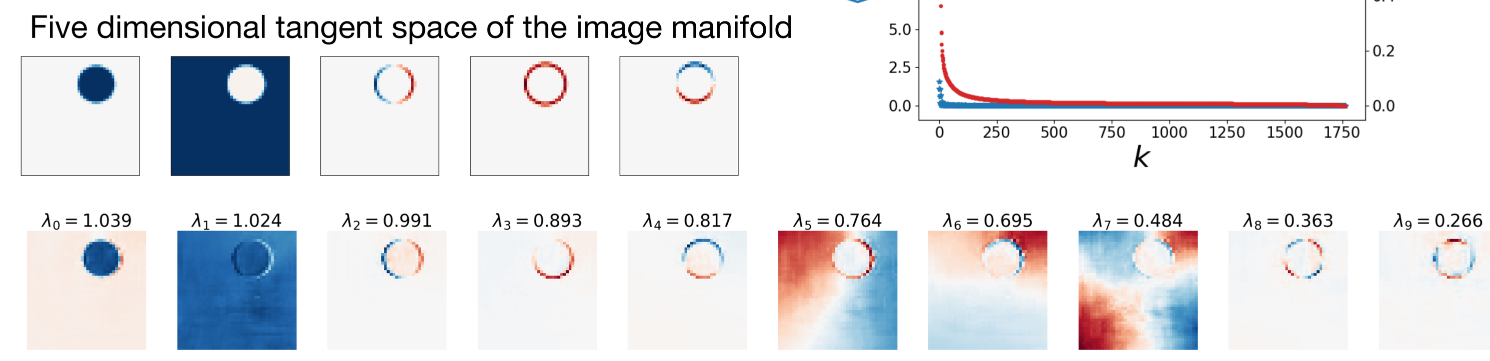
Mis-aligned inductive biases and sub-optimality

Manifold of discs: Varying positions, sizes, and foreground/background intensities. Defines a five-dimensional *curved* manifold





Five dimensional tangent space of the image manifold



closely match the optimal basis

GAHB vectors with non-zero λ_k

The rest of the basis can be arbitrary, but it produces GAHB vectors, and it does not cut them off.

DNNs achieve this generalization through an inductive bias that favors shrinkage in a GAHB.