Title: Architectural bias in a transport-based generative model : an asymptotic perspective

Speakers: Hugo Cui

Collection/Series: Theory + AI Workshop: Theoretical Physics for AI

Date: April 10, 2025 - 9:45 AM

URL: https://pirsa.org/25040092

Abstract:

We consider the problem of learning a generative model parametrized by a two-layer auto-encoder, and trained with online stochastic gradient descent, to sample from a high-dimensional data distribution with an underlying low-dimensional structure. We provide a tight asymptotic characterization of low-dimensional projections of the resulting generated density, and evidence how mode(l) collapse can arise. On the other hand, we discuss how in a case where the architectural bias is suited to the target density, these simple models can efficiently learn to sample from a binary Gaussian mixture target distribution.

Architectural bias in generative models
–an asymptotic viewpoint

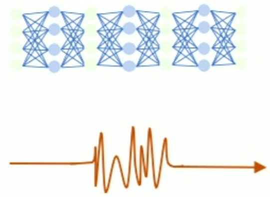
Hugo Cui

PI Theory + AI workshop

Based on

HC, Pehlevan, Lu, *Precise asymptotics of learning diffusion models: theory & insights,* **ArXiv 2025**
HC, Krzakala, Vanden-Eijnden, Zdeborová, *Analysis of a learning a flow-based generative model from finite sample complexity,* **ICLR 2024**

HARVARD UNIVERSITY CENTER OF MATHEMATICAL SCIENCES AND APPLICATIONS
HARVARD UNIVERSITY



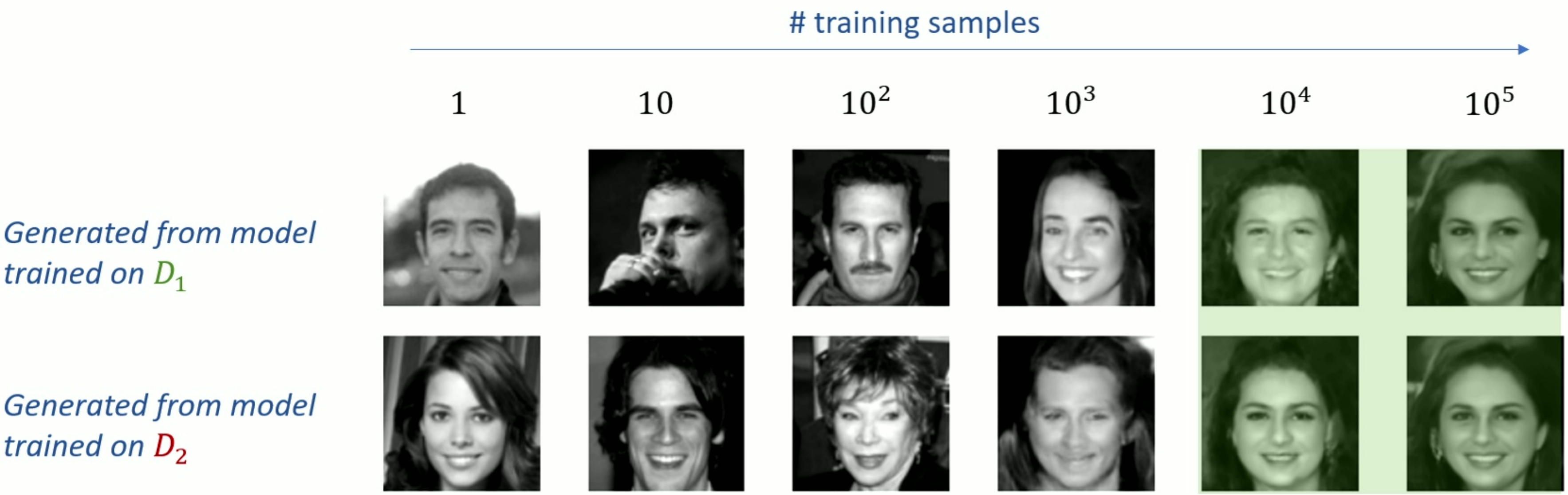
Training set $\sim \rho$

new generated samples

Transport-based generative models learn to sample (generate) complex distributions in high-dimensions from moderate training sets.

→Existence of strong *inductive biases* in the architecture.

Ho, Jain, Abbeel, *Denoising Diffusion Probabilistic Models*, NeurIPS 2020
Sohl-Dickstein et al., *unsupervised learning using nonequi-librium thermodynamics,* ICML 2015
Song and Ermon, *Generative modeling by estimating gradients of the data distribution*. NeurIPS 2019



Kadkhodaie et al., *Generalization in diffusion models arises from geometry-adaptive harmonic representation*, ICLR 2024

Two models trained on disjoint training sets D_1 and D_2 generate the **same image** from a given prompt when trained with sufficiently many samples ($\sim 2-16$x dimension).

How is the distribution of generated samples shaped by the network architecture?

→Try to understand in simple models.

Analysis of the transport only

Chen et al,. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv:2209.11215, 2022.
Biroli et al, Dynamical regimes of diffusion models. Nature Communications, 15(1):9957, 2024
...

Analysis of the transport only

Chen et al,. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv:2209.11215, 2022.
Biroli et al, Dynamical regimes of diffusion models. Nature Communications, 15(1):9957, 2024
...

Sample bounds when density can be perfectly learnt by the model class with enough samples:

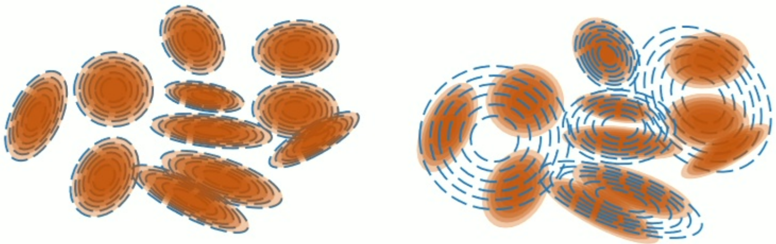
Boffi et al., Shallow diffusion networks provably learn hidden low-dimensional structure., arXiv:2410.11275, 2024.
Chen et al., *Score approx-imation, estimation and distribution recovery of diffusion models on low-dimensional data*, ICML 2023
Oko, Akiyama and Suzuki, *Diffusion models are minimax optimal distribution estimators,* ICML 2023

Analysis of the transport only

Chen et al,. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv:2209.11215, 2022.
Biroli et al, Dynamical regimes of diffusion models. Nature Communications, 15(1):9957, 2024
...

Sample bounds when density can be perfectly learnt by the model class with enough samples:

Boffi et al., Shallow diffusion networks provably learn hidden low-dimensional structure., arXiv:2410.11275, 2024.
Chen et al., *Score approx-imation, estimation and distribution recovery of diffusion models on low-dimensional data*, ICML 2023
Oko, Akiyama and Suzuki, *Diffusion models are minimax optimal distribution estimators,* ICML 2023



Analysis of the transport only

Chen et al,. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. arXiv:2209.11215, 2022.
Biroli et al, Dynamical regimes of diffusion models. Nature Communications, 15(1):9957, 2024
...

Sample bounds when density can be perfectly learnt by the model class with enough samples:

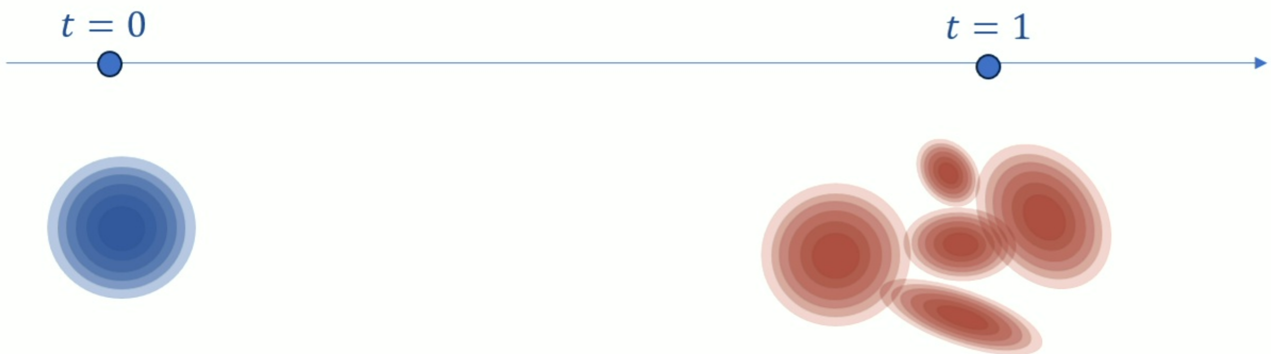
Boffi et al., Shallow diffusion networks provably learn hidden low-dimensional structure., arXiv:2410.11275, 2024.
Chen et al., *Score approx-imation, estimation and distribution recovery of diffusion models on low-dimensional data*, ICML 2023
Oko, Akiyama and Suzuki, *Diffusion models are minimax optimal distribution estimators,* ICML 2023

→*To complement these results* : a **tight** characterization of the generated density in the case **where architecture and target distribution are not perfectly matched**.

Outline

1. Generated density for an auto-encoder parametrized model
2. Failure modes : mode(l) collapse
3. Aligned case: binary isotropic Gaussian mixture distribution.

HC, Pehlevan, Lu, *Precise asymptotics of learning diffusion models: theory & insights,* **ArXiv 2025**
HC, Krzakala, Vanden-Eijnden, Zdeborová, *Analysis of a learning a flow-based generative model from finite sample complexity,* **ICLR 2024**



$X_0 \sim \mathcal{N}(0, \mathbb{I}_d)$ ⟶ $X_1 \sim \rho$

<u>Goal</u>: sample from a **target distribution** ρ

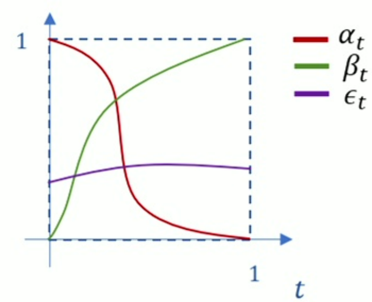
Albergo, Boffi, and Vanden-Eijnden, *Stochastic interpolants: A unifying framework for flows and diffusions*. arXiv:2303.08797, 2023.

The sampling can be done by transporting X_0 through the SDE for $t \in [0,1]$

$$\frac{d}{dt}X_t = \left(\dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2}\right) \boldsymbol{f}(t, X_t) + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2}\right) X_t + \sqrt{2\epsilon_t}\, dW_t$$

For any choice of interpolation schedules st:

$\alpha, \beta \in \mathcal{C}^2([0,1])$
$\alpha_0 = \beta_1 = 1, \alpha_1 = \beta_0 = 0$
$\epsilon_t \geq 0$



Albergo, Boffi, and Vanden-Eijnden, *Stochastic interpolants: A unifying framework for flows and diffusions*. arXiv:2303.08797, 2023.

The sampling can be done by transporting X_0 through the SDE for $t \in [0,1]$

$$\frac{d}{dt} X_t = \left(\dot{\beta}_t - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2}\right) \boldsymbol{f}(t, X_t) + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2}\right) X_t + \sqrt{2\epsilon_t} dW_t$$

Denoising function is the minimizer of a **denoising objective**

$$\boldsymbol{f} = \min_h \int_0^1 \mathbb{E}_{x_1 \sim \rho, x_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \| h(t, \alpha_t x_0 + \beta_t x_1) - x_1 \|^2 dt$$

Learnable from data

Empirical average

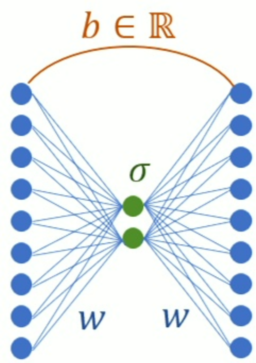
Network param.

Albergo, Boffi, and Vanden-Eijnden, *Stochastic interpolants: A unifying framework for flows and diffusions*. arXiv:2303.08797, 2023.

$$\forall x \in \mathbb{R}^d, \quad f_{b,w}(x) = bx + \frac{w}{\sqrt{d}}\sigma\left(\frac{w^\top x}{\sqrt{d}}\right)$$

Trainable skip connection $b \in \mathbb{R}$

Weight matrix $w \in \mathbb{R}^{d\times r}$

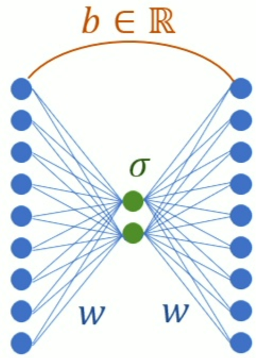


Vincent et al., *Stacked denoising AEs:Learning useful representations in a deep net. with a local denoising criterion*, JMLR 2010

$$\forall x \in \mathbb{R}^d, \quad f_{b,w}(x) = bx + \frac{w}{\sqrt{d}}\sigma\left(\frac{w^\top x}{\sqrt{d}}\right)$$

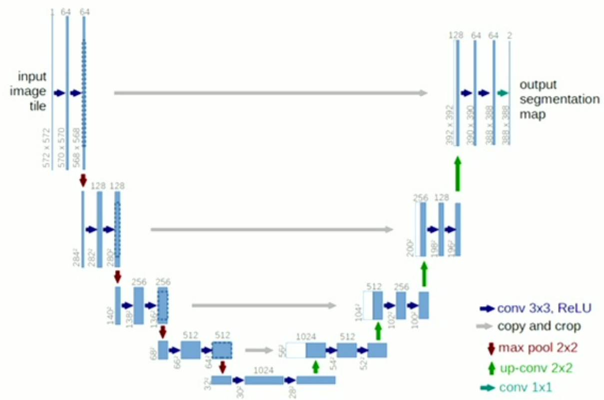
Trainable skip connection $b \in \mathbb{R}$

Weight matrix $w \in \mathbb{R}^{d \times r}$



Remark: U-Nets are used in practice.

- skip connections
- bottlenecks
- convolutional layers



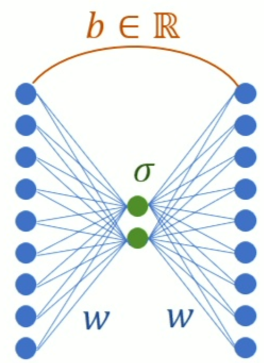
Ronneberger, Fischer, and Brox *U-net: Convolutional networks for biomedical image segmentation*. MICCAI 2015

Vincent et al., *Stacked denoising AEs:Learning useful representations in a deep net. with a local denoising criterion,* JMLR 2010

$\forall x \in \mathbb{R}^d$, $f_{b,w}(x) = bx + \frac{w}{\sqrt{d}}\sigma\left(\frac{w^\top x}{\sqrt{d}}\right)$

Trainable skip connection $b \in \mathbb{R}$

Weight matrix $w \in \mathbb{R}^{d\times r}$



Given a training set of n i.i.d samples $\left\{x_1^\mu \sim \rho, x_0^\mu \sim \mathcal{N}(0, \mathbb{I}_d)\right\}_{\mu=1}^n$ one can train the network $f_{b,w}(x)$ with *online SGD*

$$b_{\mu+1} - b_\mu = -\frac{\eta}{d^2}\left(\partial_b \mathbb{E}_t \left\| x_1^\mu - f_{b_\mu, w_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\right\|^2\right)$$
$$w_{\mu+1} - w_\mu = -\eta\left(\nabla_w \mathbb{E}_t \left\| x_1^\mu - f_{b_\mu, w_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\right\|^2 + \lambda/d\, w_\mu\right)$$

Note $\boldsymbol{\tau} = 2\eta\, {}^n/_d$ and w_τ, b_τ the trained parameters.

Vincent et al., *Stacked denoising AEs:Learning useful representations in a deep net. with a local denoising criterion,* JMLR 2010

$$\frac{d}{dt}X_t = \left(\dot{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2}\right) \boldsymbol{f}(t, X_t) + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2}\right) X_t + \sqrt{2\epsilon_t} dW_t$$

Generated density

$$\frac{d}{dt} X_t = \left(\dot{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t} \beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2} \right) \boldsymbol{f}(t, X_t) + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2} \right) X_t + \sqrt{2\epsilon_t} dW_t$$

$$\frac{d}{dt} X_t = \left(\dot{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t} \beta_t + \epsilon_t \frac{\beta_t}{\alpha_t^2} \right) \boldsymbol{f}_{b_\tau, w_\tau}(X_t) + \left(\frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2} \right) X_t + \sqrt{2\epsilon_t} dW_t$$

Using the trained AE in the generative SDE

$$\hat{\rho}_\tau(t) = \mathrm{Law}[X_t] \; ?$$

Gaussian mixture supported on *a low-dimensional latent manifold*

$$\rho = \int_{\mathbb{R}^K} d\pi(c)\mathcal{N}(\mu(c), \Sigma(c))$$

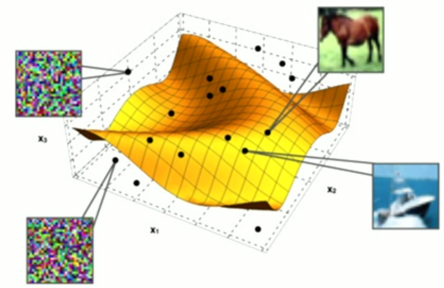


Figure from Goldt et al., *Modelling the influence of data structure in learning in neural networks: the hidden manifold model*, PRX 2020.

Tenenbaum., Silva and Langford, *A global geometric framework for nonlinear dimensionality reduction*. science, 2000
Weinberger and Saul, *Unsupervised learning of image manifolds by semidefinite programming*. Int. journal of computer vision, 2006.

Gaussian mixture supported on *a low-dimensional latent manifold*

$$\rho = \int_{\mathbb{R}^K} d\pi(c)\mathcal{N}(\mu(c), \Sigma(c))$$

centroids $\mu: \mathbb{R}^K \to \mathbb{R}^d$ $\exists\, D > 0$, w.p. 1, $\|\mu(c)\| \leq D$

$K = \dim \operatorname{span}\{\mu(c)\}_c$ is low

covariances $\Sigma: \mathbb{R}^K \to \mathcal{S}^d(\mathbb{R})$ assumed jointly diagonalizable, with a well-defined joint limiting spectral density.

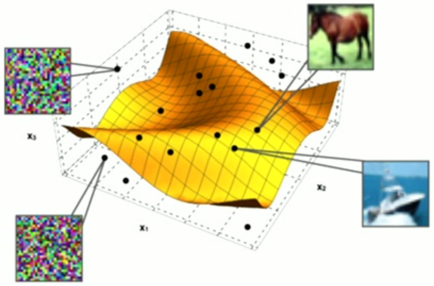


Figure from Goldt et al., *Modelling the influence of data structure in learning in neural networks: the hidden manifold model*, PRX 2020.

Tenenbaum., Silva and Langford, *A global geometric framework for nonlinear dimensionality reduction*. science, 2000
Weinberger and Saul, *Unsupervised learning of image manifolds by semidefinite programming*. Int. journal of computer vision, 2006.

Gaussian mixture supported on *a low-dimensional latent manifold*

$$\rho = \int_{\mathbb{R}^K} d\pi(c)\mathcal{N}(\mu(c), \Sigma(c))$$

centroids $\mu: \mathbb{R}^K \to \mathbb{R}^d$ $\exists\, D > 0$, w.p. 1, $\|\mu(c)\| \leq D$

$K = \dim \mathrm{span}\{\mu(c)\}_c$ is low

covariances $\Sigma: \mathbb{R}^K \to \mathcal{S}^d(\mathbb{R})$ assumed jointly diagonalizable, with a well-defined joint limiting spectral density.

Average extension of the density $\Lambda = \int d\pi(c) \frac{1}{d} \mathrm{Tr}[\Sigma(c)]$

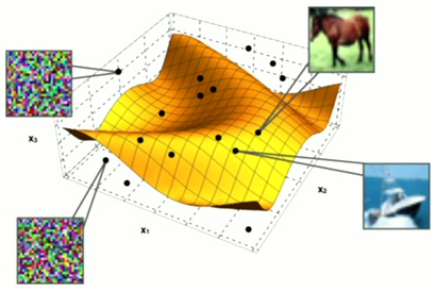


Figure from Goldt et al., *Modelling the influence of data structure in learning in neural networks: the hidden manifold model*, PRX 2020.

Tenenbaum., Silva and Langford, *A global geometric framework for nonlinear dimensionality reduction*. science, 2000
Weinberger and Saul, *Unsupervised learning of image manifolds by semidefinite programming*. Int. journal of computer vision, 2006.

Target density $\rho = \int_{\mathbb{R}^\kappa} d\pi(c)\mathcal{N}(\mu(c), \Sigma(c))$

Architecture $f_{b,w}(x) = bx + \frac{w}{\sqrt{d}}\sigma\left(\frac{w^\top x}{\sqrt{d}}\right)$

Learning
$$b_{\mu+1} - b_\mu = -\frac{\eta}{d^2}\left(\partial_b \mathbb{E}_t \left\| x_1^\mu - f_{b_\mu, w_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\right\|^2\right)$$
$$w_{\mu+1} - w_\mu = -\eta\left(\nabla_w \mathbb{E}_t \left\| x_1^\mu - f_{b_\mu, w_\mu}(\alpha_t x_0^\mu + \beta_t x_1^\mu)\right\|^2 + \lambda/d w_\mu\right)$$

for a time τ

Sampling
$$\frac{d}{dt}X_t = \Gamma_t \frac{w_\tau}{\sqrt{d}}\sigma\left(\frac{w_\tau^\top X_t}{\sqrt{d}}\right) + \Delta_t^\tau X_t + \sqrt{2\epsilon_t}dW_t$$

for a time t

with $\Gamma_t = \left(\dot{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t + \epsilon_t\frac{\beta_t}{\alpha_t^2}\right)$ $\quad \Delta_t^\tau = b_\tau \Gamma_t + \frac{\dot{\alpha}_t}{\alpha_t} - \frac{\epsilon_t}{\alpha_t^2}$

Generated density $\hat{\rho}_\tau(t) = \mathrm{Law}[X_t]$

21

Asymptotic limit $n, d \to \infty$ with $n/d, \kappa, r, D, K = \Theta_d(1)$

Finite width, large amount of data, large dimension

Saad and Solla, *Exact solution for on-line learning in multilayer neural networks*, PRL 1995,
…

Gabrié, Mean-Field inference methods for neural networks, J. Phys. A 2020.
HC, High-dimensional learning of narrow networks, J.Stat Mech 2025

Tight characterization of low-dimensional projections of the generated density $\hat{\rho}_\tau(t)$

Consider a low-dimensional subspace $\mathcal{E} \subset \mathbb{R}^d$, with $\dim\mathcal{E} = R = \Theta_d(1)$. The distribution of the R −dimensional projection $\Pi_\mathcal{E} X_t$ is given by

$$\Pi_\mathcal{E} X_t =^d \Theta_\tau^\top Q_\tau^+ Z_t + Y_t$$

Tight characterization of low-dimensional projections of the generated density $\hat{\rho}_\tau(t)$

Consider a low-dimensional subspace $\mathcal{E} \subset \mathbb{R}^d$, with $\dim\mathcal{E} = R = \Theta_d(1)$. The distribution of the R −dimensional projection $\Pi_\mathcal{E} X_t$ is given by

$$\Pi_\mathcal{E} X_t =^d \Theta_\tau^\top Q_\tau^+ Z_t + Y_t$$

Where:

- $Y_t \in \mathbb{R}^R$ is Gaussian

$$Y_t \sim \mathcal{N}\left(0_R, e^{2\int_0^t \Delta_s^\tau ds}\left[1 + 2\int_0^t \epsilon_s e^{-2\int_0^s \Delta_z^\tau dz} ds\right](\mathbb{I}_R - \Theta_\tau^\top Q_\tau^+ \Theta_\tau)\right)$$

Tight characterization of low-dimensional projections of the generated density $\hat{\rho}_\tau(t)$

Consider a low-dimensional subspace $\mathcal{E} \subset \mathbb{R}^d$, with $\dim\mathcal{E} = R = \Theta_d(1)$. The distribution of the R −dimensional projection $\Pi_\mathcal{E} X_t$ is given by

$$\Pi_\mathcal{E} X_t =^d \Theta_\tau^\top Q_\tau^+ Z_t + Y_t$$

Where:

- $Y_t \in \mathbb{R}^R$ is Gaussian

$$Y_t \sim \mathcal{N}\left(0_R, e^{2\int_0^t \Delta_s^\tau ds}\left[1 + 2\int_0^t \epsilon_s e^{-2\int_0^s \Delta_z^\tau dz} ds\right](\mathbb{I}_R - \Theta_\tau^\top Q_\tau^+ \Theta_\tau)\right)$$

- $Z_t \in \mathbb{R}^r$ is distributed as the solution of the SDE

$$\frac{d}{dt} Z_t = \Delta_t^\tau Z_t + \Gamma_t Q_\tau \sigma(Z_t) + \sqrt{2\epsilon_t} Q_\tau^{1/2} W_t$$

From initialization $Z_0 \sim \mathcal{N}(0_r, Q_\tau)$

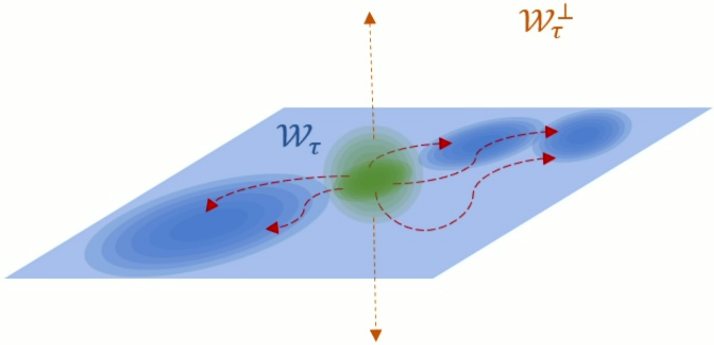
Tight characterization of low-dimensional projections of the generated density $\hat{\rho}_\tau(t)$

Consider a low-dimensional subspace $\mathcal{E} \subset \mathbb{R}^d$, with $\dim\mathcal{E} = R = \Theta_d(1)$. The distribution of the R −dimensional projection $\Pi_{\mathcal{E}} X_t$ is given by

$$\Pi_{\mathcal{E}} X_t =^d \Theta_\tau^\top Q_\tau^+ Z_t + Y_t$$

Where:

- $Y_t \in \mathbb{R}^R$ is Gaussian $$Y_t \sim \mathcal{N}\left(0_R, e^{2\int_0^t \Delta_s^\tau ds}\left[1 + 2\int_0^t \epsilon_s e^{-2\int_0^s \Delta_z^\tau dz} ds\right](\mathbb{I}_R - \Theta_\tau^\top Q_\tau^+ \Theta_\tau)\right)$$
- $Z_t \in \mathbb{R}^r$ is distributed as the solution of the SDE $$\frac{d}{dt} Z_t = \Delta_t^\tau Z_t + \Gamma_t Q_\tau \sigma(Z_t) + \sqrt{2\epsilon_t} Q_\tau^{1/2} W_t$$ From initialization $Z_0 \sim \mathcal{N}(0_r, Q_\tau)$
- The parameters $\Theta_\tau \in \mathbb{R}^{r\times R}$, $Q_\tau \in \mathbb{R}^{r\times r}$ are the solutions of a set of 5 coupled **low-dimensional** deterministic **ODEs**.

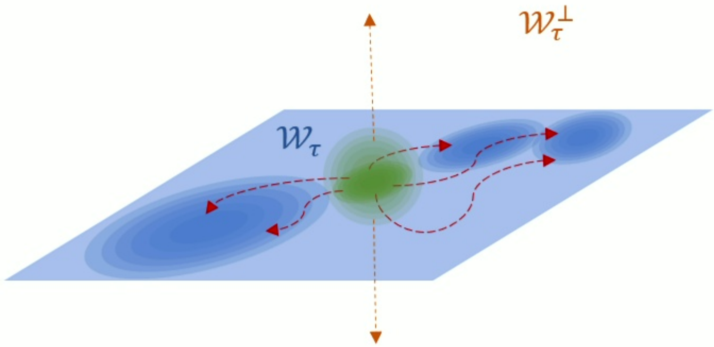


$$\frac{d}{dt}X_t = \Gamma_t \frac{w_\tau}{\sqrt{d}}\sigma\left(\frac{w_\tau^\top X_t}{\sqrt{d}}\right) + \Delta_t^\tau X_t + \sqrt{2\epsilon_t}dW_t$$

Non-linear transport in $\mathcal{W}_\tau = \mathrm{span}(\{w_i\}_{i=1}^r)$

Linear in $\mathcal{W}_\tau^\perp$

27



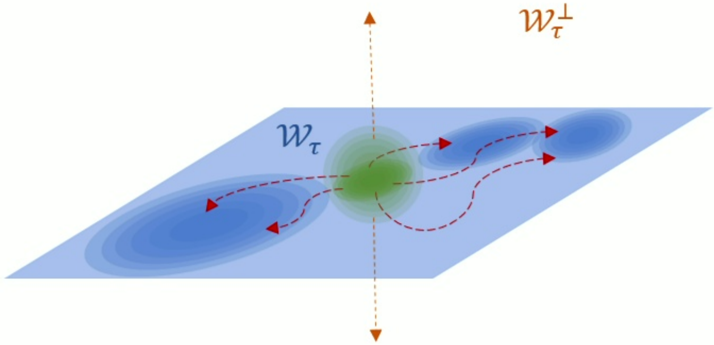
$$\frac{d}{dt}X_t = \boxed{\Gamma_t \frac{w_\tau}{\sqrt{d}}\sigma\left(\frac{{w_\tau}^\top X_t}{\sqrt{d}}\right)} + \Delta_t^\tau X_t + \sqrt{2\epsilon_t}dW_t$$

Non-linear transport in $\mathcal{W}_\tau = \mathrm{span}(\{w_i\}_{i=1}^r)$

Linear in $\mathcal{W}_\tau^\perp$

Dynamics of $Z_t = \frac{{w_\tau}^\top X_t}{\sqrt{d}}$

$$\frac{d}{dt}Z_t = \Delta_t^\tau Z_t + \Gamma_t \frac{w_\tau^\top w_\tau}{d}\sigma(Z_t) + \sqrt{2\epsilon_t}\left(\frac{w_\tau^\top w_\tau}{d}\right)^{\frac{1}{2}} W_t$$



$$\frac{d}{dt}X_t = \boxed{\Gamma_t \frac{w_\tau}{\sqrt{d}}\sigma\left(\frac{{w_\tau}^\top X_t}{\sqrt{d}}\right)} + \Delta_t^\tau X_t + \sqrt{2\epsilon_t}dW_t$$

Non-linear transport in $\mathcal{W}_\tau = \mathrm{span}(\{w_i\}_{i=1}^r)$

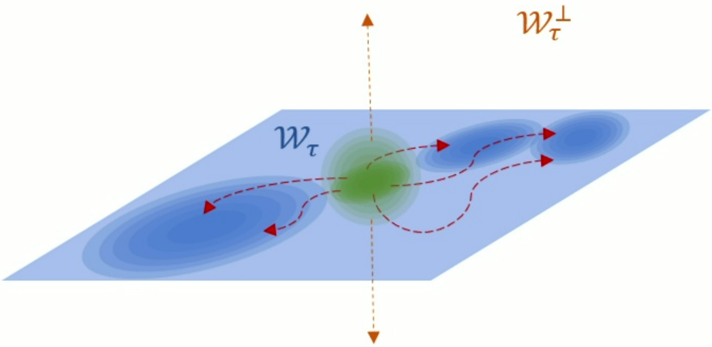
Linear in $\mathcal{W}_\tau^\perp$

Dynamics of $Z_t = \frac{{w_\tau}^\top X_t}{\sqrt{d}}$

$$\frac{d}{dt}Z_t = \Delta_t^\tau Z_t + \Gamma_t \frac{w_\tau^\top w_\tau}{d}\sigma(Z_t) + \sqrt{2\epsilon_t}\left(\frac{w_\tau^\top w_\tau}{d}\right)^{\frac{1}{2}} W_t$$

The SGD dynamics of the summary statistic $Q_\tau = \frac{w_\tau^\top w_\tau}{d}$ (and others) self-average and can be characterized in closed-form by a set of low-dimensional ODEs.

Saad and Solla, *Exact solution for on-line learningin multilayer neural networks*, PRL 1995



Intuition:

- The network **identifies** a r −dimensional $\mathcal{W}_\tau$ subspace where the target ρ has important structure, and implements a non-linear transport.
- It approximates ρ in the orthogonal space by an **isotropic Gaussian**, whose variance is tuned by the skip connection strength.

Special case: linear networks $\sigma(x) = x$

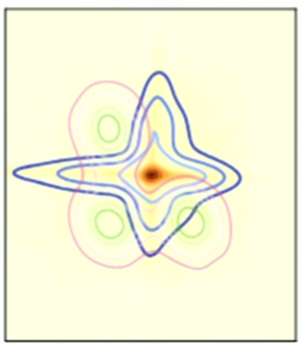
- Linear networks approximately learn ≈ **principal components** $\mathcal{W}_\tau \approx \mathrm{PCA}_r\left[\{x_1^\mu\}_\mu\right]$

Pretorius et al,. *Learning dynamics of linear denoising autoencoders*. ICML, 2018. , …..

- The linear diffusion model does a Gaussian approximation in the principal space. In the orthogonal space, approximates by an isotropic Gaussian.

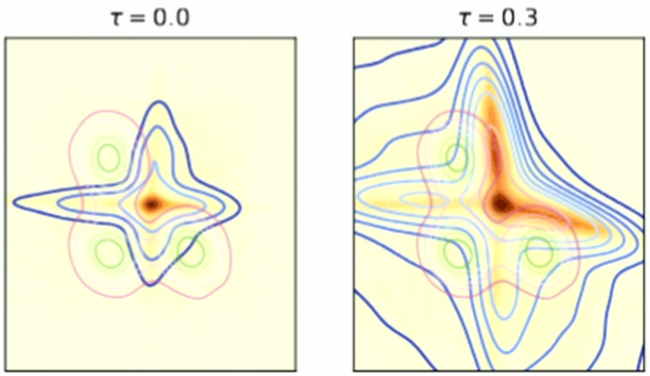
Generated density with training time : Gaussian mixture

σ = ReLU activation,
r = 4 hidden units



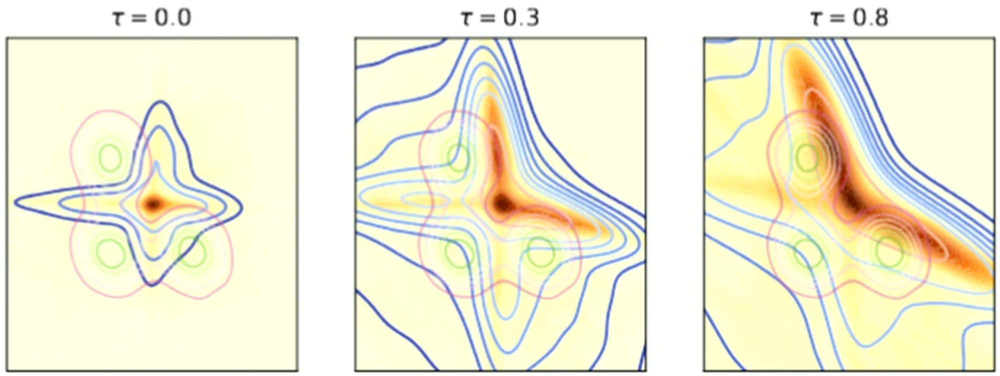
Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$

σ = ReLU activation,
$r = 4$ hidden units



Target density ρ (theory) Generated density $\hat{\rho}$ (exp) Generated density $\hat{\rho}$

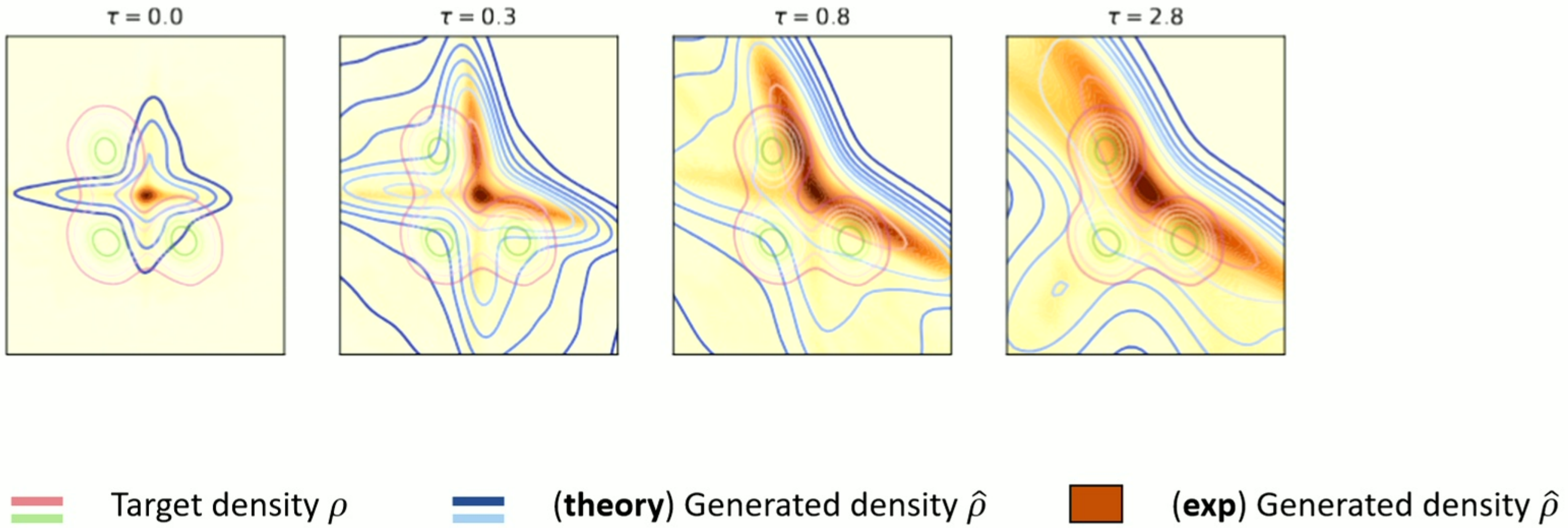
σ = ReLU activation,
$r = 4$ hidden units



Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$

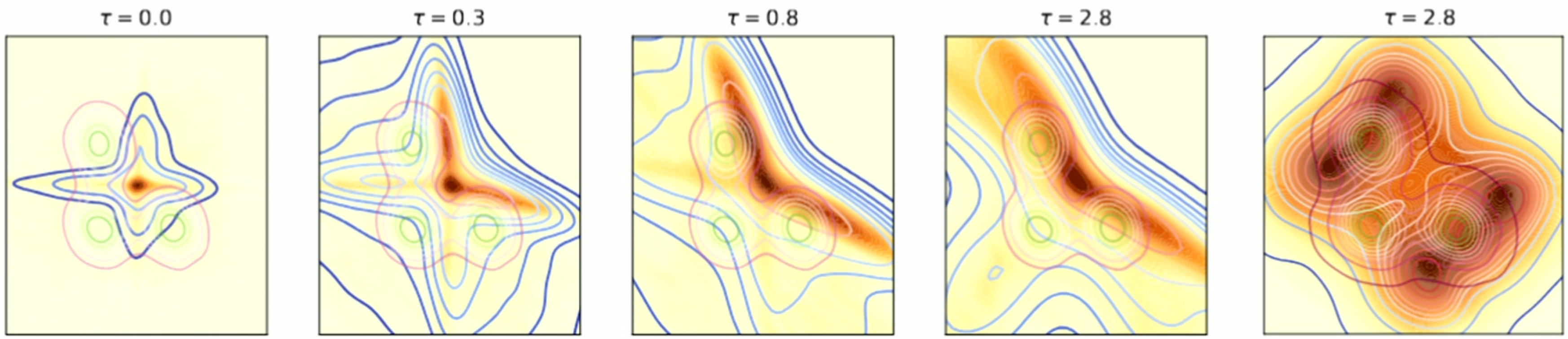
Generated density with training time : Gaussian mixture

$\sigma =$ ReLU activation,
$r = 4$ hidden units

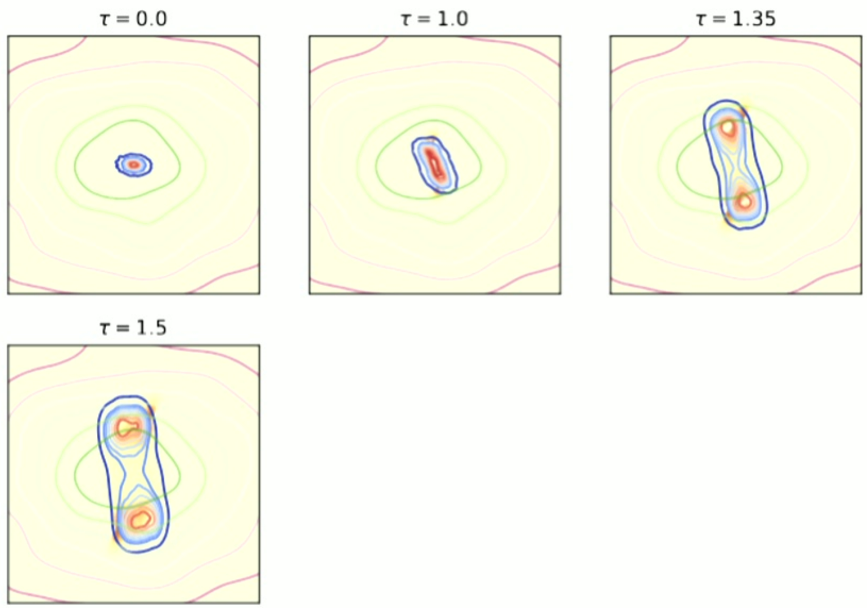


σ = ReLU activation,
r = 4 hidden units

σ = tanh activation,
r = 2 hidden units

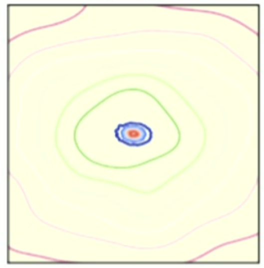


Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$



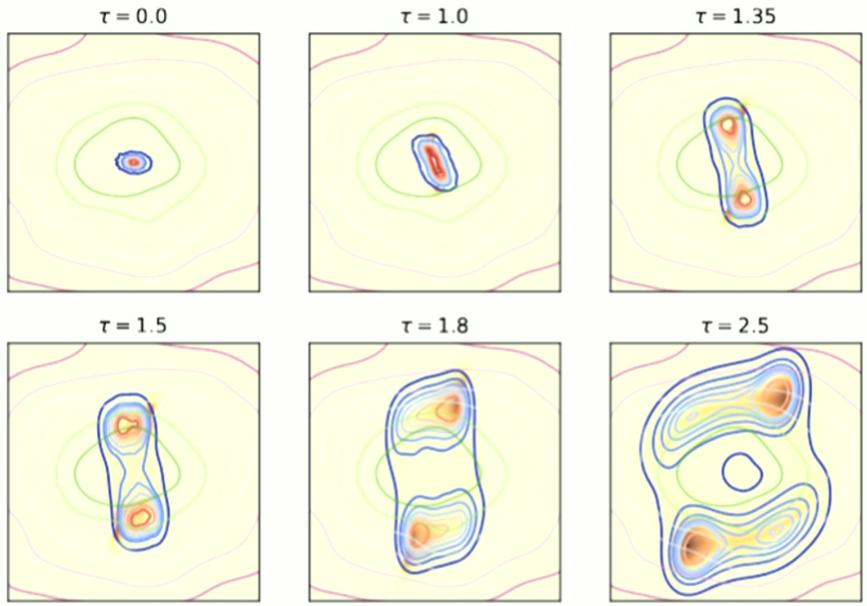
σ = tanh activation,
$r = 2$ hidden units
Gaussian ρ with MNIST covariance

Target density ρ | (**theory**) Generated density $\hat{\rho}$ | (**exp**) Generated density $\hat{\rho}$



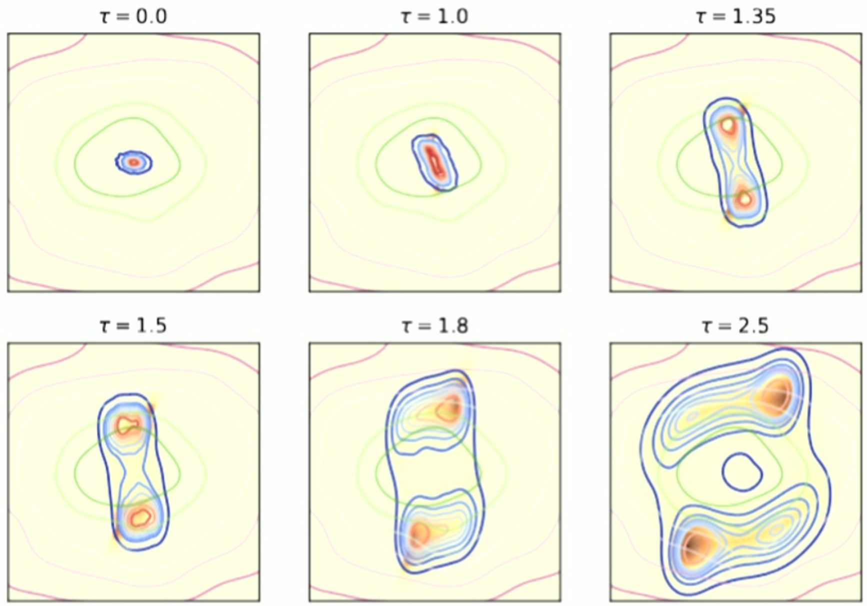
σ = tanh activation,
r = 2 hidden units
Gaussian ρ with MNIST covariance

Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$

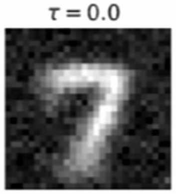


σ = tanh activation,
$r = 2$ hidden units
Gaussian ρ with MNIST covariance

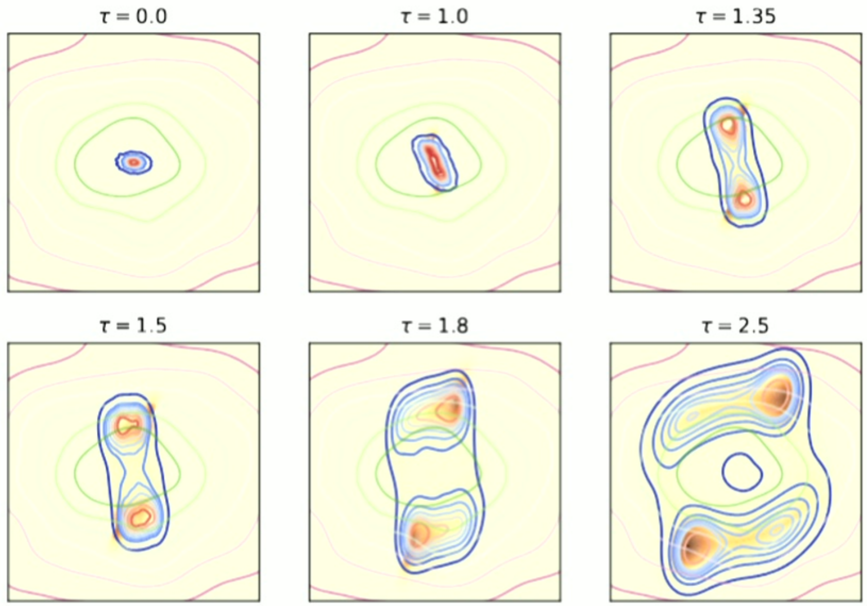
Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$



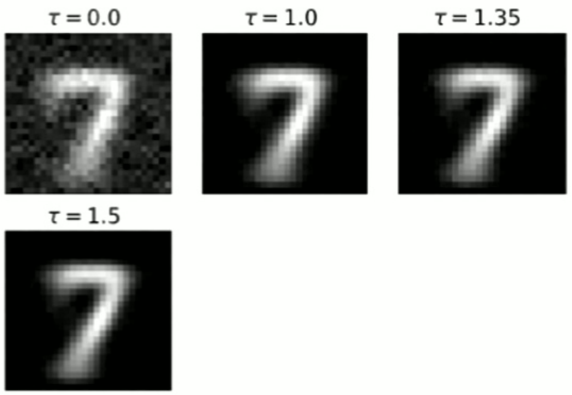
σ = tanh activation,
$r = 2$ hidden units
Gaussian ρ with MNIST covariance



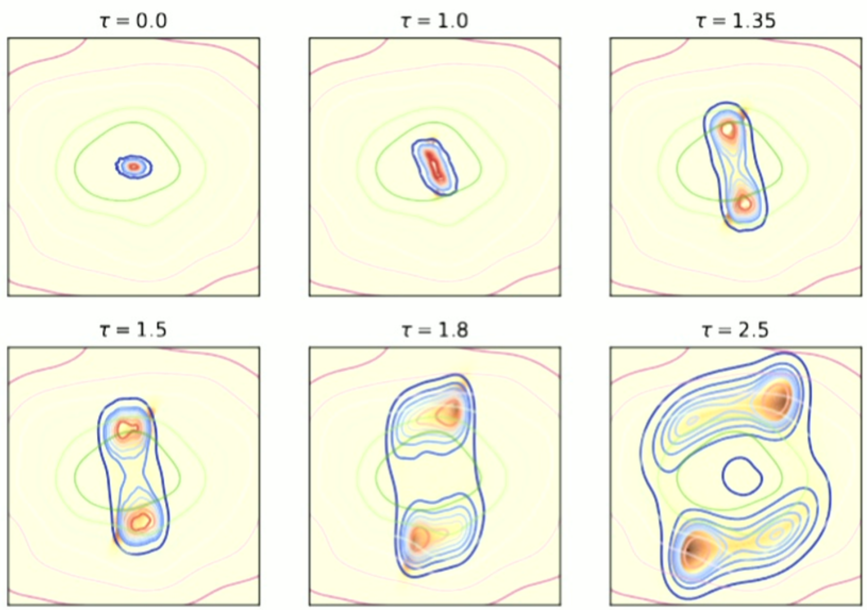
Target density ρ | (**theory**) Generated density $\hat{\rho}$ | (**exp**) Generated density $\hat{\rho}$



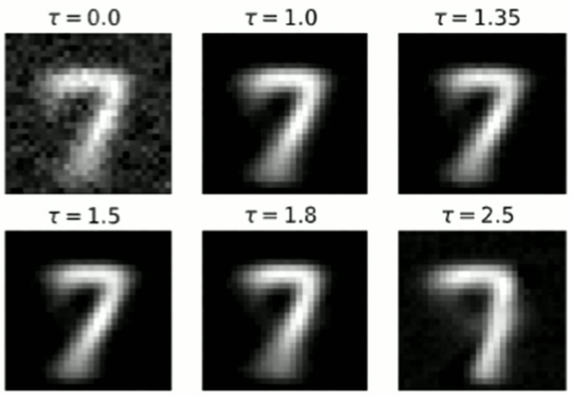
σ = tanh activation,
r = 2 hidden units
Gaussian ρ with MNIST covariance



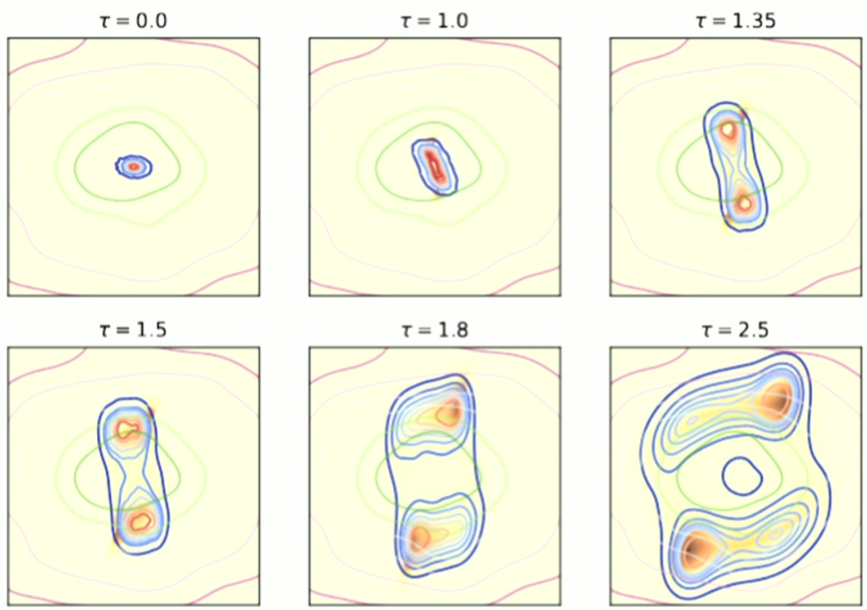
Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$



σ = tanh activation,
$r = 2$ hidden units
Gaussian ρ with MNIST covariance



Target density ρ (**theory**) Generated density $\hat{\rho}$ (**exp**) Generated density $\hat{\rho}$



Closed-form expression for the trained skip connection

$$b_\tau = \frac{\Lambda \mathbb{E}_t[\beta_t]\left[1-(1-b_0)e^{-(\Lambda \mathbb{E}_t[\beta_t^2]+\mathbb{E}_t[\alpha_t^2])\tau}\right]}{\Lambda \mathbb{E}_t[\beta_t^2]+\mathbb{E}_t[\alpha_t^2]}$$

Average cov. eigenvalue $\boldsymbol{\Lambda} = \int d\pi(c)\frac{1}{d}\mathrm{Tr}[\Sigma(c)]$

is typically **small** in real datasets, causing
≈***mode collapse***

Goodfellow et al., *Generative adversarial nets.* NeurIPS 2014.

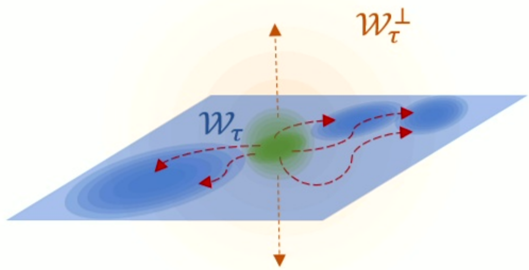
Target density ρ — (**theory**) Generated density $\hat{\rho}$ — (**exp**) Generated density $\hat{\rho}$

Can this bias be ***aggravated*** when using synthetic data to train a new generative model ?

$$\rho \to \hat{\rho}^{(1)} \to \hat{\rho}^{(2)} \to \cdots \to \hat{\rho}^{(g)}$$

Can this bias be ***aggravated*** when using synthetic data to train a new generative model ?

$$\rho \to \hat{\rho}^{(1)} \to \hat{\rho}^{(2)} \to \cdots \to \hat{\rho}^{(g)}$$



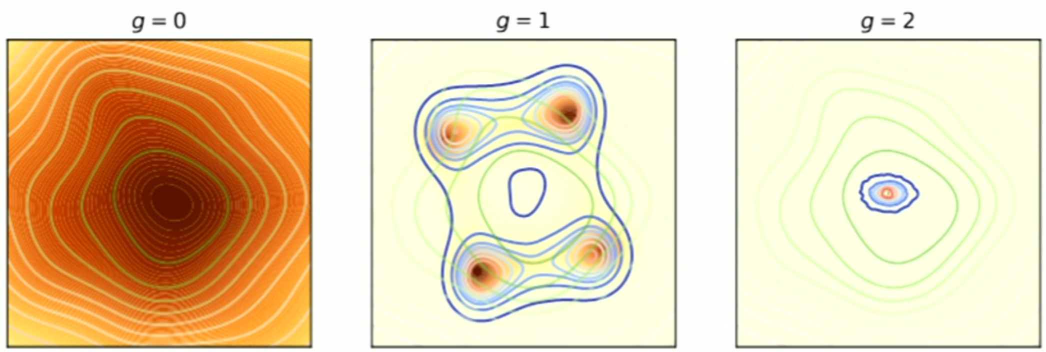
Remark: *Manifold form of the generated density*

$\hat{\rho}^{(1)}$ is still of the form $\int d\pi(c)\mathcal{N}\big(\mu(c), \Sigma(c)\big)$, with $\mu(c) = c$ and

$$\pi = \Pi_{\mathcal{W}_\tau} \hat{\rho}^{(1)}$$

$$\Sigma(c) = e^{2\int_0^t \Delta_s^\tau ds}\left[1 + 2\int_0^t \epsilon_s e^{-2\int_0^s \Delta_z^\tau dz} ds\right]\Pi_{\mathcal{W}_\tau^\perp}$$

Thus the analysis *carries over iteratively* to generations $\hat{\rho}^{(2)}$, ...



Target density $\hat{\rho}^{(g)}$ **(theory)** Generated density $\hat{\rho}^{(g+1)}$ **(exp)** Generated density $\hat{\rho}^{(g+1)}$

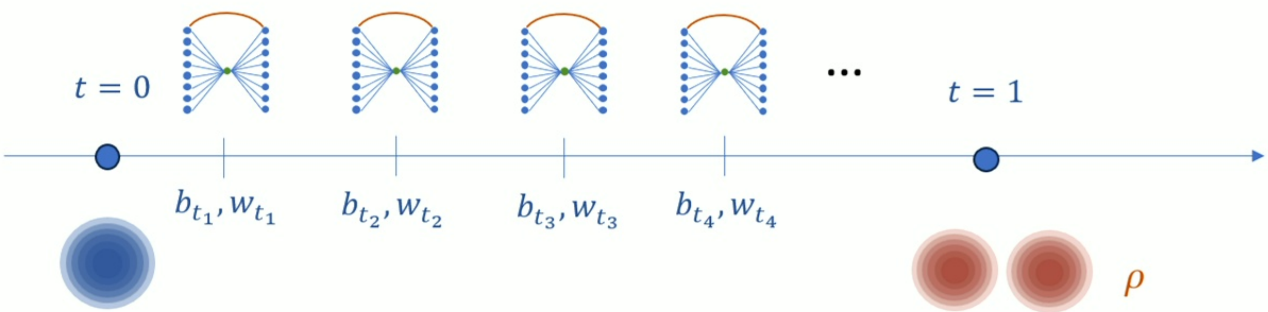
Shumailov et al., *Ai models collapse when trained onrecursively generated data.* Nature, 2024

Binary, isotropic Gaussian mixture $\rho = {}^1\!/_2\, \mathcal{N}(-\mu, \sigma^2\, \mathbb{I}_d) + {}^1\!/_2\, \mathcal{N}(+\mu, \sigma^2 \mathbb{I}_d)$

Binary, isotropic Gaussian mixture $\rho = {}^1\!/_2\,\mathcal{N}(-\mu, \sigma^2\,\mathbb{I}_d) + {}^1\!/_2\,\mathcal{N}(+\mu, \sigma^2\mathbb{I}_d)$

At *each sampling time*, train a **separate** AE with $r = 1$ hidden unit and $\sigma =$ sign activation

$$b_t, w_t = \mathrm{argmin}_{\theta\in\mathbb{R}^{d\times r}} \sum_{\mu=1}^{n} \left\|f_{b,w}(\alpha_t x_0^\mu + \beta_t x_1^\mu) - x_1^\mu\right\|_2^2 + \lambda\|w\|^2$$



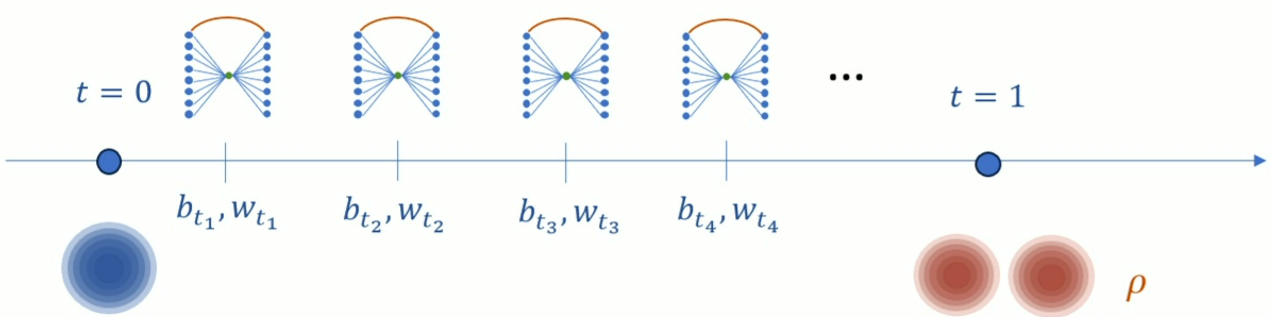
Binary, isotropic Gaussian mixture $\rho = {}^1\!/_2\,\mathcal{N}(-\mu, \sigma^2\,\mathbb{I}_d) + {}^1\!/_2\,\mathcal{N}(+\mu, \sigma^2\mathbb{I}_d)$

At *each sampling time*, train a **separate** AE with $r = 1$ hidden unit and $\sigma =$ sign activation

$$b_t, w_t = \mathrm{argmin}_{\theta \in \mathbb{R}^{d\times r}} \sum_{\mu=1}^{n} \left\| f_{b,w}(\alpha_t x_0^\mu + \beta_t x_1^\mu) - x_1^\mu \right\|_2^2 + \lambda \|w\|^2$$

Sampling :

$$\frac{d}{dt} X_t = \left(\dot{\beta_t} - \frac{\dot{\alpha}_t}{\alpha_t}\beta_t\right) f_{b_t, w_t}(X_t) + \frac{\dot{\alpha}_t}{\alpha_t} X_t$$



Closed form characterization of the dynamics

In the asymptotic limit $d \to \infty$ with $n = \Theta_d(1)$, $\|\mu\| = \Theta_d(\sqrt{d})$, the sampling dynamic is non-linear in $\mathrm{span}(\mu, \xi, \eta)$ where

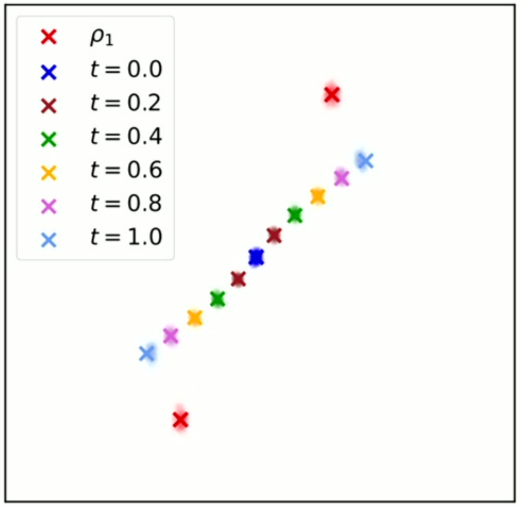
$$\boldsymbol{\xi} \equiv \sum_{\mu=1}^{n} s^\mu \boldsymbol{x}_0^\mu, \qquad \boldsymbol{\eta} \equiv \sum_{\mu=1}^{n} s^\mu (\boldsymbol{x}_1^\mu - s^\mu \boldsymbol{\mu}),$$

The coordinates M_t, Q_t^ξ, Q_t^η of a sample X_t follow the ODEs

$$\begin{cases} \frac{d}{dt} M_t = \frac{\left(\dot\beta(t)\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)+\dot\alpha(t)\alpha(t)(\lambda+n-1)\right)M_t + \left(\alpha(t)\dot\beta(t)-\dot\alpha(t)\beta(t)\right)\frac{n\alpha(t)(\lambda+n-1)}{\lambda+n}}{\alpha(t)^2(\lambda+n-1)+\beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \\ \frac{d}{dt} Q_t^\xi = \frac{\left(\dot\beta(t)\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)+\dot\alpha(t)\alpha(t)(\lambda+n-1)\right)Q_t^\xi - \left(\alpha(t)\dot\beta(t)-\dot\alpha(t)\beta(t)\right)\frac{\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)}{\lambda+n}}{\alpha(t)^2(\lambda+n-1)+\beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \\ \frac{d}{dt} Q_t^\eta = \frac{\left(\dot\beta(t)\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)+\dot\alpha(t)\alpha(t)(\lambda+n-1)\right)Q_t^\eta + \left(\alpha(t)\dot\beta(t)-\dot\alpha(t)\beta(t)\right)\frac{\alpha(t)(\lambda+n-1)}{\lambda+n}}{\alpha(t)^2(\lambda+n-1)+\beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)} \end{cases}$$

The component $X_t^\perp$ orthogonal to $\mathrm{span}(\mu, \xi, \eta)$ evolves linearly

$$\frac{d}{dt}\boldsymbol{X}_t^\perp = \frac{\left(\dot\beta(t)\beta(t)(\lambda(1+\sigma^2)+(n-1)\sigma^2)+\dot\alpha(t)\alpha(t)(\lambda+n-1)\right)}{\alpha(t)^2(\lambda+n-1)+\beta(t)^2(\lambda(1+\sigma^2)+(n-1)\sigma^2)}\boldsymbol{X}_t^\perp$$



HC, Krzakala, Vanden-Eijnden, Zdeborová, *Analysis of a learning a flow-based generative model from finite sample complexity,* **ICLR 2024**

Corollary

The *mixture Wasserstein distance* between the target ρ and the generated density $\hat{\rho}$ decays as

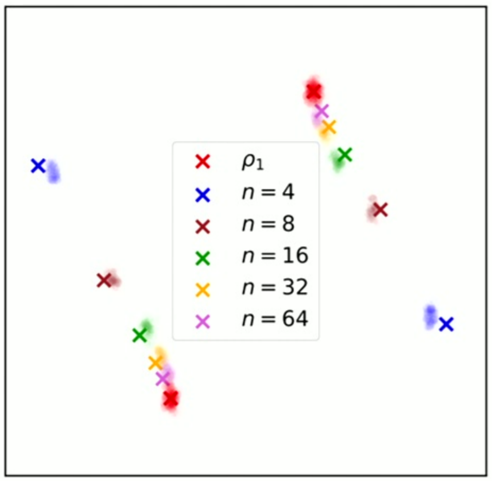
$$\mathrm{M}\mathcal{W}_2[\rho,\hat{\rho}] = O\left(\frac{1}{n}\right)$$

HC, Krzakala, Vanden-Eijnden, Zdeborová, *Analysis of a learning a flow-based generative model from finite sample complexity,* **ICLR 2024**

Corollary

The *mixture Wasserstein distance* between the target ρ and the generated density $\hat{\rho}$ decays as

$$\mathrm{M}\mathcal{W}_2[\rho, \hat{\rho}] = O\left(\frac{1}{n}\right)$$

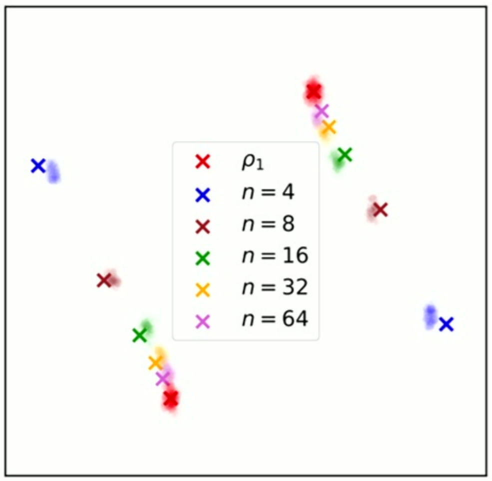


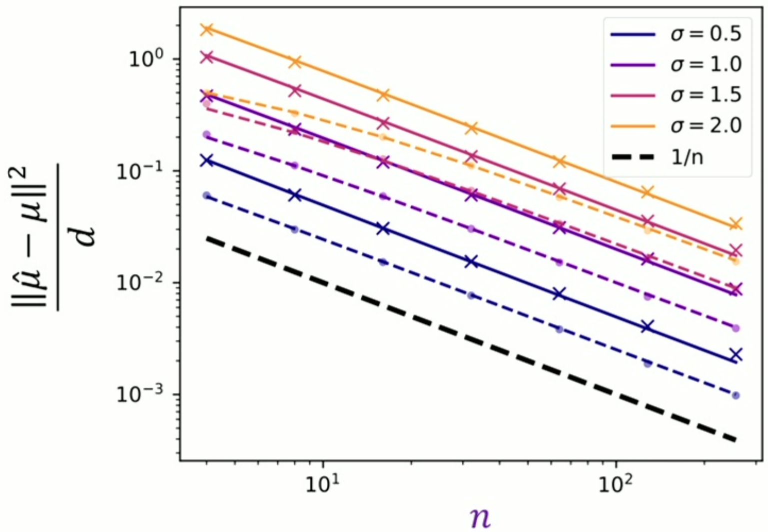
HC, Krzakala, Vanden-Eijnden, Zdeborová, *Analysis of a learning a flow-based generative model from finite sample complexity,* **ICLR 2024**

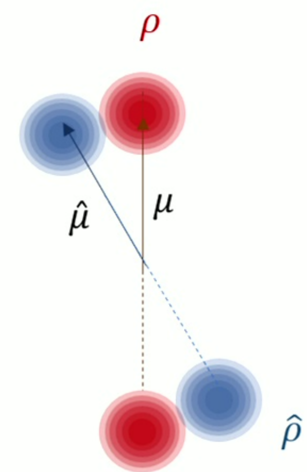
Corollary

The *mixture Wasserstein distance* between the target ρ and the generated density $\hat{\rho}$ decays as

$$\mathrm{M}\mathcal{W}_2[\rho, \hat{\rho}] = O\left(\frac{1}{n}\right)$$







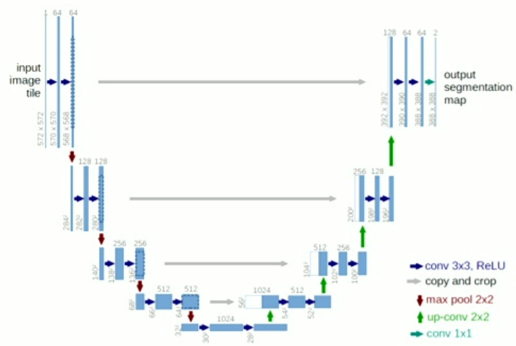
HC, Krzakala, Vanden-Eijnden, Zdeborová, *Analysis of a learning a flow-based generative model from finite sample complexity,* **ICLR 2024**

Intuition : The optimal denoising function follows from *Tweedie's formula* (Empirical Bayes) and is of the **same functional form** as the AE

$$f_t^\star(x) = \frac{\beta(t)\sigma^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} x + \frac{\alpha(t)^2}{\alpha(t)^2 + \beta(t)^2\sigma^2} \mu \times \tanh\left(\frac{\beta(t)}{\alpha(t)^2 + \beta(t)^2\sigma^2} \mu^\top x \right)$$

→ The architectural bias *is **aligned** with the target distribution*.

Bradley Efron. *Tweedie's formula and selection bias*. Journal of the American Statistical Association, 2011
Robbins, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Contributions to the Theory of Statistics.
Koichi Miyasawa. *An empirical Bayes estimator of the mean of a normal population*. Bulletin of the International Statistical Institute, 1961



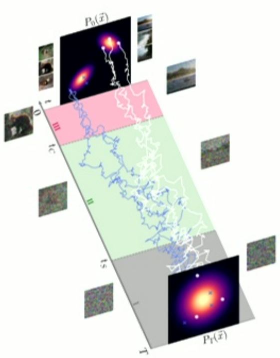
<u>Inductive bias of **Unets**?</u>

U-nets are suited to data with a *hierarchical structure*

Kadkhodaie et al., *Generalization in diffusion models arises from geometry-adaptive harmonic representation*, ICLR 2024
Mei, *S. U-nets as belief propagation: Efficient classification,denoising, and diffusion in generative hierarchical models,* arXiv:2404.18444, 2024.

(Recall also Alessandro's talk!)

Ronneberger, Fischer, and Brox *U-net: Convolutional networks for biomedical image segmentation*. MICCAI 2015



For infinitely expressive networks who can perfectly overfit the data, dynamical transitions in the sampling process.

Biroli et al, *Dynamical Regimes of Diffusion Models,* Nature Comm. 2024

<u>How are they altered for networks with finite expressivity?</u>

54

Collaborators

Yue M Lu
(*Harvard*)

Cengiz Pehlevan
(*Harvard*)

Lenka Zdeborová
(*EPFL*)

Florent Krzakala
(*EPFL*)

Eric Vanden-Eijnden
(*NYU*)

Thank you for your attention !