Schedule for today

8:30 – 9:10 Intro + notations

Coffee Break

9:20 – 10:30 What is the optimal denoiser? What is the impact of smoothing?

Coffee Break

10:50 - 12:00 How far can the linear perspective get us?

Lunch Break

13:30 - 14:30 Seeing generalization through locality

Coffee Break

14:40 - 15:10 Why do diffusion models learn to be local?

Coffee Break

15:20 - 16:00 How do diffusion models learn and encode global structures?

Coffee Break

16:10 - 17:00 Open question in the field + Q&A

Social and chat

for schedule, code, slides, and recordings!

analytic-diffusion.github.io

Diffusion Models Through the Linear Lens

Sampling, Consistency, Receptive Fields, and Learning

Binxu Wang

Kempner Institute; Harvard Medical School

CVPR 2026

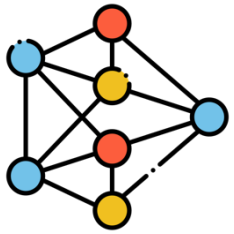
CVPR JUNE 3-7, 2026 DENVER COLORADO

Paradigm of Generative AI

Data pile

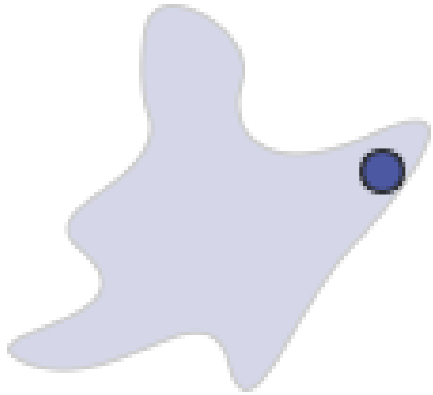
Image, Video, Molecule, …

$p(\mathbf{x}_0)$



Learning system
(Diffusion, Autoregressive / LLM)

Learned distribution



$\tilde{\mathbf{x}} \sim p_{gen}$

Sampling generative model

How does generated distribution relate to training one, due to learning system?

Training Diffusion Model

Learning to denoise

Noisy sample

$\mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$

$\mathbf{D}_\theta(., \sigma)$

Neural network

Estimated Clean Sample

$\mathbf{D}_\theta(\mathbf{x_0} + \sigma\boldsymbol{\epsilon}, \sigma)$

$$\mathcal{L}_{DSM,\sigma} = \mathbb{E}_{\substack{\mathbf{x_0} \sim p_0(x) \\ \boldsymbol{z} \sim \mathcal{N}(0,I)}} \|\mathbf{D}_\theta(\mathbf{x_0} + \sigma\boldsymbol{\epsilon}, \sigma) - \mathbf{x_0}\|^2$$

Conventions:

EDM, variance exploding parametrization

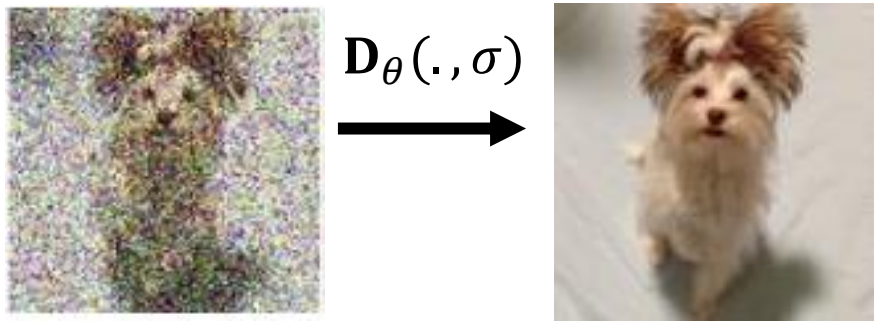
Xo prediction

Denoising score matching

Learning to denoise $\Leftrightarrow$ Smooth the density $p(\mathbf{x};\sigma)$ and learning its gradient, i.e. **score function** $\nabla \log p(\mathbf{x};\sigma)$

$$\nabla \log p(\mathbf{x};\sigma) = \frac{\mathbf{D}^*(\mathbf{x},\sigma) - \mathbf{x}}{\sigma^2}$$

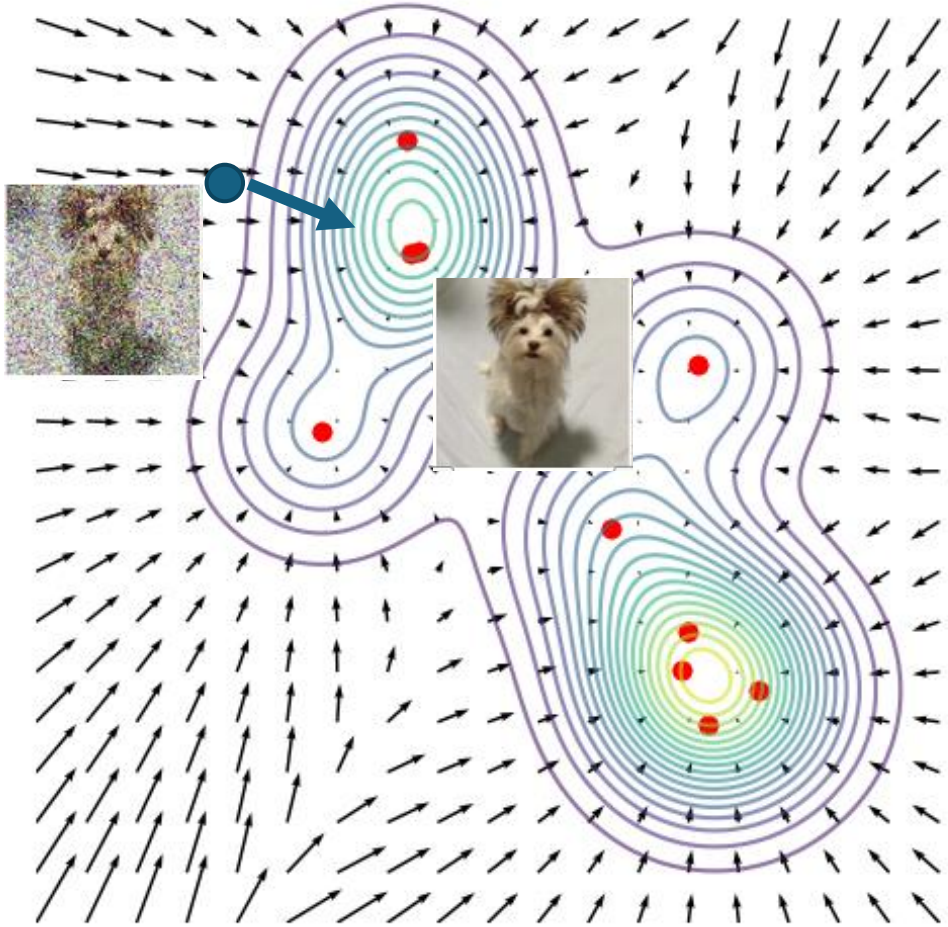
Tweedie's formula



Noisy sample $\mathbf{x}_0 + \sigma\boldsymbol{\epsilon}$

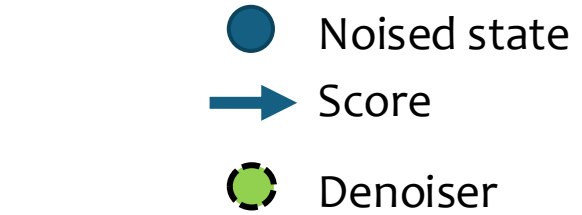
Estimated Clean Sample $\mathbf{D}_\theta(\mathbf{x_0} + \sigma\boldsymbol{\epsilon}, \sigma)$

$$\arg\min \mathcal{L}_{DSM,\sigma} \Rightarrow \mathbf{D}^*$$



Robbins, 1956, An empirical Bayes approach to statistics.
Vincent, 2011, *Neural Computation*

Score and denoiser



- Score

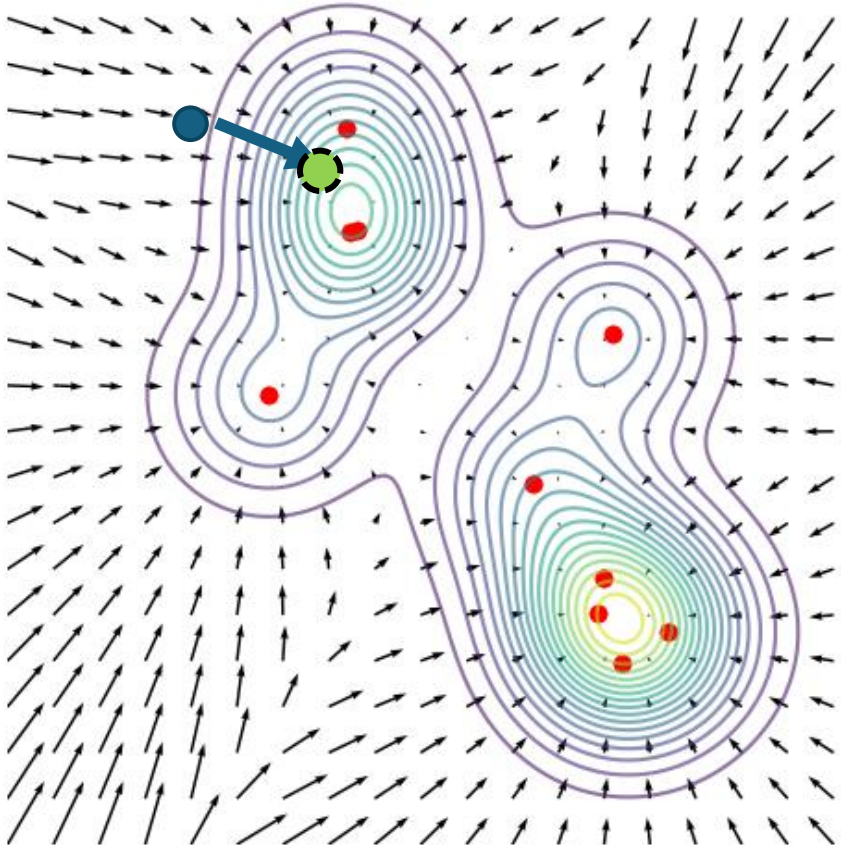
$$\mathbf{s}(\mathbf{x}, \sigma) \coloneqq \nabla \log p(\mathbf{x}; \sigma)$$

- Denoiser

$$\mathbf{D}^*(\mathbf{x}, \sigma) \coloneqq \mathbb{E}[\mathbf{x}_0 | \mathbf{x}] = \mathbf{x} + \sigma^2 \nabla \log p(\mathbf{x}; \sigma)$$

Bayes optimal estimate of clean sample

"Posterior mean"



Robbins, (1992) "An empirical Bayes approach to statistics."
Jonathan, Jain, Abbeel, (2020) NeurIPS, DDPM

Sampling Diffusion models

- Deterministic sampler: Probability-Flow Ordinary Differential Equation (PF-ODE)

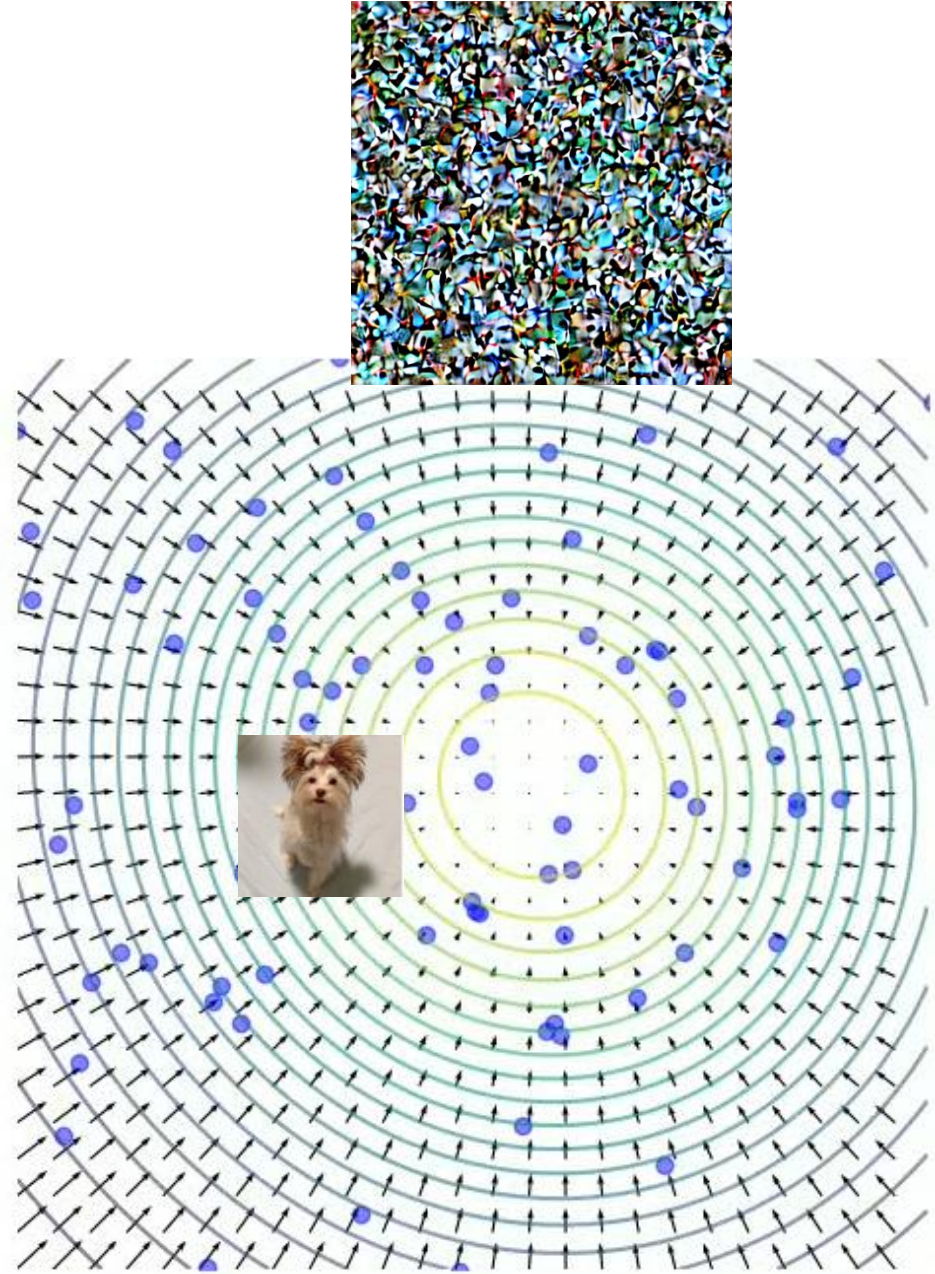
Gradient of noised data distribution

$$\frac{d\mathbf{x}}{d\sigma} = -\sigma \nabla \log p(\mathbf{x}; \sigma)$$

$$\approx -\frac{\mathbf{D}_{\boldsymbol{\theta}}(\mathbf{x}, \sigma) - \mathbf{x}}{\sigma}$$

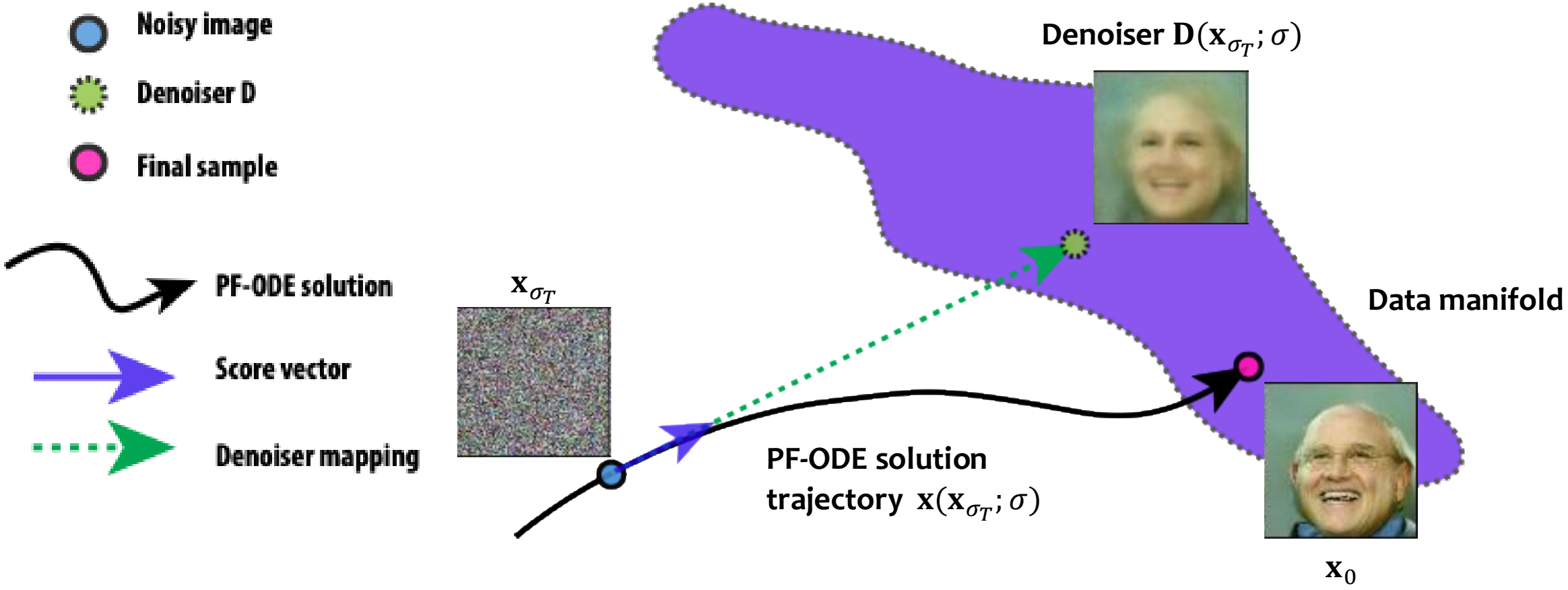
Neural network trained to denoise (DSM)

Initialize $\mathbf{x}_{\sigma_T} \sim \mathcal{N}(0, \sigma_T^2 I), \sigma_T \gg 1$
Integrate from $\sigma_T \to \sigma_0 \approx 0$



Karras, T., Aittala, M., Aila, T., & Laine, S. (2022), *NeurIPS*. EDM

Relation between Denoiser and Generated Sample



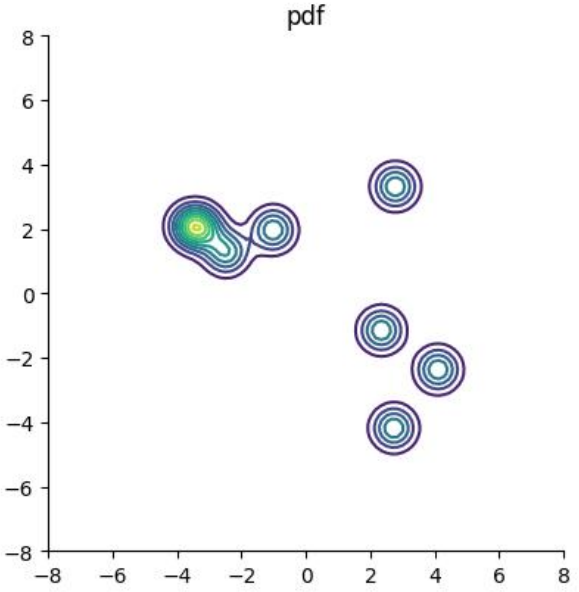
Denoiser as one-step look ahead of final sample.

Mystery of diffusion generalization

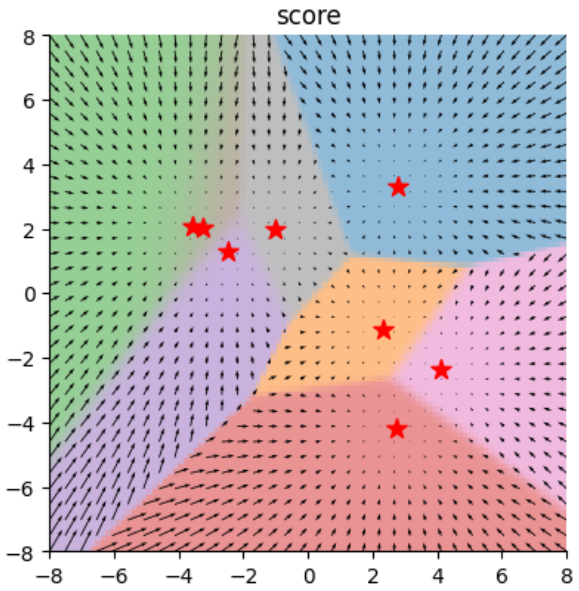
Training data
$p(\mathbf{x}_0)$

$$p_{delta}(\mathbf{x}) = \frac{1}{N}\sum_i \delta(\mathbf{x} - \mathbf{x}_i)$$

$$\mathbf{s}_{delta}(\mathbf{x};\sigma) = \frac{1}{\sigma^2}\sum_i w_i(\mathbf{x})(\mathbf{x}_i - \mathbf{x})$$



Global minimizer of DSM without constraint
$\arg\min_{\boldsymbol{s}(.)} \mathcal{L}_{DSM}$



Sampling with PFODE

$p_{delta}(\mathbf{x})$

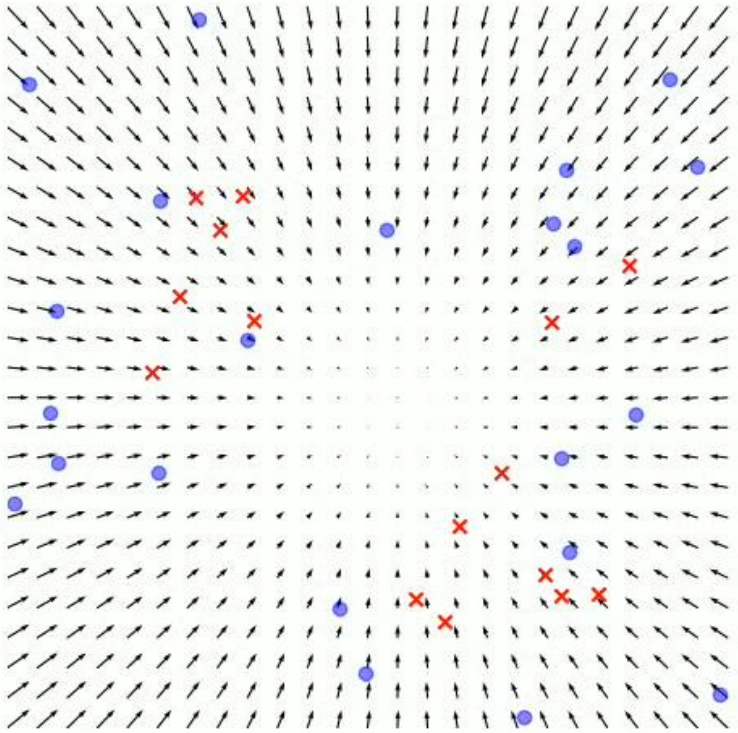
Generate training dataset again?
Useless!!

This does not happen (generally)!
But what does model actually learn?

Learner with infinite capacity: diffusion model = associative memory

Diffusion model ≡ **Memory retrieval**

$$\arg\min \mathcal{L} \Rightarrow \mathbf{D}_\infty$$



Generating sample = converging to the closest example from training set

x training samples $\{\mathbf{x_i}\}$

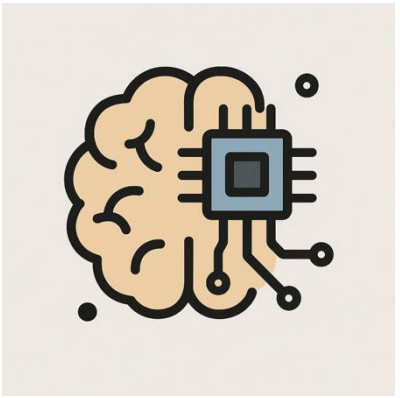
o dynamics of generated sample

c.f. Modern Hopfield Network

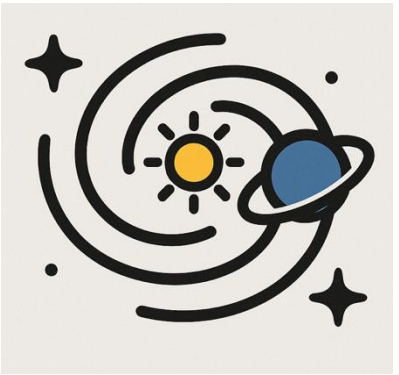
Krotov, D., & Hopfield, J. (2020).
Ramsauer, H., …& Hochreiter, S. (2020).

Simplify reality into minimal nontrivial models.

Achieve analytical understanding of it.



Physics of AI



Linear case of Diffusion model

Hidden linear structure in diffusion score

- Sampling dynamics
- Receptive Field Structure
- Cross split Consistency
- Learning dynamics

Why linear is interesting? (I)

Linear Score ~ Gaussian statistics of data

Gaussian

$$p_{gauss}(\mathbf{x}) = \mathcal{N}(\mu, \Sigma)$$

Energy function is Quadratic

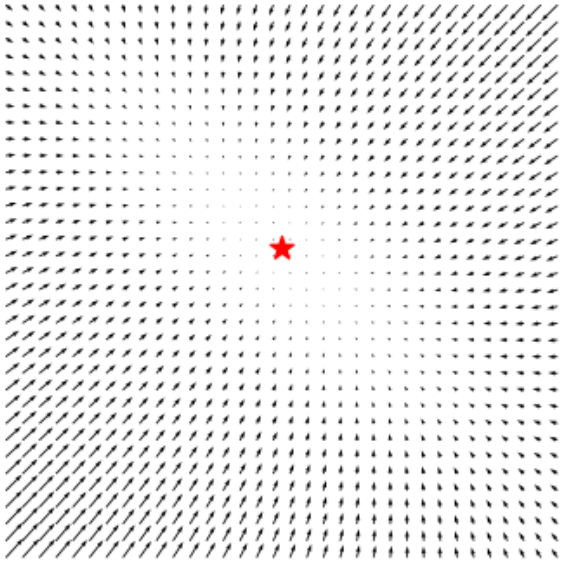
$$E_{gauss}(\mathbf{x}) \propto \frac{1}{2}(\mu - \mathbf{x})^{\top}\Sigma^{-1}(\mu - \mathbf{x})$$

pdf

Score function is Linear

$$\mathbf{s}_{gauss}(\mathbf{x}) = \Sigma^{-1}(\mu - \mathbf{x})$$

score



Gaussian score recovers Wiener filter

Noising model
(Var Exploding)

$$\mathbf{x}_\sigma = \mathbf{x} + \sigma \boldsymbol{z}$$
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

At noise scale σ, $p_{gauss}(\mathbf{x}_\sigma; \sigma) = \mathcal{N}(\mu, \Sigma + \sigma^2 I)$

Score function $\mathbf{s}_{gauss}(\mathbf{x}_\sigma; \sigma) = (\Sigma + \sigma^2 I)^{-1}(\mu - \mathbf{x}_\sigma)$

Denoiser function $\mathbf{D}_{gauss}(\mathbf{x}_\sigma; \sigma) = \mu + \Sigma(\Sigma + \sigma^2 I)^{-1}(\mathbf{x}_\sigma - \mu)$

Wiener filter (1967)

Denoiser under an arbitrary Gaussian prior

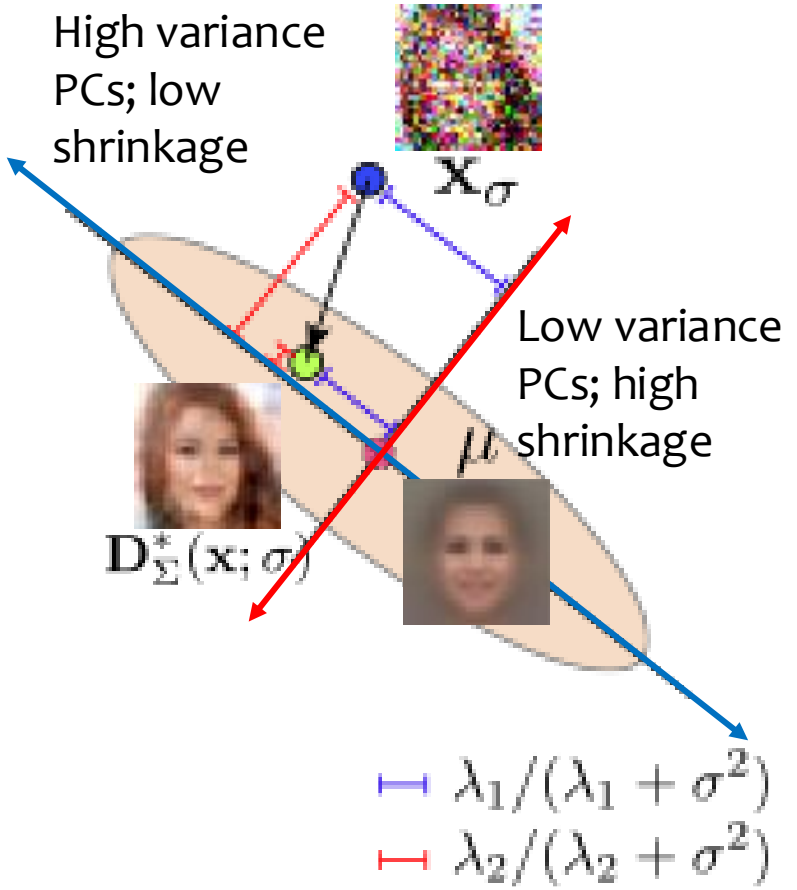
Geometry of Wiener filter

- Wiener filter shrink the projection along each PC according to their signal to noise ratio.

$$\mathbf{D}_{gauss}(\mathbf{x};\sigma) = \mu + \sum_k \frac{\lambda_k}{\lambda_k + \sigma^2} \mathbf{u}_k \mathbf{u}_k^\top (\mathbf{x} - \mu)$$

Shrinkage factor (λ_k variance of PC k)

$\mathbf{u}_k$ PC of dataset



Why linear is interesting? (II)

Gaussian score ⟺ Optimal under linear constraint

For arbitrary dataset,

Linear function approximator

$$\mathbf{D}_\theta(\mathbf{x};\sigma) = W_\sigma \mathbf{x} + b_\sigma$$

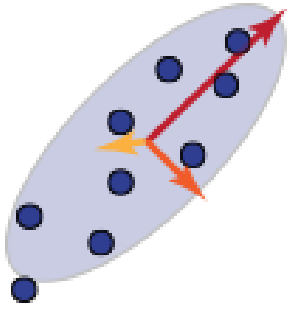
Optimize denoise score matching (DSM) loss

$$W_\sigma^*, b_\sigma^* = \arg\min_{W,b} \mathcal{L}_{DSM,\sigma}$$

$$\mathbf{D}^*(\mathbf{x};\sigma) = \boldsymbol{\mu} + \Sigma(\sigma^2 I + \Sigma)^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Wiener filter (1967)

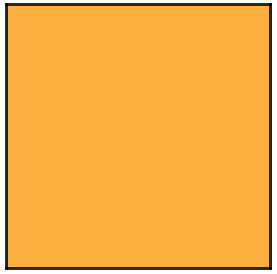
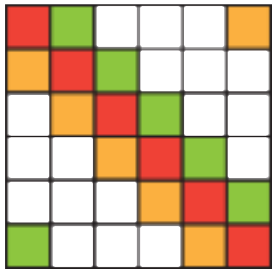
From the blurry eye of linear "network", any data looks Gaussian.



Li, Dai, Qu, (2024). *NeurIPS*

Wang, Pehlevan, (2025), *NeurIPS spotlight*

Duality of Architecture Constraint and Learned distribution

	Score network architecture		Learned distribution	
Linear network	$W_\sigma \mathbf{x} + b_\sigma$		Gaussian	$\mathcal{N}(\mu, \Sigma)$
Linear conv network	$w_\sigma * \mathbf{x}$		Stationary Gaussian Process	$\mathcal{GP}(\tilde{\Sigma})$
Local Patch conv network	$w_\sigma * \mathbf{x}$		Stationary Gaussian Process (local kernel)	$\mathcal{GP}(\tilde{\Sigma}')$

Wang, Pehlevan, (2025), *NeurIPS spotlight*

Dual view of linear score

Architectural constraints

Dataset statistics

Constrained architecture can only learn certain order of statistics (e.g. Gaussian).

Gaussian statistics can induce linear structure in score

Why linear is interesting? (III)

Gaussian score admits closed-form sampling path solution

Probability flow ODE Score of Gaussian $\mathbf{s}_{gauss}$

$$\frac{d\mathbf{x}}{d\sigma} = -\frac{1}{\sigma}\sum_k \frac{\sigma^2}{\lambda_k + \sigma^2}\mathbf{u}_k\mathbf{u}_k^T(\mu - \mathbf{x})$$

Linear time-varying ODE

Dynamics matrices commute!

Sampling trajectory solution

$$\mathbf{x}(\sigma_t) = \mu + \sum_{\boldsymbol{k}} \sqrt{\frac{\sigma_t^2 + \lambda_k}{\sigma_T^2 + \lambda_k}}\mathbf{u}_k\mathbf{u}_k^T\left(\mathbf{x}_{\sigma_T} - \mu\right)$$

$\mathbf{u}_k$ PC of dataset
λ_k variance of PC

Wang, Vastola, TMLR, 2024
Pierret, Galerne, 2024

Implication for sampling dynamics
Generation follows variance order

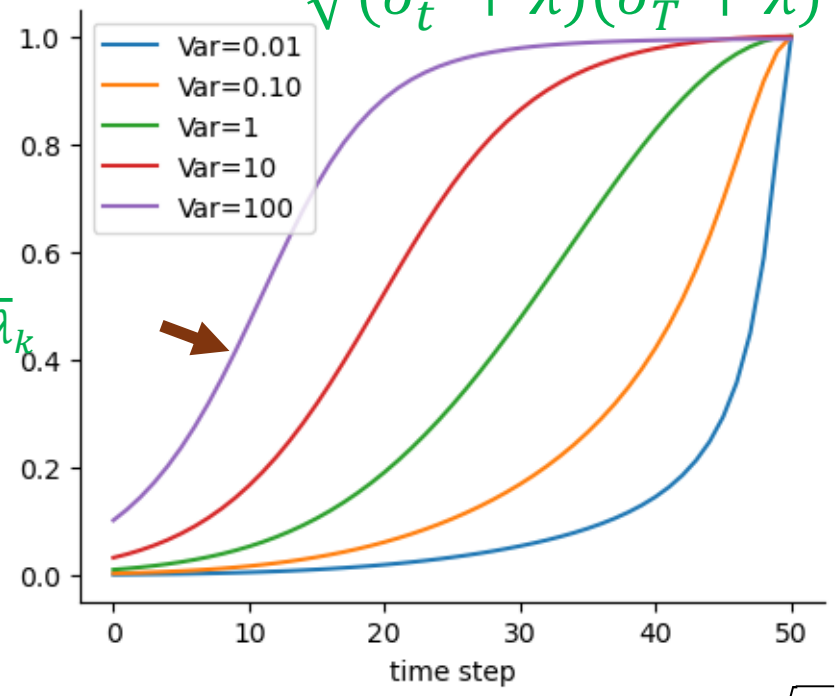
Dynamics of denoiser:

$$\mathbf{D}_{\text{gauss}}(\mathbf{x}_{\sigma_t};\sigma_t) = \underbrace{\mu}_{\text{Data mean}} + \sum_{k=1}^{r} \underbrace{\xi(t,\lambda_k)\mathbf{u}_k\mathbf{u}_k^T(\mathbf{x}_{\sigma_T}-\mu)}_{\text{Scaling of PC features}}$$

Scaling coefficient of each PC

$$\xi(t,\lambda) := \frac{\lambda}{\sqrt{(\sigma_t^2+\lambda)(\sigma_T^2+\lambda)}}$$

$\xi(t,\lambda)\sigma_T \ /\sqrt{\lambda_k}$



Normalized by $\sqrt{\lambda_k}/\sigma_T$

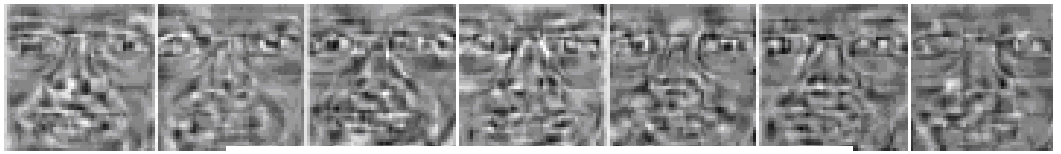
Statistical structure of natural image

$$\mathbf{x} \sim p_{face}$$

Distribution mean
represents average sample

Principal components
(PC) represent variations
of different levels / spatial
frequency

Higher variance PC



Lower variance PC

Burton, Moorhead (1987);
Field (1987, 1994);
Tolhurst et al. (1992)
Torralba, Oliva (2003)

Turk, Pentland (1991) Eigenface

Implication for image sampling dynamics

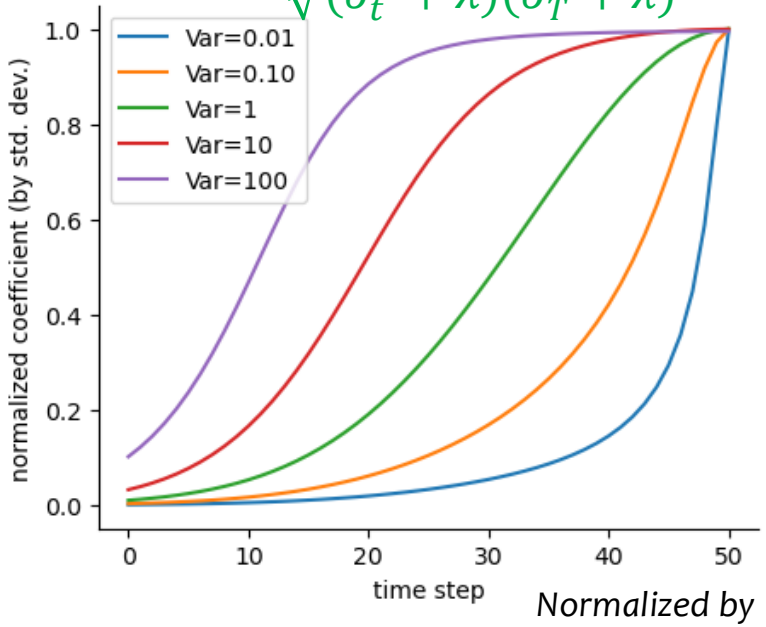
Gaussian theory predicts the spectral order of diffusion generation

Dynamics of denoiser:

$$\mathbf{D}_{\text{gauss}}(\mathbf{x}_{\sigma_t};\sigma_t) = \underbrace{\mu}_{\text{Data mean}} + \sum_{k=1}^{r} \underbrace{\xi(t,\lambda_k)\mathbf{u}_k\mathbf{u}_k^T(\mathbf{x}_{\sigma_T}-\mu)}_{\text{Scaling of PC features}}$$

$$\xi(t,\lambda) := \frac{\lambda}{\sqrt{(\sigma_t^2+\lambda)(\sigma_T^2+\lambda)}}$$

$\xi(t,\lambda)\sigma_T /\sqrt{\lambda_k}$



Normalized by $\sqrt{\lambda_k}/\sigma_T$

Starting from "Average" sample

Specifying low frequency info (hairstyle, face orientation)

Adding high frequency finer details.

Denoiser $\mathbf{D}(\mathbf{x}_t;\sigma_t)$

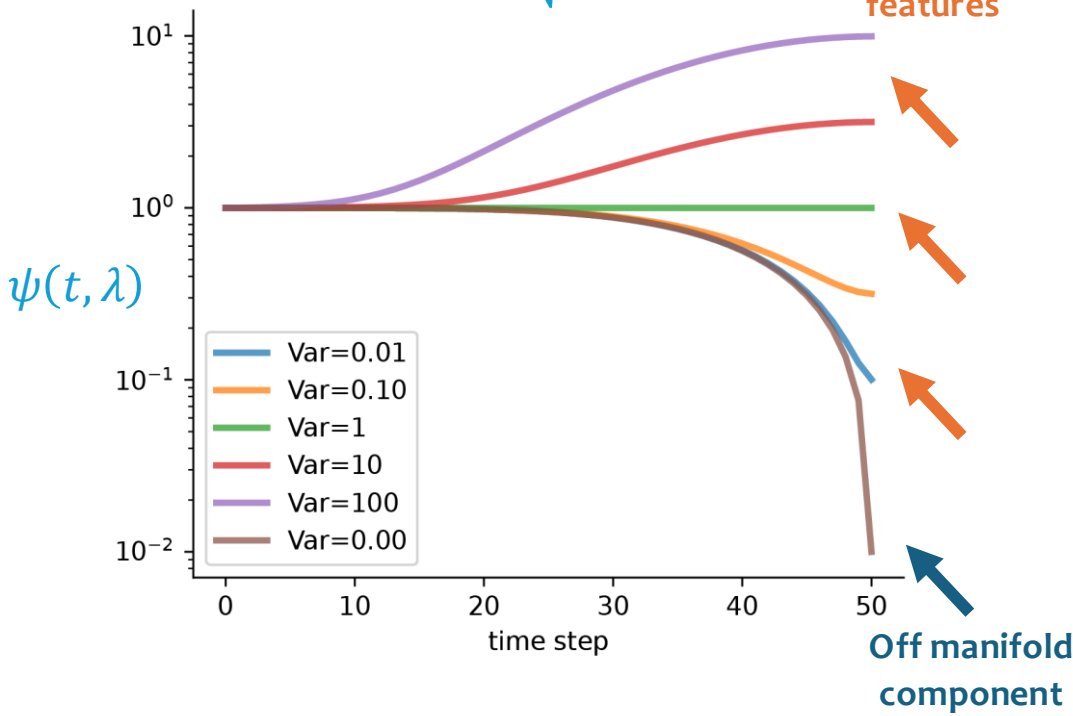
Gaussian analytical solution of $\mathbf{x}_t$ predicts evolution of noisy states

Dynamics of state (VP/ DDIM):

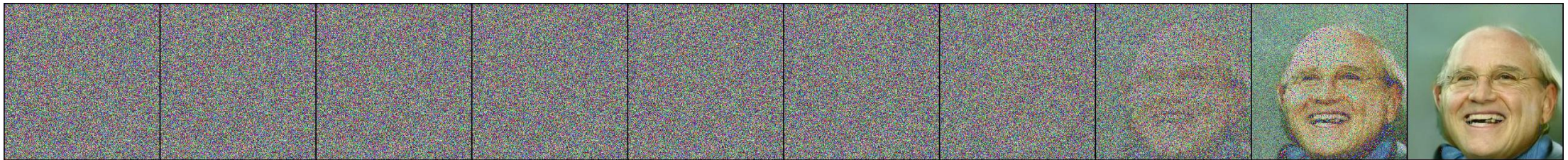
$$\mathbf{x}_t = \underbrace{\alpha_t \boldsymbol{\mu}}_{\text{Scaling mean}} + \underbrace{\psi(t,0)\mathbf{x}_T^{\perp}}_{\text{Off-manifold component}} + \underbrace{\sum_{k=1}^{r} \psi(t,\lambda_k) c_k(T)\mathbf{u}_k}_{\text{On-manifold feature components}}$$

$$\mathbf{x}_T^{\perp} \coloneqq (I - U^T U)(\mathbf{x}_T - \alpha_T \mu) \qquad c_k(T) \coloneqq \mathbf{u}_k^T(\mathbf{x}_T - \alpha_T \mu)$$

$$\psi(t,\lambda) \coloneqq \sqrt{\frac{\sigma_t^2 + \lambda\alpha_t^2}{\sigma_T^2 + \lambda\alpha_T^2}}$$



State $\mathbf{x}_t$



Validating linear diffusion
Comparing analytical score approximations

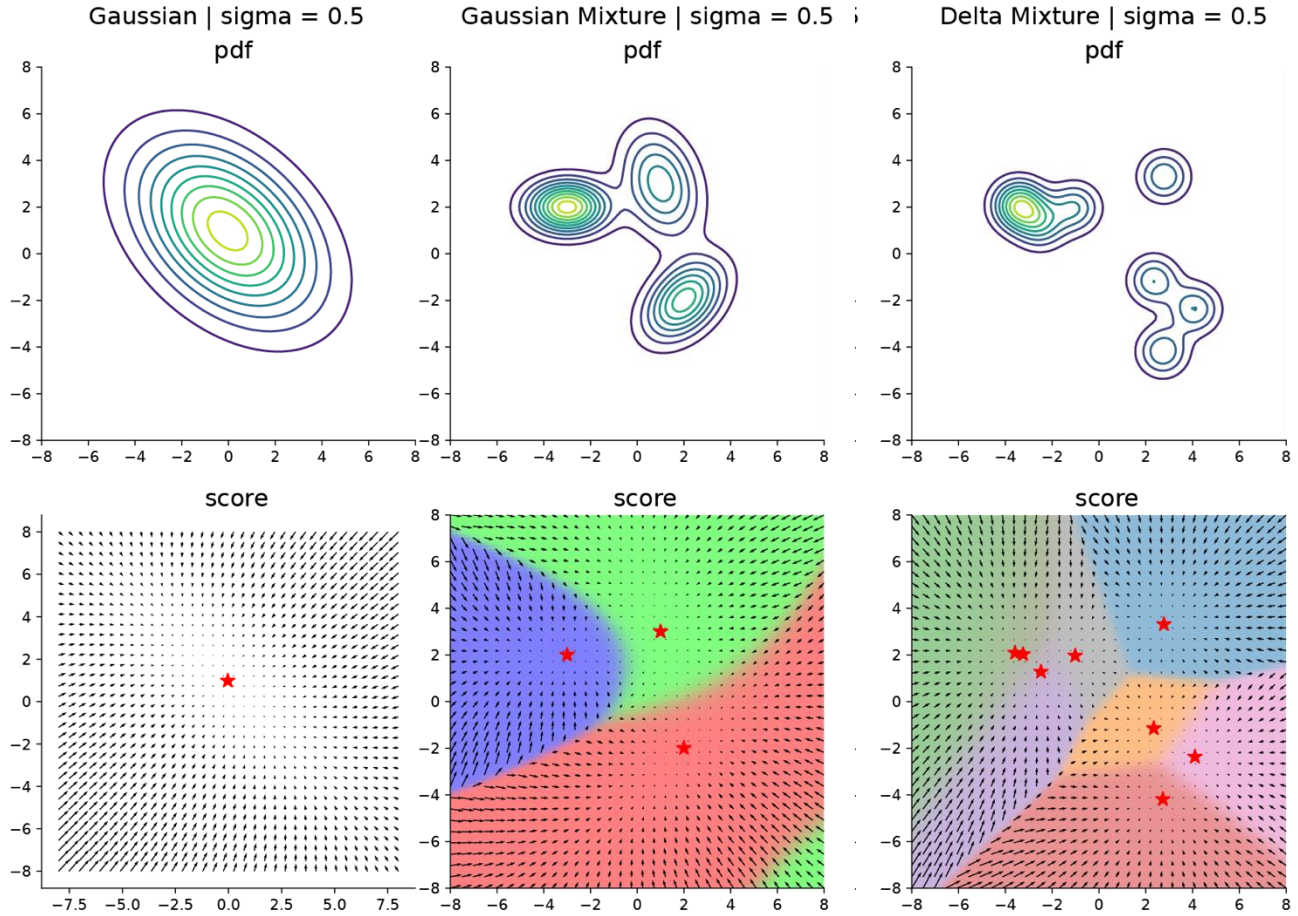
- How well does analytical score explain pretrained DNN scores?

- The analytical score candidates are

  - Gaussian (linear)
    $$(\Sigma + \sigma^2 I)^{-1}(\boldsymbol{\mu} - \mathbf{x})$$

  - Gaussian mixture
    $$\sum_i w_i(\mathbf{x})(\Sigma_\mathbf{i} + \sigma^2 I)^{-1}(\boldsymbol{\mu_i} - \mathbf{x})$$

  - Delta mixture (memorizing)
    $$-\mathbf{x} + \sum_i w_i(\mathbf{x})\, y_i \;\; ;$$
    $$w_\mathbf{i}(\mathbf{x}) = \mathrm{softmax}\left(-\frac{\|y_i - \mathbf{x}\|^2}{\sigma^2}\right)$$



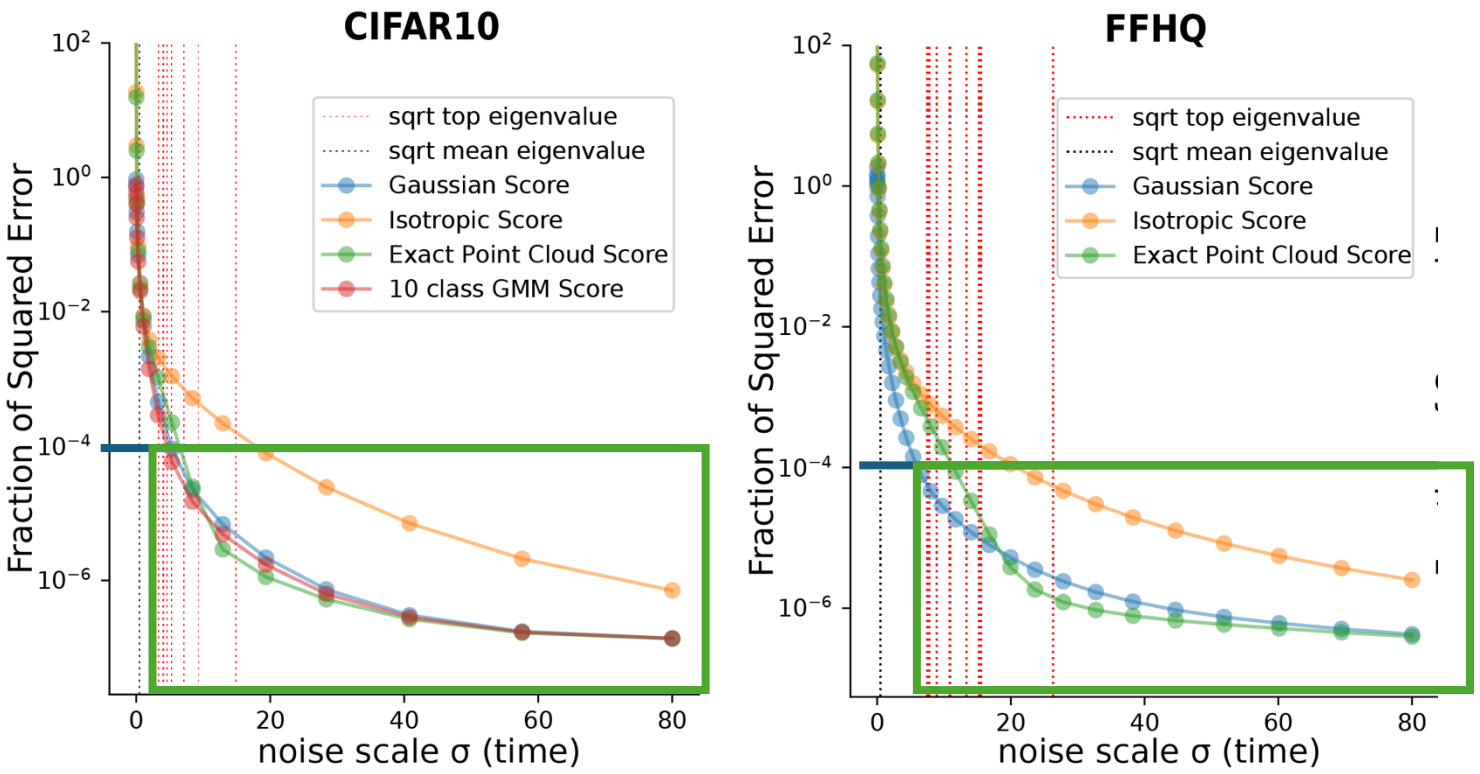
Dataset $\{\mathbf{y_i}\}$,

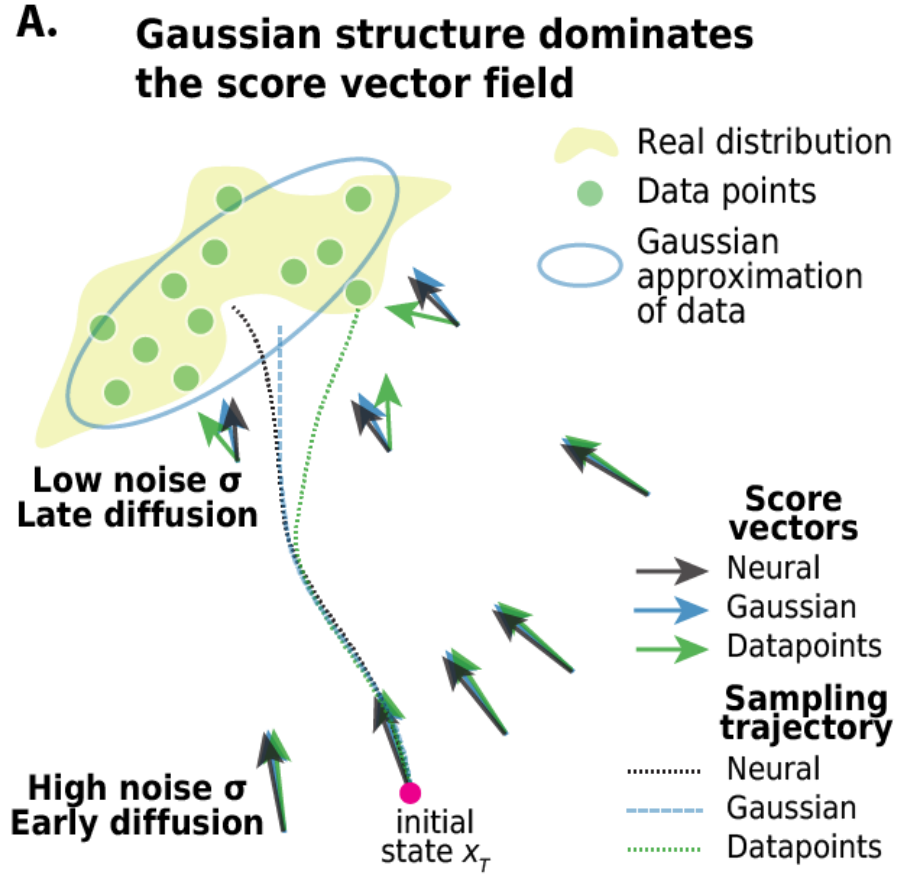
Validation

Gaussian structure dominates the learned neural score function at high noise

Evaluation metric

$$Fraction\ of\ Squared\ Err = \mathbb{E}_{\mathbf{x}}\left[\frac{\left\|\mathbf{s}_{NN}(\mathbf{x};\sigma) - \mathbf{s}_{analy}(\mathbf{x};\sigma)\right\|^2}{\left\|\mathbf{s}_{NN}(\mathbf{x};\sigma)\right\|^2}\right]$$



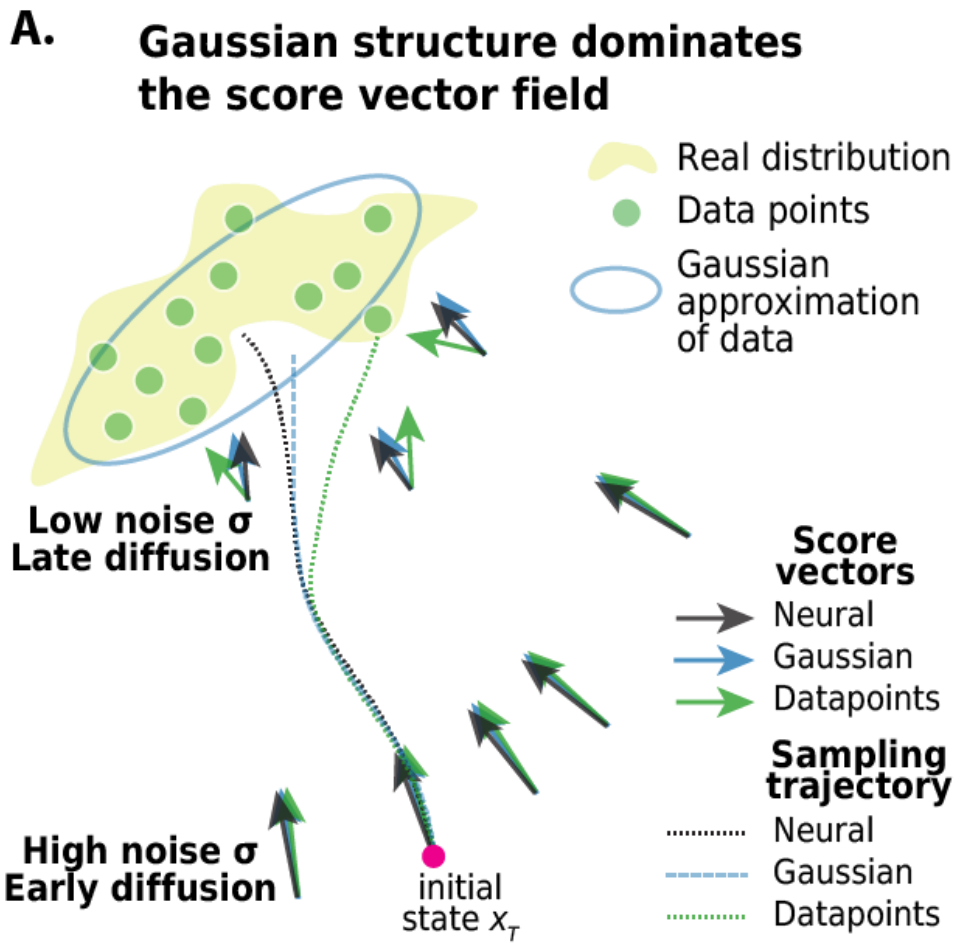


Theory

Gaussian score should be a good approximation at far-field

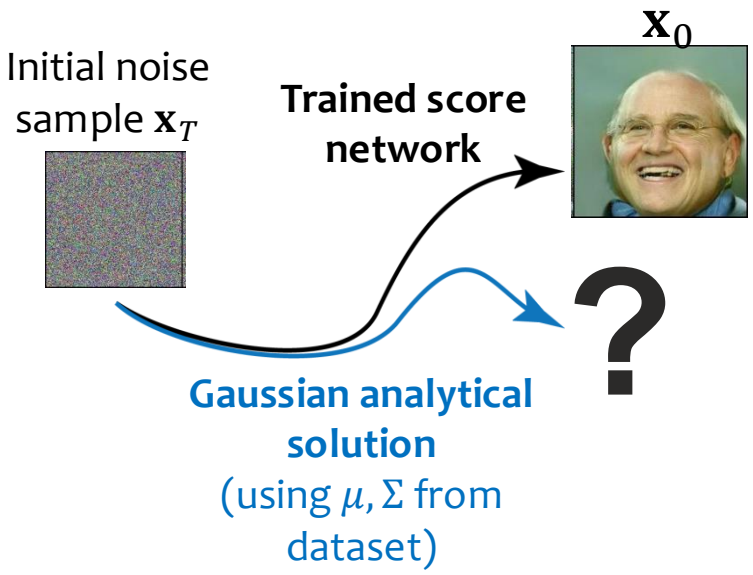
Theorem (*informal*)
For arbitrary *bounded data*, its score function at large enough σ is closely matched to the score of the Gaussian approximation of data.

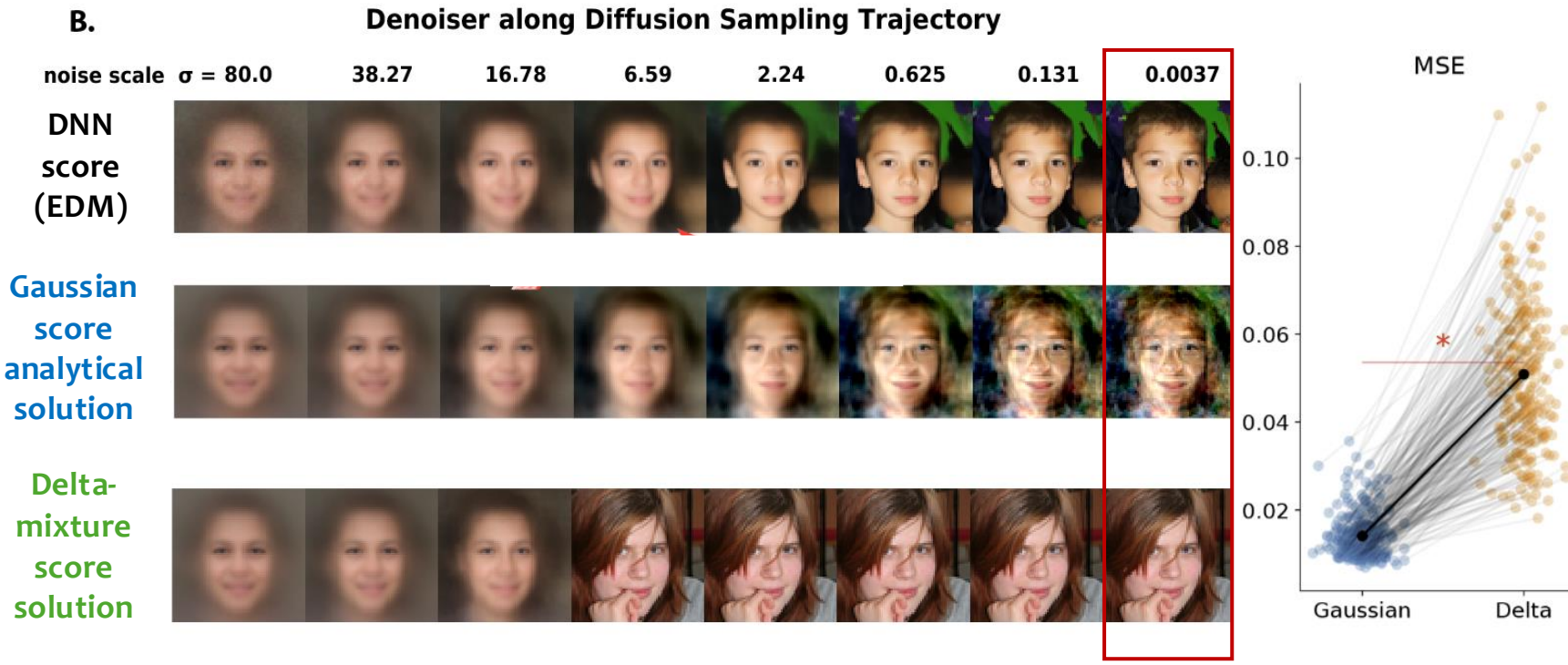
c.f. multi-pole expansion in electromagnetism



Wang, Vastola, TMLR, 2024

Gaussian linear score quantitatively approximates the sampling path and samples of pretrained diffusion



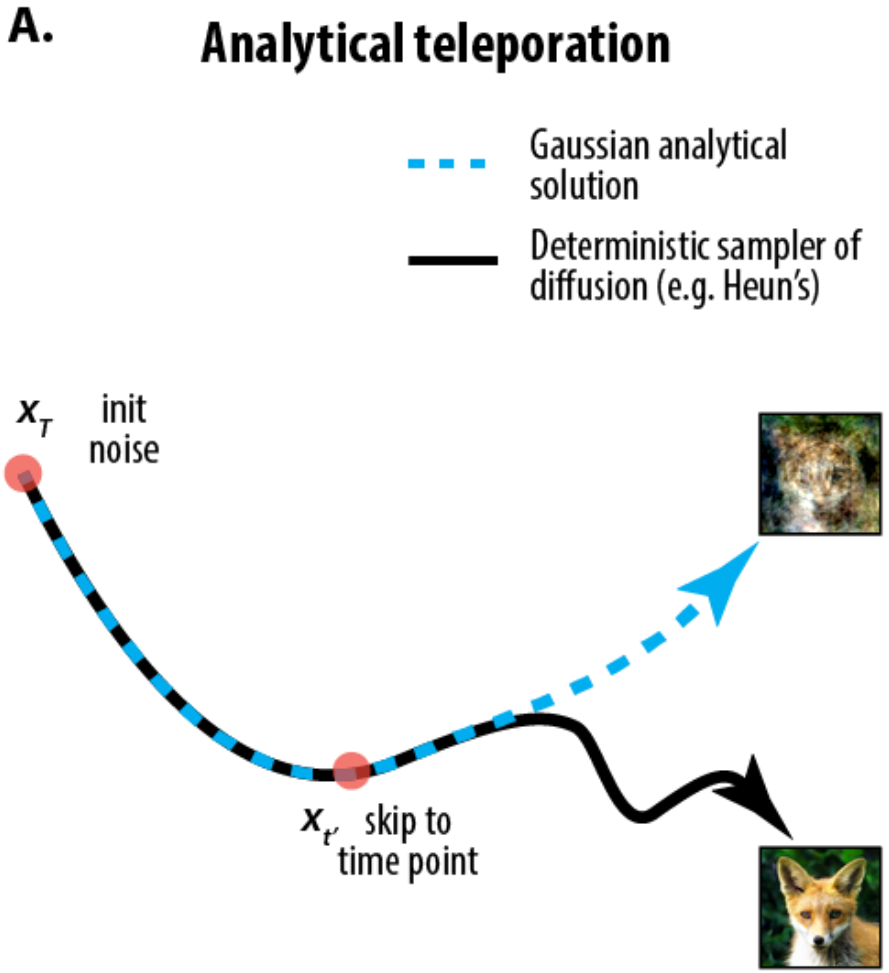


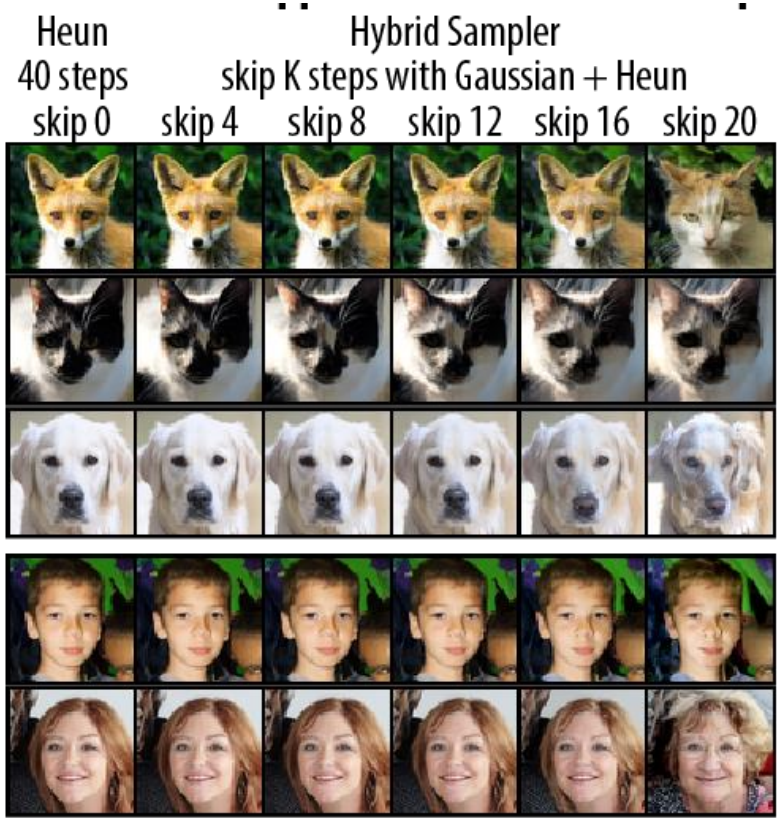
$$\mathbf{D}_{\text{gauss}}(\mathbf{x}_{\sigma_t}; \sigma_t) = \mu + \sum_{k=1}^{r} \xi(t, \lambda_k) \mathbf{u}_k \mathbf{u}_k^T (\mathbf{x}_{\sigma_T} - \mu)$$

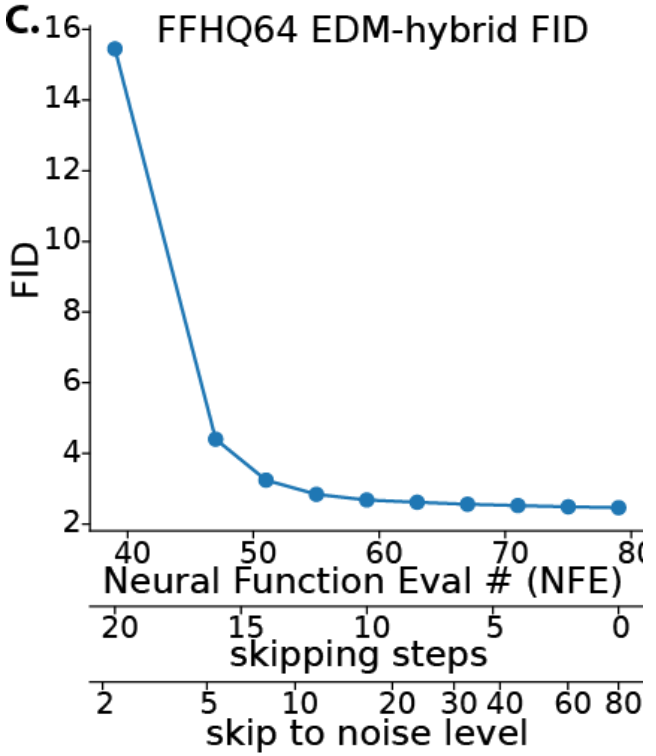
Wang, Vastola, TMLR, 2024

Application

Gaussian solution enables analytical teleportation to accelerate sampling





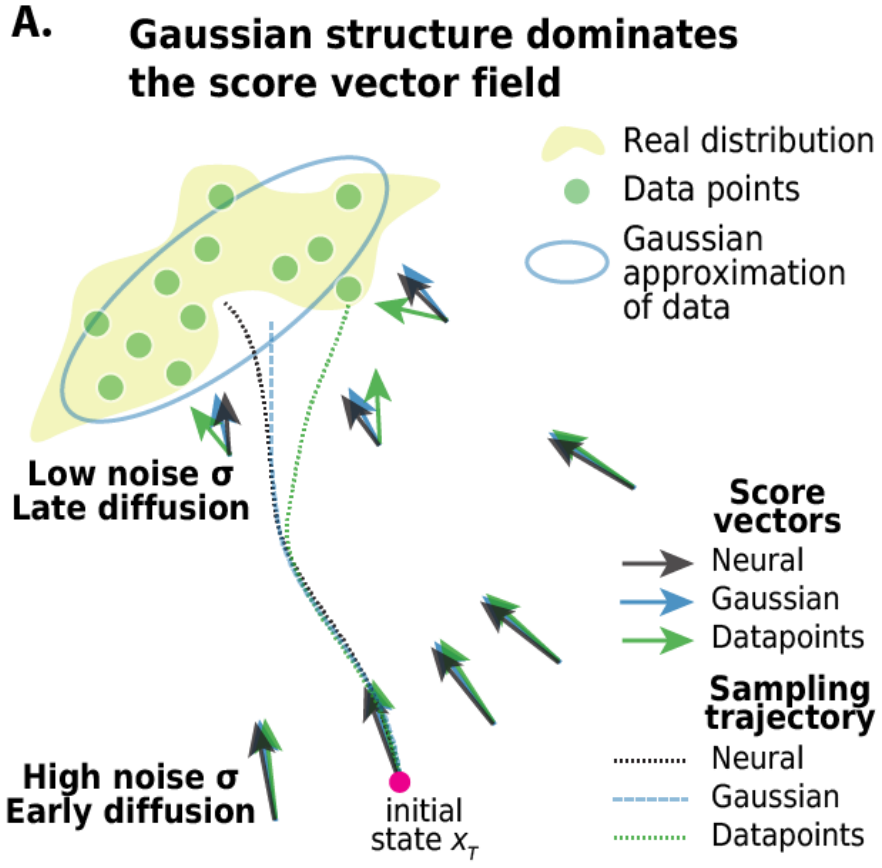


Wang, Vastola, TMLR, 2024

Part I summary

Why linear score is interesting?

- Gaussian score recovers Wiener filter and admits close-form sampling trajectory.
- Score of trained diffusion model is very close to Gaussian linear score for a wide range, esp. at high noise regime.
- Gaussian sampling trajectory predicts the early generation trajectory and final samples, better than empirical delta score solution.



How far could the linear lens get us?

Implications for diffusion phenomena

Sampling dynamics

Receptive Field Structure

Cross-split Consistency

Learning dynamics *

Linear lens implication II

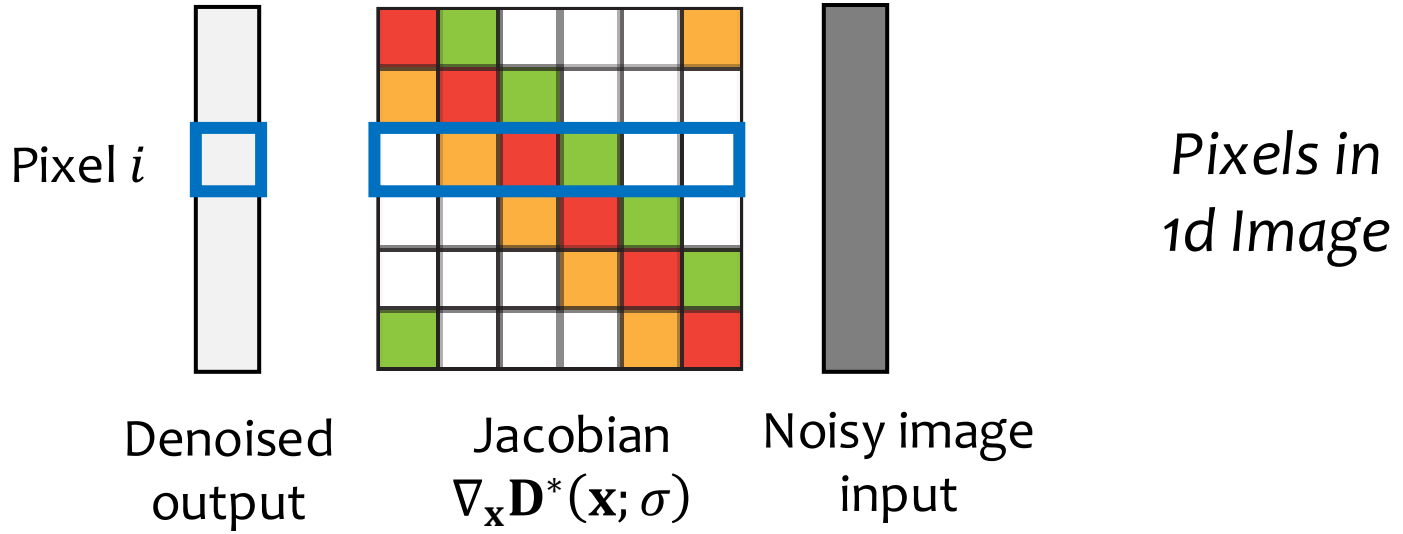
Receptive field depends on noise scale

Jacobian and receptive field of linear denoiser

$$\mathbf{D}^*(\mathbf{x};\sigma) = \boldsymbol{\mu} + \Sigma(\sigma^2 I + \Sigma)^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Jacobian $\nabla_{\mathbf{x}}\mathbf{D}^*(\mathbf{x};\sigma) = \Sigma(\sigma^2 I + \Sigma)^{-1}$

Receptive field. $\mathrm{RF_i} = \mathbf{e}_i^{\mathsf{T}}\,\Sigma(\sigma^2 I + \Sigma)^{-1}$



Theory:
Receptive field in linear denoiser

$$\mathbf{D}^*(\mathbf{x};\sigma) = \boldsymbol{\mu} + \Sigma(\sigma^2 I + \Sigma)^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Receptive field. $\mathrm{RF_i} = \mathbf{e}_i^\top \Sigma(\sigma^2 I + \Sigma)^{-1}$

Dependency on noise scale

Dependency on spatial probe point

Low noise limit
$\boldsymbol{\sigma \to 0}$

$\mathrm{RF_i} = \mathbf{e}_i^\top I = \mathbf{e}_i$

Depending on pixel itself
Local and equivariant!

High noise limit
$\boldsymbol{\sigma \to \infty}$

$\mathrm{RF_i} = \mathbf{e}_i^\top \Sigma/\sigma^2$

Depending on all correlated pixels,
Non-local.
Not-necessarily equivariant

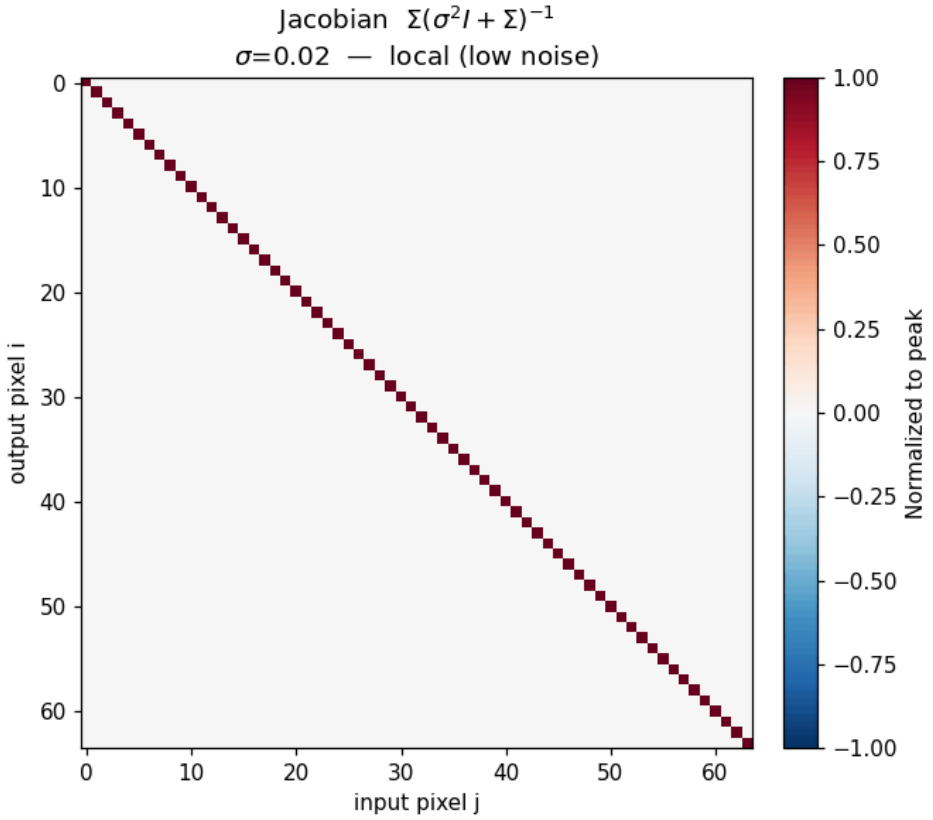
Theory
RF size under natural image statistics

- Idealized assumption
  - 1d “image”
  - Stationary (circulant) covariance matrix.
  - Power law spectrum: eigen values
    $\lambda_0 = \lambda_0, \lambda_m = \lambda_{N-m} = \frac{A}{m^2}$

- We have analytical description of the RF

$$RF(\sigma)[k] = \frac{1}{N}\left[-\frac{\sigma^2}{\sigma^2+\lambda_0} + \frac{\frac{\pi\sqrt{A}}{\sigma}}{\sinh\frac{\pi\sqrt{A}}{\sigma}}\cosh\left(\frac{\pi\sqrt{A}}{\sigma}\left(1-\left|\frac{k}{N/2}\right|\right)\right)\right]$$

$$k = -\frac{N}{2}, \ldots 0, \ldots, \frac{N}{2}$$



Wang, Pehlevan, 2025

Theory
RF size under natural image statistics

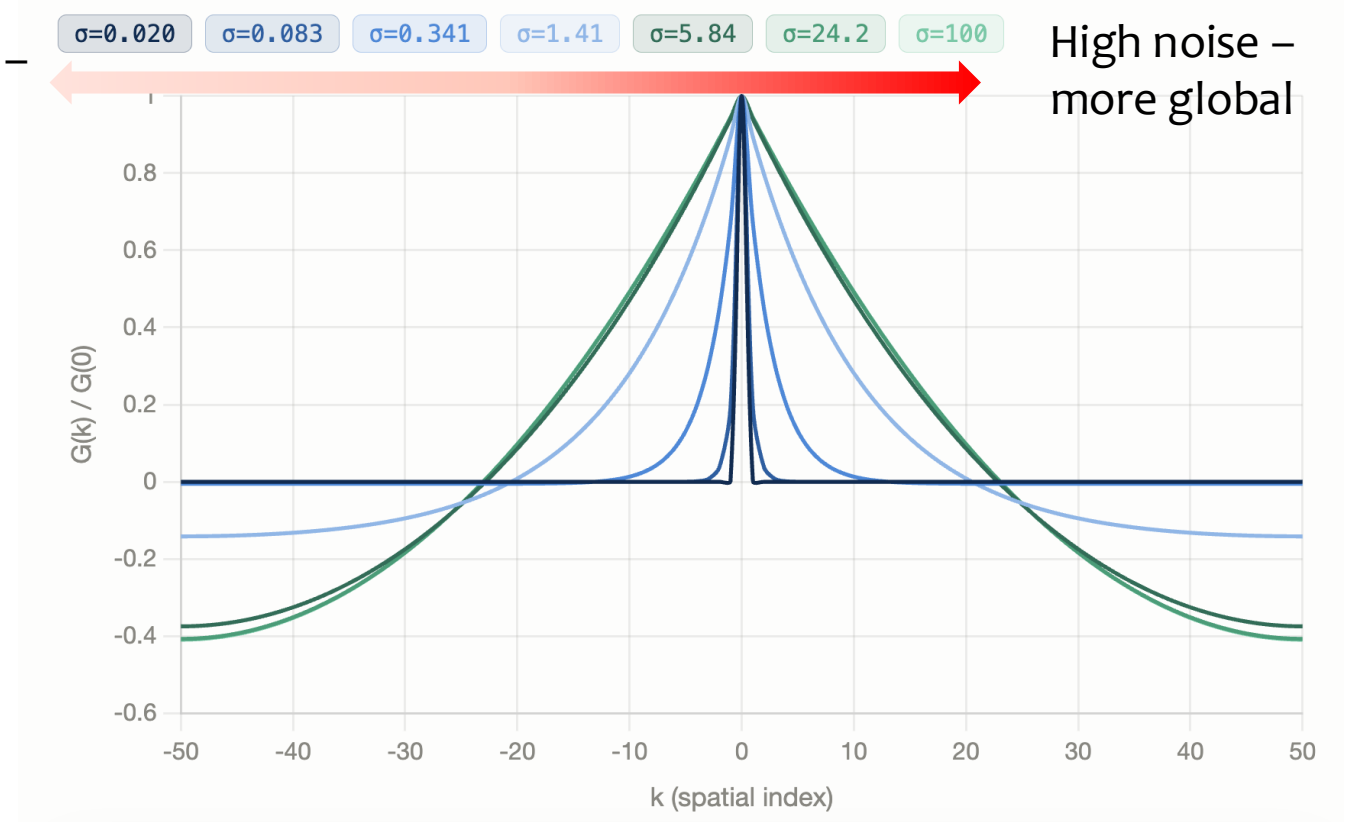
- Idealized assumption
  - 1d “image”
  - Stationary (circulant) covariance matrix.
  - Power law spectrum: eigen values
    $\lambda_0 = \lambda_0, \lambda_m = \lambda_{N-m} = \frac{A}{m^2}$

- We have analytical description of the RF

$$RF(\sigma)[k] = \frac{1}{N}\left[-\frac{\sigma^2}{\sigma^2+\lambda_0} + \frac{\frac{\pi\sqrt{A}}{\sigma}}{\sinh\frac{\pi\sqrt{A}}{\sigma}}\cosh\left(\frac{\pi\sqrt{A}}{\sigma}\left(1-\left|\frac{k}{N/2}\right|\right)\right)\right]$$

$$k = -\frac{N}{2}, \ldots 0, \ldots, \frac{N}{2}$$

Low noise – local



Wang, Pehlevan, 2025

Teaser for the next talk:
Locality from image statistics

Artem Lukoianov

Locality in Image Diffusion Models Emerges from Data Statistics

Artem Lukoianov
Massachusetts Institute of Technology
artem1@mit.edu

Chenyang Yuan
Toyota Research Institute
chenyang.yuan@tri.global

Justin Solomon
Massachusetts Institute of Technology
jsolomon@mit.edu

Vincent Sitzmann
Massachusetts Institute of Technology
sitzmann@mit.edu

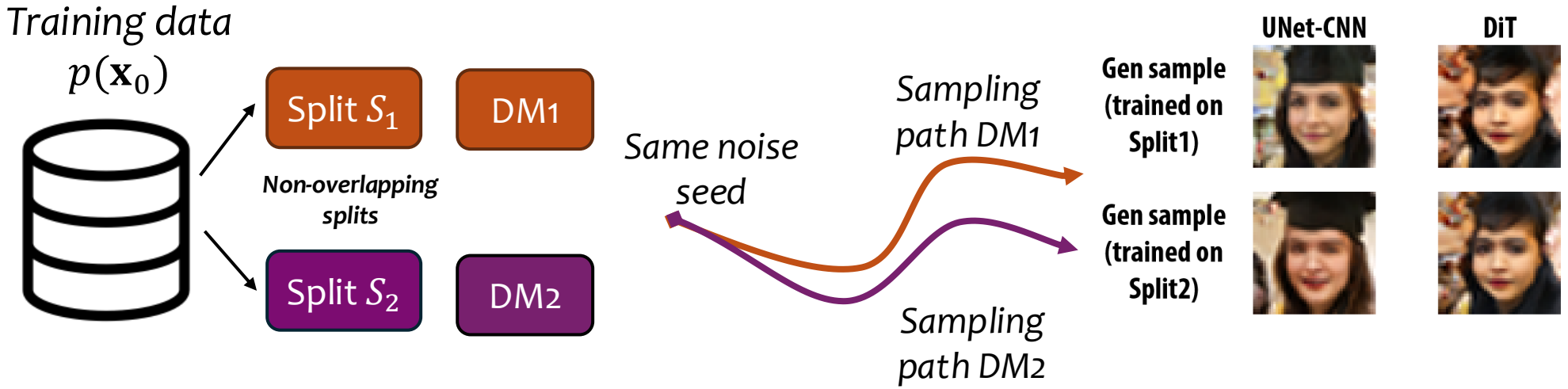
https://locality.lukoianov.com

NeurIPS 2025 spotlight!

Linear lens implication III

Consistency across training splits

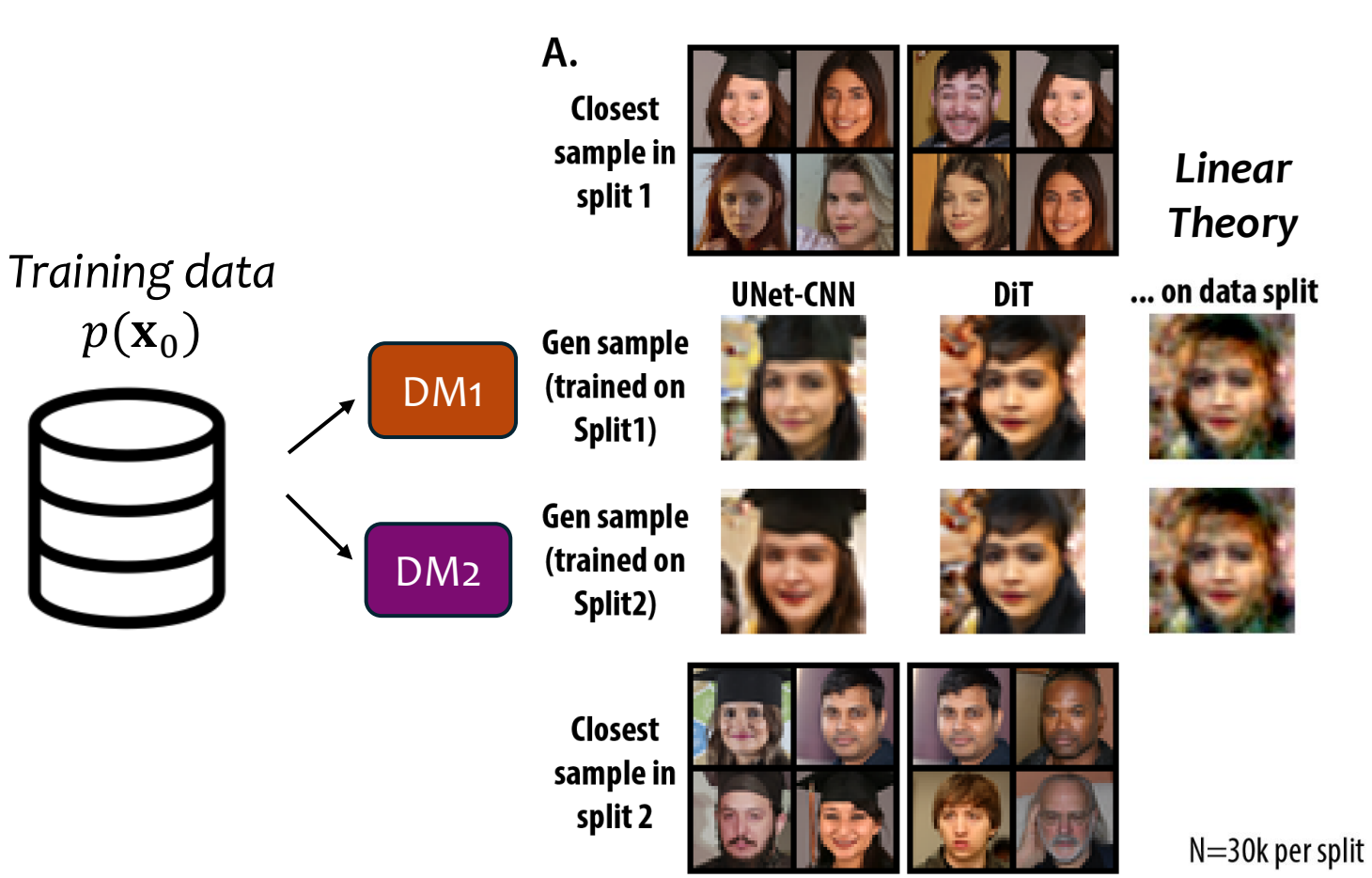
The puzzle of consistency :
Same seed, different training splits, similar images

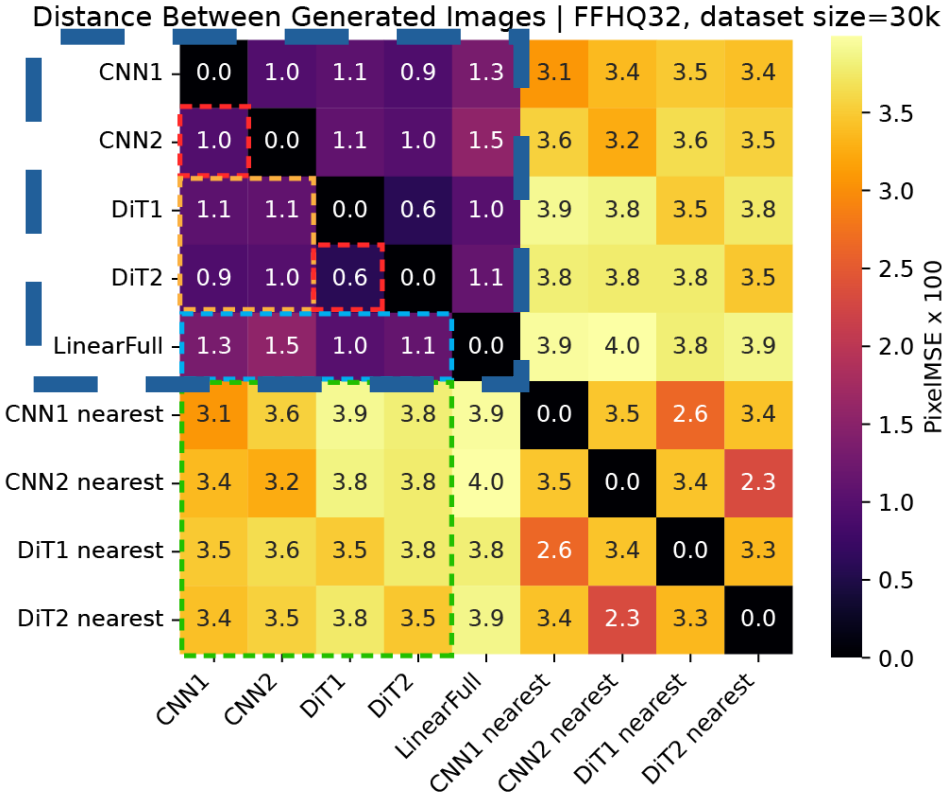


Kadkhodaie, Guth, Simoncelli, Mallat, (2024). *ICLR*
Zhang, Zhou, Lu, Guo, Wang, Shen, Qu, (2024). *ICML*

Linear theory captures consistency across training splits

DNN generation from both splits are close to the linear solution





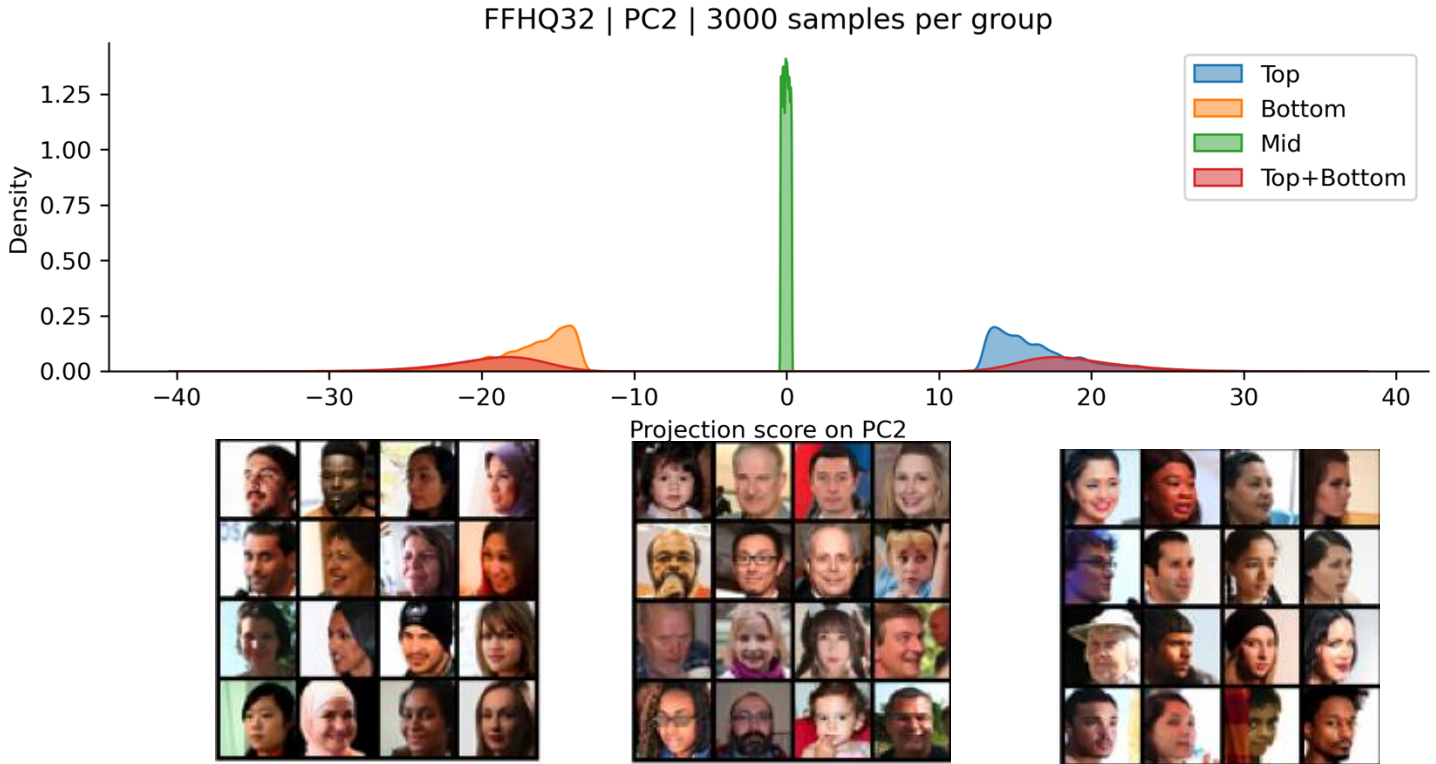
Type of distances

- Consistency across training data split
- Similarity with linear solution
- Consistency across model architecture
- Distance from nearest sample in training set

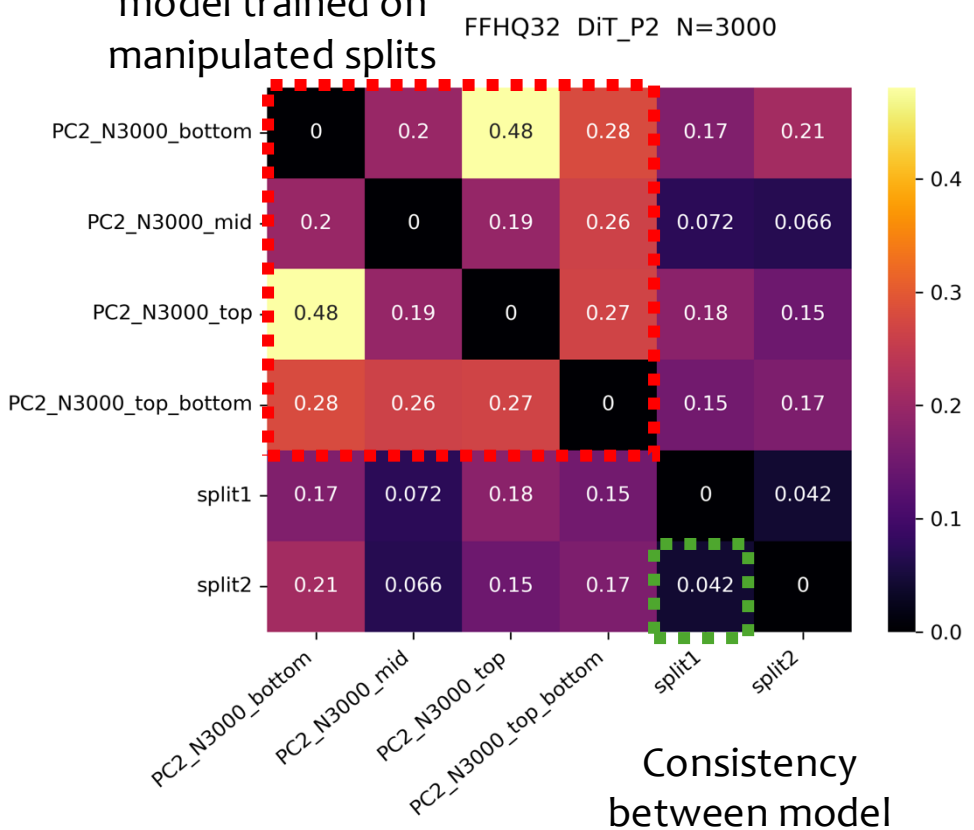
Wang, Zavatone-Veth, Pehlevan, (2026). ICML *oral*
Kadkhodaie, Guth, Simoncelli, Mallat, (2024). *ICLR oral*

When moments differ, consistency breaks

- Manipulate the mean and covariance of splits, by splitting of data along PC.



Consistency between model trained on manipulated splits

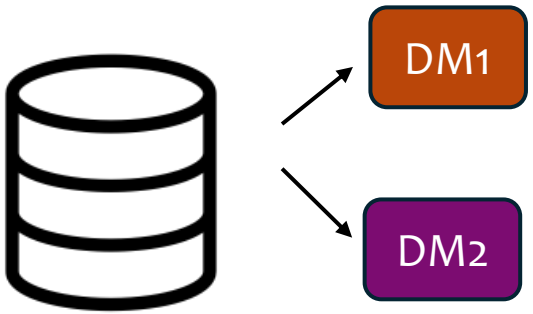


Consistency between model trained on i.i.d. split

Linear intuition of consistency

Relatively stable

Different splits / realization of datasets



Summary statistics
Empirical Gaussian statistics
$\hat{\mu}, \hat{\Sigma}$

Observables
Linear denoiser & Sampling map

$$\mathbf{x}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}_{\sigma_T};\sigma) = \hat{\boldsymbol{\mu}} + \left(\sigma^2 I + \hat{\boldsymbol{\Sigma}}\right)^{\frac{1}{2}}\left(\sigma_T^2 I + \hat{\boldsymbol{\Sigma}}\right)^{-\frac{1}{2}}\left(\mathbf{x}_{\sigma_T} - \hat{\boldsymbol{\mu}}\right)$$

$$\mathbf{D}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x};\sigma) = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\Sigma}}\left(\sigma^2 I + \hat{\boldsymbol{\Sigma}}\right)^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}})$$

Linear / Far field approximation

Neural diffusion model

(High noise & generalization regime)

RMT quantifies the stability of the linear denoiser and sampler

Training distribution $p(\mathbf{x}_0)$

n i.i.d. samples $X = \{\mathbf{x}_i\}$

$X \in \mathbb{R}^{n \times d}$

Empirical mean & covariance

$\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$

Random variables

Denoiser

$\mathbf{D}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}; \sigma)$

Sampling map

$\mathbf{x}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}_{\sigma_T}; \sigma)$

Nonlinear function of random matrix

Variance due to dataset realization

$\mathrm{Var}_{\hat{\Sigma}}\left[\mathbf{D}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}; \sigma)\right]$

$\mathbb{E}_{\hat{\Sigma}}\left[\mathbf{D}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}; \sigma)\right]$

$\mathrm{Var}_{\hat{\Sigma}}\left[\mathbf{x}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}_{\sigma_T}; \sigma)\right]$

$\mathbb{E}_{\hat{\Sigma}}\left[\mathbf{x}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}_{\sigma_T}; \sigma)\right]$

Bias due to finite dataset realization

Random matrix theory tools (deterministic equivalence) can precisely predict the bias and variance due to finite data realization!

Wang, Zavatone-Veth, Pehlevan, (2026). ICML *oral*

Random Matrix Theory to the rescue

- Deterministic equivalence (DE)

$$\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-1} \asymp \Sigma(\Sigma+\kappa(\lambda)I)^{-1}$$

Fluctuation due to X

Isotropic effective regularization
(function of sample size N, spectrum μ, and λ)

DE circumvents expectation!

$$\mathbb{E}_{\hat{\Sigma}}\left[\hat{\Sigma}\left(\hat{\Sigma}+\sigma^2 I\right)^{-1}\right] \asymp \Sigma(\Sigma+\kappa(\sigma^2)I)^{-1}$$

Atanasov, A., Zavatone-Veth, J. A., & Pehlevan, C. (2024). Scaling and renormalization in high-dimensional regression
Bach, F. (2024). High-dimensional analysis of double descent for linear regression with random projections.
Atanasov, A., Bordelon, B., Zavatone-Veth, J. A., Paquette, C., & Pehlevan, C. (2025).

Computing renormalization effect: Silverstein equation

- Deterministic equivalence

$$\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-1} \asymp \Sigma(\Sigma+\kappa(\lambda)I)^{-1}$$

Isotropic regularization

- Self consistent equation (Silverstein)

$$\kappa(\lambda)-\lambda=\gamma\kappa(\lambda)\int_0^\infty \frac{s d\mu(s)}{\kappa(\lambda)+s}=\gamma\kappa(\lambda)\,\mathrm{tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa(\lambda)I)^{-1}]$$

Population covariance Σ
Population limit spectrum μ
Aspect ratio $\gamma = d/n$

} Silverstein Eq. Fixed point

$\lambda \to \kappa$

Silverstein, J. W. (1995).
Marčenko, V. A., & Pastur, L. A. (1967).

Summary of theoretical results

RMT toolkit | **Overshrinkage** | **Anisotropy** | **Inhomogeneity**

1-point DE — **Denoiser** — *Expectation*

$$\mathbb{E}_{\hat{\boldsymbol{\Sigma}}}\left[\mathbf{v}^\top \mathbf{D}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x};\sigma)\right] \asymp \mathbf{v}^\top \mathbf{D}^*_{\boldsymbol{\Sigma}}(\mathbf{x};\kappa(\sigma^2)) = \mathbf{v}^\top\left[\mu + \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa(\sigma^2)I)^{-1}(\mathbf{x}-\mu)\right]$$

2-point DE — **Denoiser** — *Fluctuation*

$$\begin{aligned}\mathbf{v}^\top \mathcal{S}_D(\mathbf{x})\mathbf{v} &= \mathrm{Var}_{\hat{\boldsymbol{\Sigma}}}[\mathbf{v}^\top \mathbf{D}^*_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x};\sigma)] \\ &\asymp \frac{\kappa(\sigma^2)^2}{n-\mathrm{df}_2(\kappa(\sigma^2))} \underbrace{\left(\mathbf{v}^\top\left(\boldsymbol{\Sigma}+\kappa(\sigma^2)I\right)^{-2}\boldsymbol{\Sigma}\mathbf{v}\right)}_{\text{anisotropy: } \square(\mathbf{v},\kappa,\boldsymbol{\Sigma})} \underbrace{\left((\mathbf{x}-\mu)^\top\left(\boldsymbol{\Sigma}+\kappa(\sigma^2)I\right)^{-2}\boldsymbol{\Sigma}(\mathbf{x}-\mu)\right)}_{\text{inhomogeneity: } \square(\mathbf{x}-\mu,\kappa,\boldsymbol{\Sigma})}\end{aligned} \tag{6}$$

Fractional DE — **Generation map** — *Expectation*

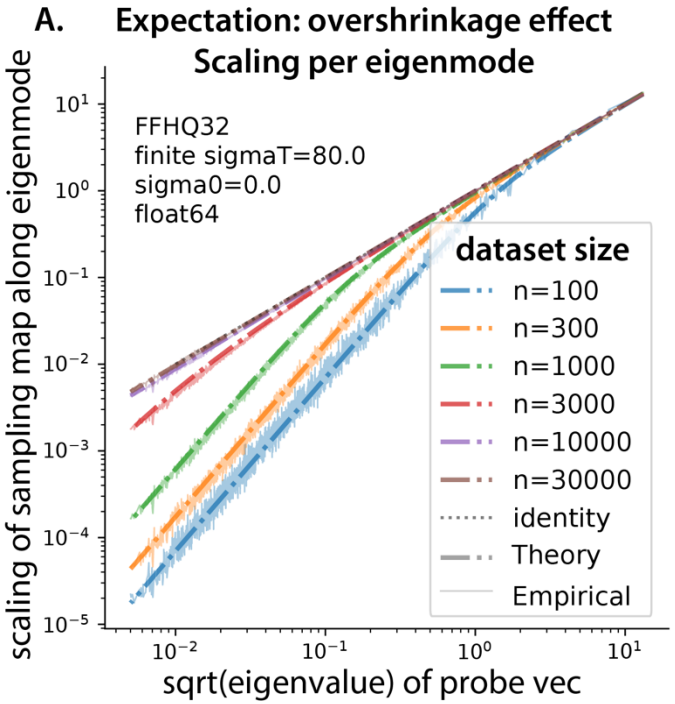
$$\mathbb{E}_{\hat{\boldsymbol{\Sigma}}}[\mathbf{x}_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}_{\sigma_T},0)] \approx \mu + \mathbb{E}_{\hat{\boldsymbol{\Sigma}}}\left[\hat{\boldsymbol{\Sigma}}^{1/2}\right]\frac{\mathbf{x}_{\sigma_T}-\mu}{\sigma_T} \asymp \mu + \frac{2}{\pi}\int_0^\infty \boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\kappa(u^2)I\right)^{-1}\bar{\mathbf{x}}\,du$$

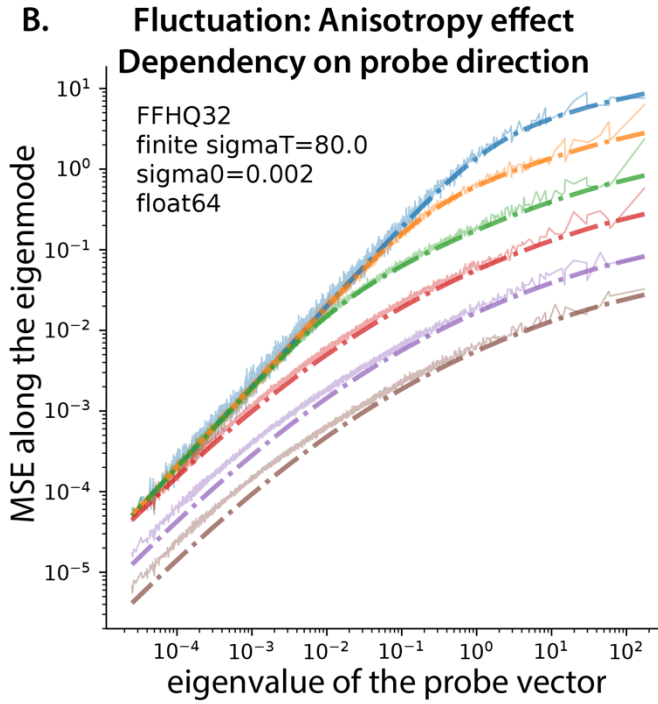
Fractional DE — **Generation map** — *Fluctuation*

$$\begin{aligned}\mathrm{Var}_{\hat{\boldsymbol{\Sigma}}}[\mathbf{v}^\top \mathbf{x}_{\hat{\boldsymbol{\Sigma}}}(\mathbf{x}_{\sigma_T},0)] &= \mathrm{Var}_{\hat{\boldsymbol{\Sigma}}}[\mathbf{v}^\top \hat{\boldsymbol{\Sigma}}^{1/2}\bar{\mathbf{x}}] \\ &\asymp \frac{4}{\pi^2}\int_0^\infty\int_0^\infty \frac{\kappa\,\kappa'}{n-\mathrm{df}_2(\kappa,\kappa')} \underbrace{\pentagon(\mathbf{v};\kappa,\kappa',\boldsymbol{\Sigma})}_{\text{anisotropy}} \underbrace{\pentagon(\bar{\mathbf{x}};\kappa,\kappa',\boldsymbol{\Sigma})}_{\text{inhomogeneity}} du\,dv,\end{aligned}$$

RMT Theory Prediction: detailed structure of consistency

- Bias - Overshrinkage
  - Over-shrink lower variance dimension towards the mean.
- Fluctuation - Anisotropy:
  - More disagreements along top PCs.
- Fluctuation - Inhomogeneity:
  - More disagreement when initial noise is along top PCs.



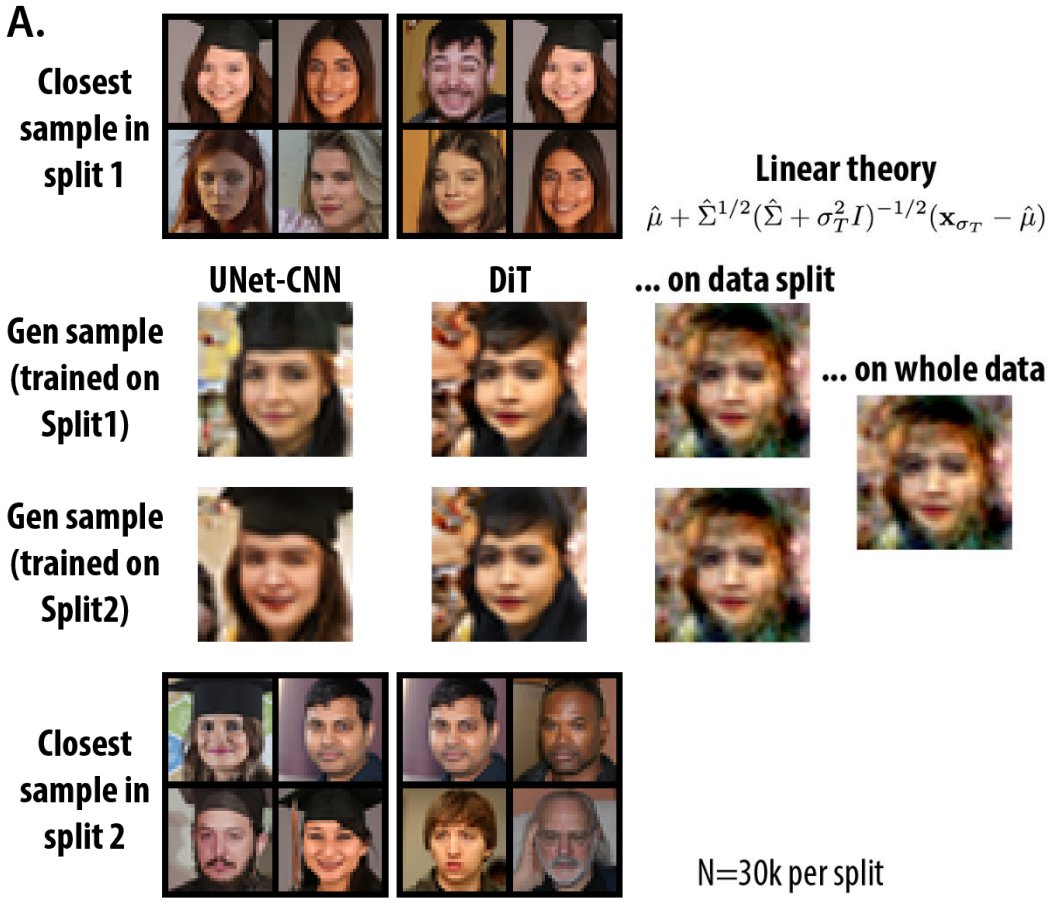


Wang, Zavatone-Veth, Pehlevan, (2026). ICML *oral*

Part III summary
Why samples are consistent and reproducible?

- Consistency across data split is related to the shared Gaussian statistics, which is quite stable.

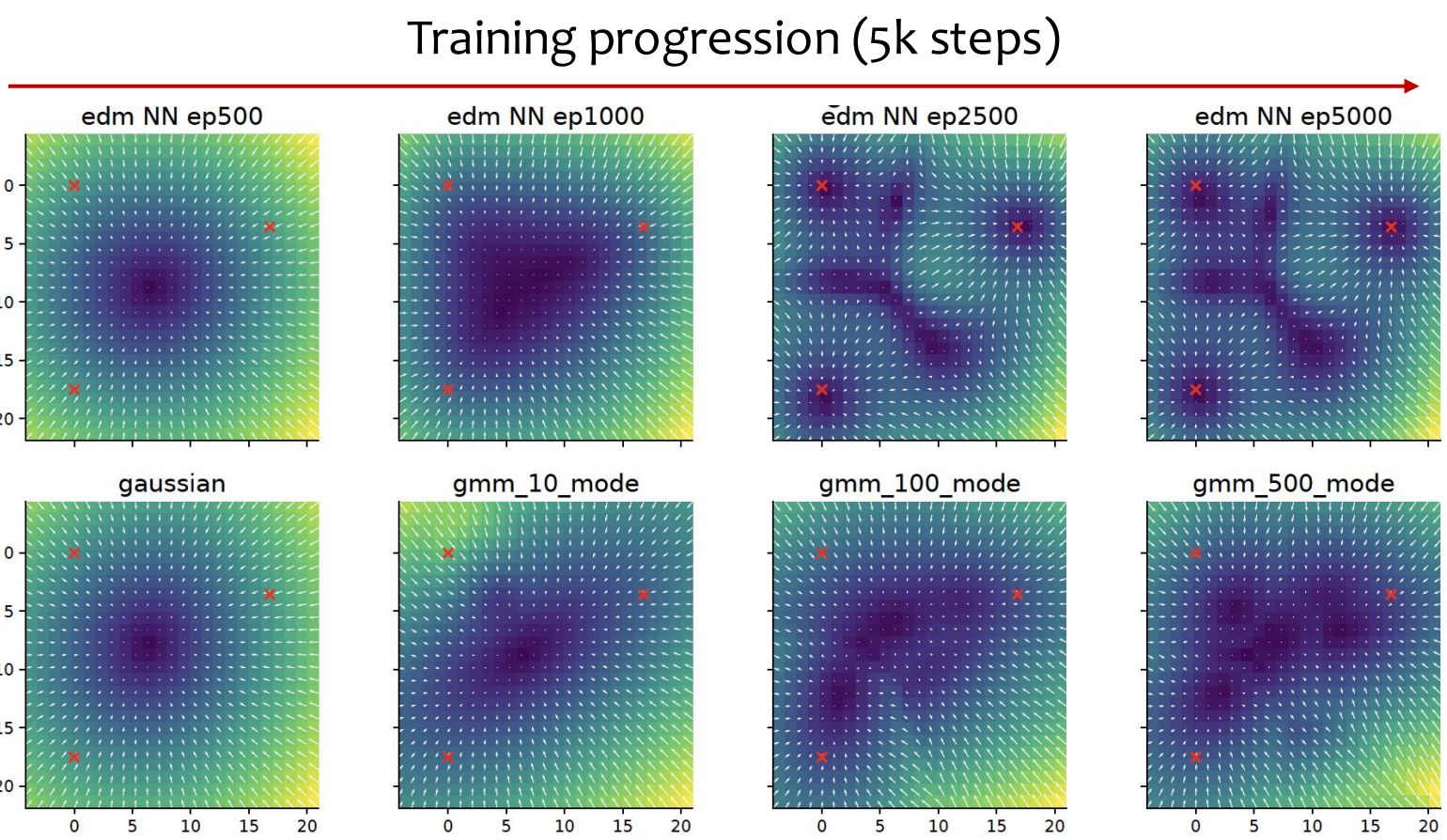
- Variance of generation map reveals detailed structure of consistency: inhomogeneous and anisotropic. Some noise seeds and directions are more inconsistent than others.

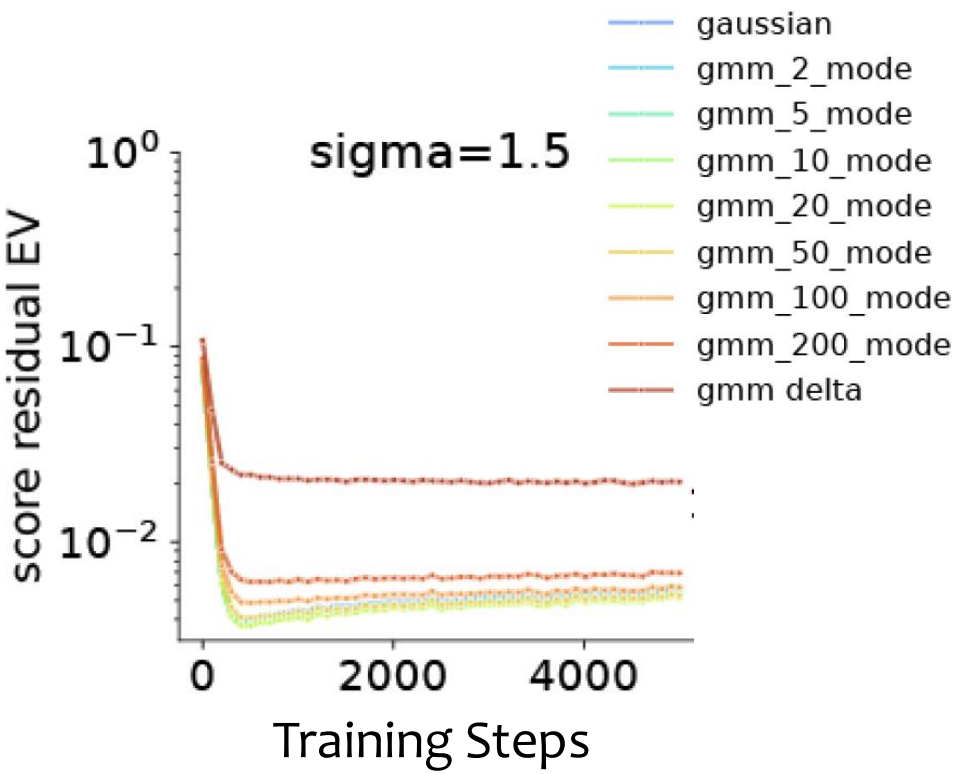


Linear lens implication IV *
Learning dynamics

Empirical clue

diffusion first learns a simpler Gaussian-like score





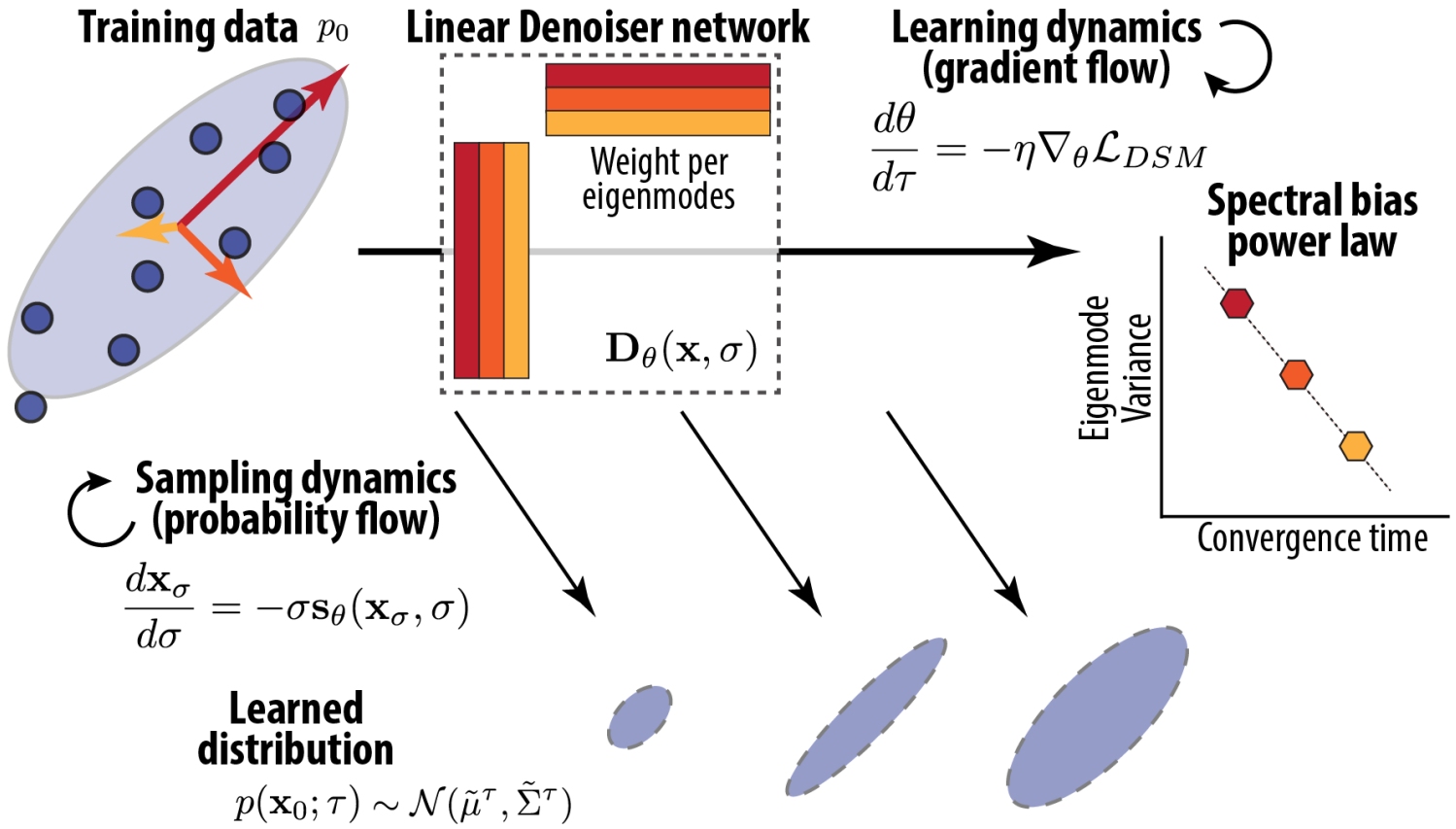
Wang, Vastola, TMLR, 2024

EDM, UNet, learning MNIST

Linear diffusion allows us to solve nested learning and sampling dynamics

$$\mathbf{D}_\theta(\mathbf{x}, \sigma) = W_\sigma \mathbf{x} + b_\sigma$$

- Linear denoiser set up allows us to solve the coupled learning and sampling dynamics.
- Tracking the generated distribution during training.



Wang, Pehlevan, (2025), *NeurIPS spotlight*

Theory outline

Dataset statistics μ, Σ

Objective * $\mathcal{L}_{DSM}$

Network architecture ** $\mathbf{D}_\theta(\mathbf{x}, \sigma) = \mathbf{W}_\sigma \mathbf{x} + \mathbf{b}_\sigma$

$\frac{d\theta}{d\tau} = -\eta \nabla \mathcal{L}_{DSM}$

Gradient flow ODE

Dynamics of denoiser weights $\mathbf{W}_\sigma$

$\mathbf{W}_\sigma(\tau) = \sum_k g(\lambda_k, \sigma, \tau) \mathbf{u}_k \mathbf{u}_k^\top$

$\frac{d\mathbf{x}}{d\sigma} = -\sigma \boldsymbol{s}_\theta(\mathbf{x}, \sigma)$

Probability flow ODE

Dynamics of generated samples covariance $\tilde{\Sigma}$

$\tilde{\Sigma}(\tau) = \sum_k \Phi(\lambda_k, \tau) \mathbf{u}_k \mathbf{u}_k^\top$

Wang, Pehlevan, (2025), *NeurIPS spotlight*

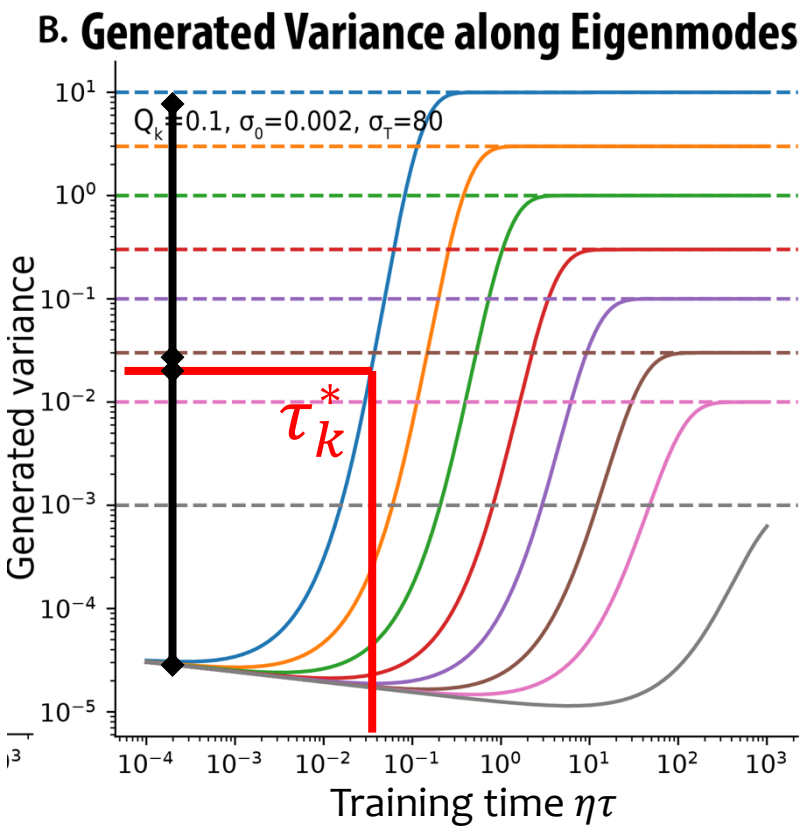
Linear theory prediction
Spectral bias in distribution learning

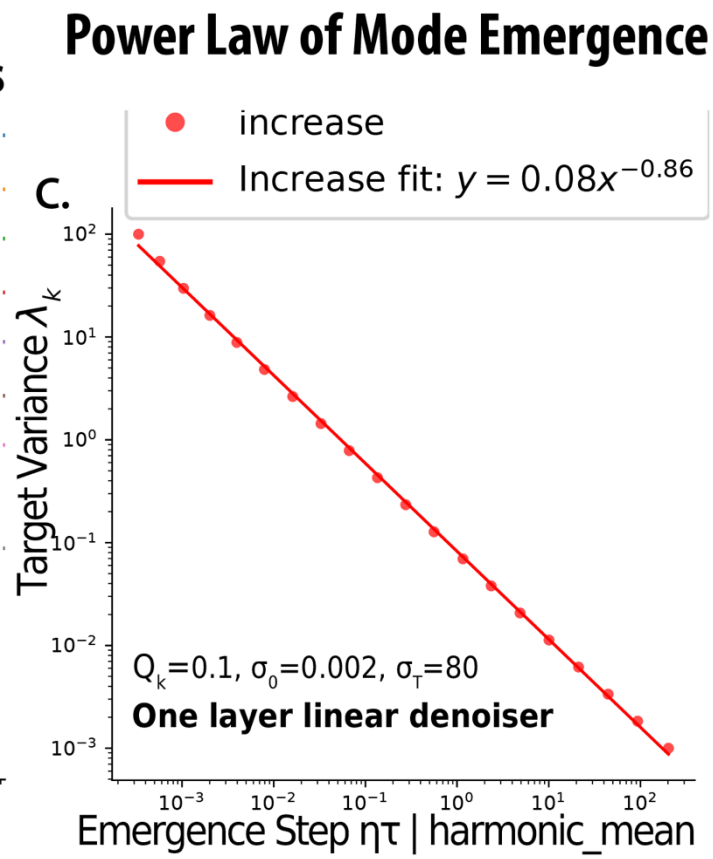
- Emergence time τ_k^* follows inverse power law with the mode variance λ_k

$$\tau_k^* \propto \lambda_k^{-\alpha}$$

- **Implication:** Eigenmodes with 1/10 variance will take 10 times training time to converge!

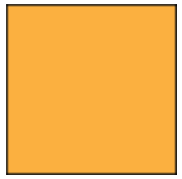
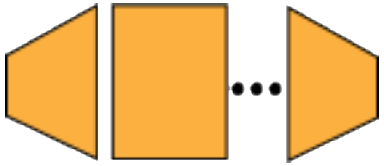
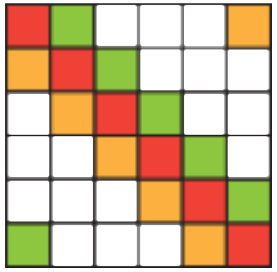
$\Phi(\lambda_k, \tau)$





Wang, Pehlevan, (2025), *NeurIPS spotlight*

Solvable linear architectures

	$\mathbf{D}(\mathbf{x}, \sigma) =$	Structure of $\mathbf{W}_\sigma$	**Weight convergence dynamics**	**Evolution of learned distribution**
Simple linear network	$W\mathbf{x} + b$		Exponential, along data PCs	Inverse var. spectral bias along PCs
Deep linear network	$W_l \ldots W_1 \mathbf{x}$		Sigmoidal, along data PCs	Inverse var. spectral bias along PCs
Linear conv network	$w_\sigma * \mathbf{x}$		Exponential, along Fourier modes	Inverse var. spectral bias along Fourier modes
Local patch conv network	$w_\sigma * \mathbf{x}$		Exponential, along patch covariance PCs	**N.A. , Simultaneous emergence of Fourier modes…**

Wang, Pehlevan , 2025, *NeurIPS spotlight*

Empirical Validation

Approach

- Compute mean μ and covariance Σ of training dataset $\mathcal{D} = \{\mathbf{x}_i\}$.

- Train (deep) diffusion model and generate samples throughout training time τ, $\{\mathbf{x}_j^\tau\}$.

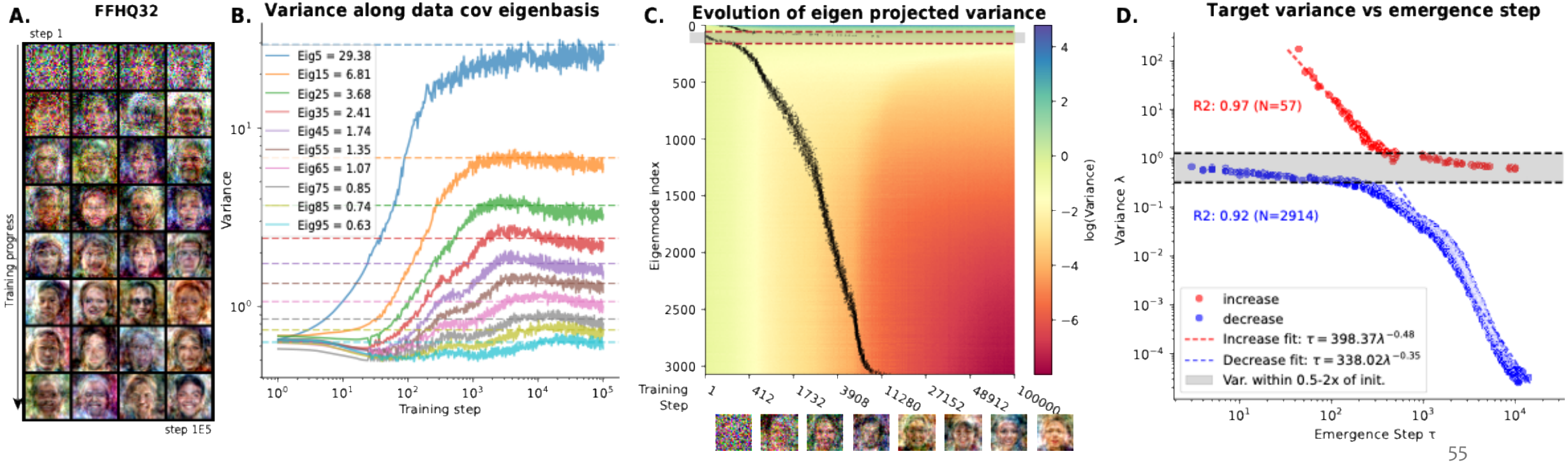
- Track the empirical variance of generated samples along PCs $\tilde{\lambda}_k^\tau := \mathbf{u}_k^T \tilde{\Sigma}^\tau \mathbf{u}_k$.

Empirical Validation

MLP-based diffusion shows power law spectral bias

- Model $\mathbf{D}_\theta$: MLP
- Dataset $\mathcal{D}$: FFHQ32

$$\tau_k^* \propto \lambda_k^{-\alpha}$$
$$\alpha \sim 0.4$$

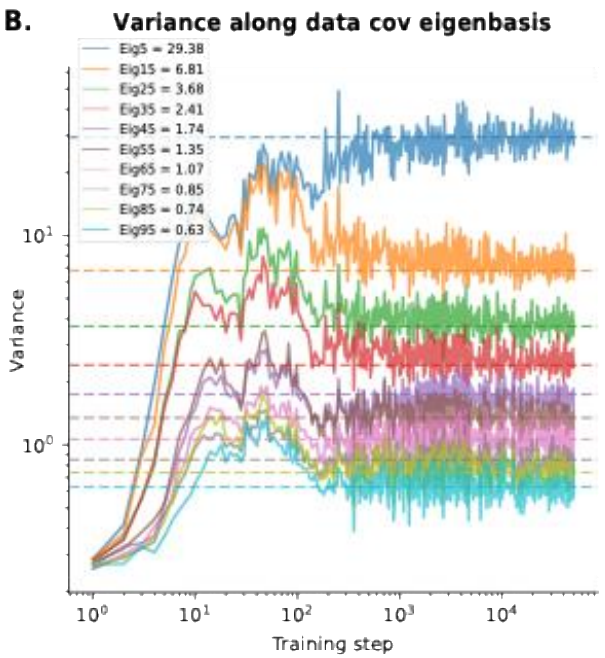


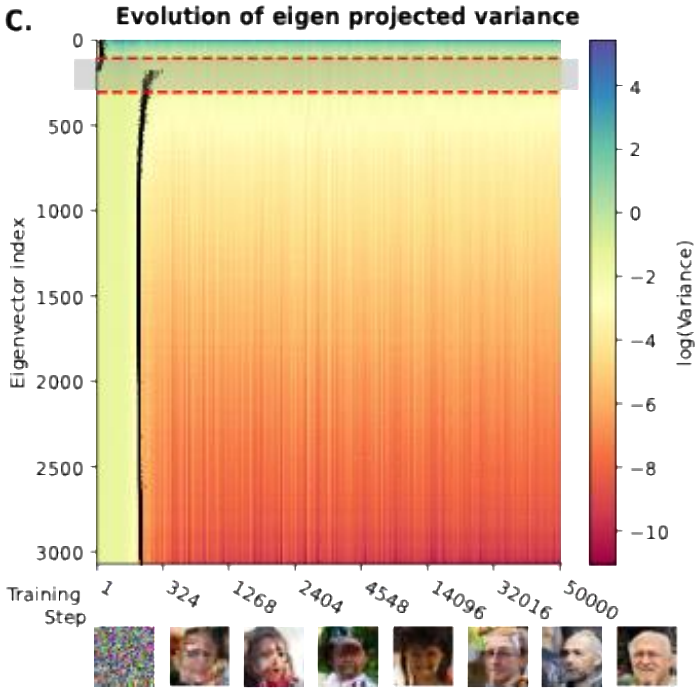
Where the linear lens reaches its limit: UNet architecture circumvent spectral bias and learn many modes at once

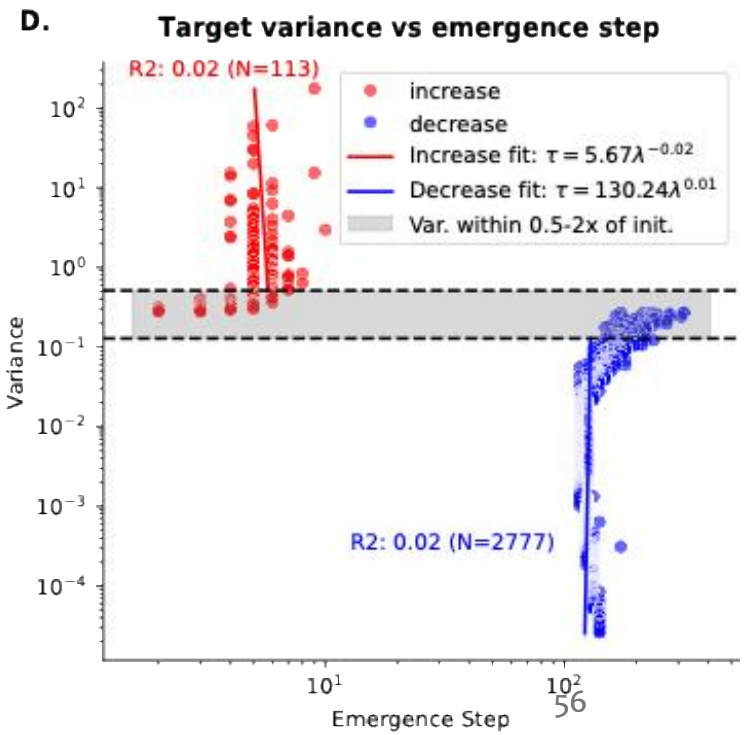
- Model $\mathbf{D}_\theta$: UNet
- Dataset $\mathcal{D}$: FFHQ32

$$\tau_k^* \propto \lambda_k^{-\alpha}$$
$$\alpha \sim 0$$

No spectral bias!



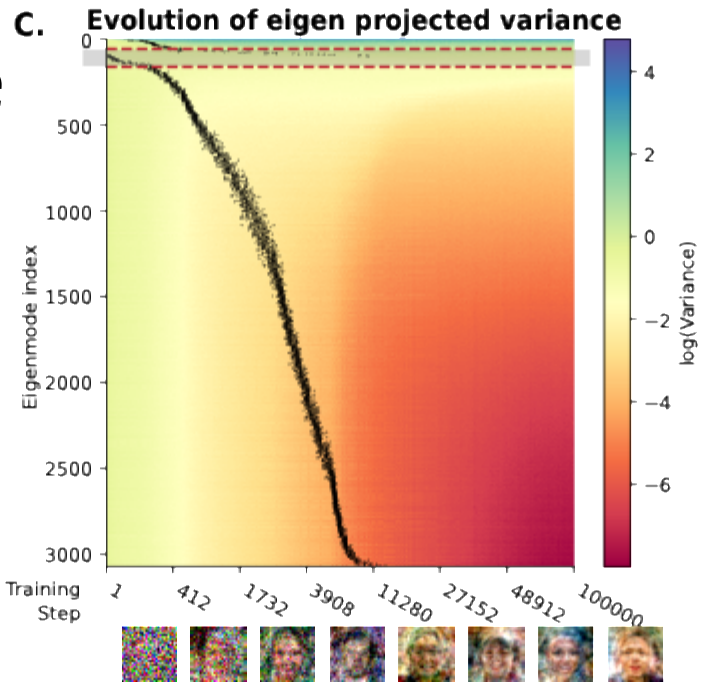


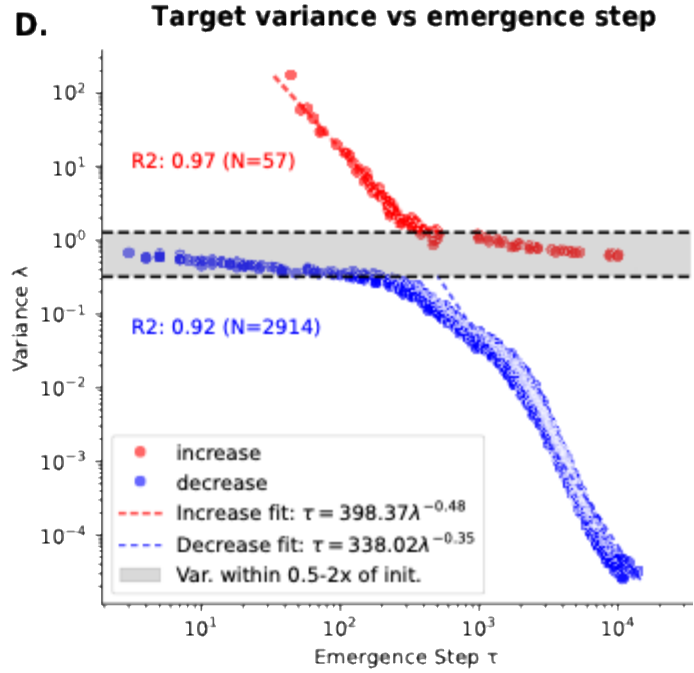


Part IV Summary
What is the order of learning?

- For MLPs, learned distribution converges to the target from higher variance to lower variance modes, exhibit spectral bias predicted by linear theory.

- Model architecture affects the spectral bias, with local convolution attenuates or annihilates it.





What can linear score explain well?

Gaussian statistics of dataset μ, Σ

Linear score / denoiser

$$\mathbf{s}_{gauss}(\mathbf{x};\sigma) = (\Sigma + \sigma^2 I)^{-1}(\mu - \mathbf{x})$$

$$\mathbf{D}_{gauss}(\mathbf{x};\sigma) = \mu + \Sigma(\Sigma + \sigma^2 I)^{-1}(\mathbf{x} - \mu)$$

Far field approximation / Inductive bias

Deep neural networks

Sampling	Gradient Structure (Receptive Field)	Consistency	Learning
• Spectral order of feature generation. • Early sampling trajectory and score. • Predicting rough layout of samples.	• Noise scale and position dependent of receptive field and its link to data covariance.	• Gaussian statistics of data relates to consistency of denoiser and sampling map.	• Spectral bias of distribution learning in fully connected networks (MLPs).

Where does linear theory fall short?

Sampling dynamics

- Score, denoiser and sampling trajectory of middle to low noise scale.
- Higher frequency details of the generated samples.

Jacobian Structure (Receptive Field)

Consistency

- Linear theory is overly consistent, not absorbing enough heterogeneity from higher order stats.

Learning Dynamics

- Architecture dependent learning dynamics, esp. that for UNet, DiT.
- Learning beyond the Gaussian stats.
- Memorization dynamics

Reference

TMLR24, arxiv: 2412.09726

The Unreasonable Effectiveness of Gaussian Score Approximation for Diffusion Models and its Applications

Binxu Wang *binxu_wang@hms.harvard.edu*
Kempner Institute for the Study of Natural and Artificial Intelligence
Harvard University
Boston, MA 02134, USA

John J. Vastola *john_vastola@hms.harvard.edu*
Department of Neurobiology
Harvard Medical School
Boston, MA 02115, USA

NeurIPS25 *spotlight*, arxiv: 2503.03206

An Analytical Theory of Spectral Bias in the Learning Dynamics of Diffusion Models

Binxu Wang
Kempner Institute, Harvard University
Boston, MA, USA
binxu_wang@hms.harvard.edu

Cengiz Pehlevan
SEAS, Harvard University
Cambridge, MA, USA
cpehlevan@seas.harvard.edu

ICML26 *oral*, arxiv: 2602.02908

A Random Matrix Theory Perspective on the Consistency of Diffusion Mod

Binxu Wang [1] **Jacob Zavatone-Veth** [2 3] **Cengiz Pehlevan** [1 3 4]

Co-authors

John Vastola

Cengiz Pehlevan

Jacob Zavatone-Veth

Wang, Vastola, (2024) *TMLR*
Wang, Pehlevan, (2025), *NeurIPS spotlight*
Wang, Zavatone-Veth, Pehlevan, (2026), *ICML oral*

References

Karras, Aittala, Aila, Laine, (2022). *NeurIPS*
Li, Dai, Qu, (2024). *NeurIPS*
Pierret, Galerne, 2024
Kamb, Ganguli, (2025). *ICML*
Niedoba, Zwartsenberg, Murphy, Wood, (2025). *ICML spotlight*
Lukoianov, Yuan, Solomon, Sitzmann, (2025). *NeurIPS spotlight*
Kadkhodaie, Guth, Simoncelli, Mallat, (2024). *ICLR oral*
Zhang, Zhou, Lu, Guo, Wang, Shen, Qu, (2024). *ICML*

Tutorial notebook: Gaussian Linear Score and Analytical Teleportation

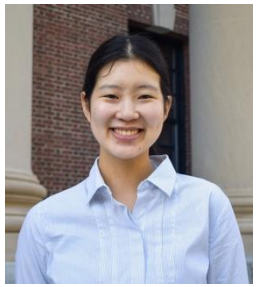
Notebook link

Acknowledgment

Mentee

Emma Finn

Hannah Kim



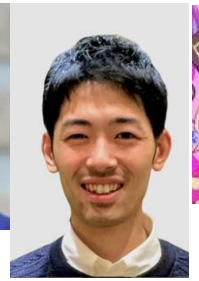
Collaborators

Haim Sompolinsky

Michael Albergo

Martin Wattenberg

Hidenori Tanaka



Yongyi Yang

Ekdeep Singh Lubana

Andrew Lee

Talia Konkle

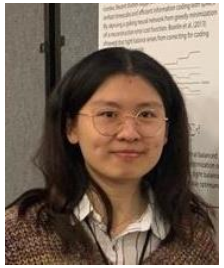
George Alvarez

Shane Shang

Xu Pan

Jingxuan Fan

Qianyi Li



Ann Huang

Bingbin Liu

Naomi Saphra

Ilenna Jones

Thomas Fel

Andy Keller