Generative AI and Diffusion Models a Statistical Physics Analysis

Giulio Biroli

In collaboration with Marc Mézard (U. Bocconi)

and Tony Bonnaire (ENS) , Valentin De Bortoli (DeepMind)

JSTAT 2023+ ArXiv 2402.18491 to appear in Nature Comm

Generative AI & Diffusion Models

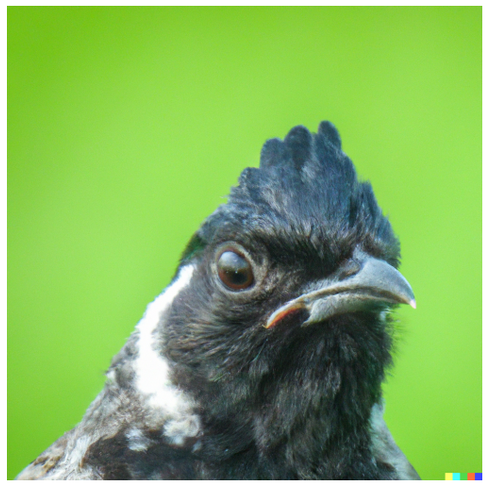
- Generation of images, videos, audios, 3-d scenes,… e.g. Dall-E

Images (fake celebrities)

Audios

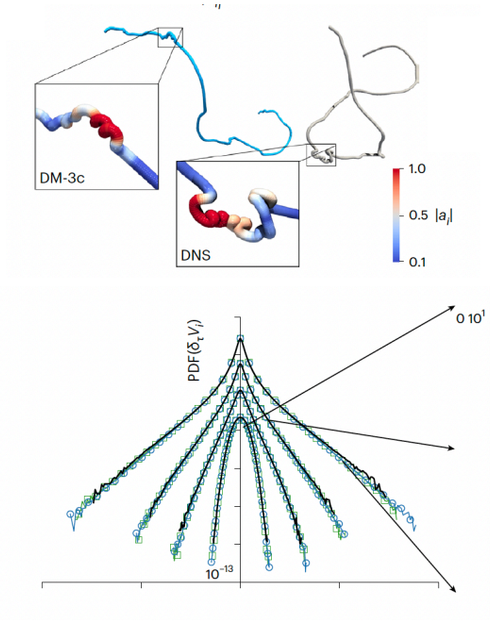
Videos

- Wide range of applications: from text-to-image to science (generation & sampling)



An angry bird

Statistical Physics and Machine Learning



Lagrangian turbulence
Li, Biferale et al 2024

Diffusion Models

Training set: a set of images $\vec{a}^{\mu} \in \mathbb{R}^d$ $\mu = 1, \ldots, n$

d is the dimension of the data, n their number

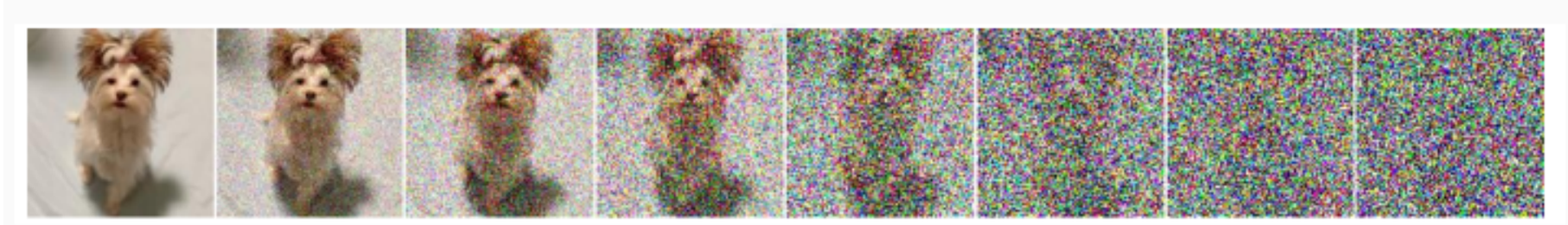
Langevin equation for an Ornstein-Uhlenbeck process (equilibrium relaxation in a well)

$$\frac{d\vec{x}}{dt} = -\vec{x} + \vec{\eta}(t) \qquad \langle \eta_i(t)\eta_j(t') \rangle = 2T\delta_{ij}\delta(t-t')$$

$\vec{x}^{\mu}(t=0) = \vec{a}^{\mu}$ It transforms the data in iid Gaussian $\mathcal{N}(0,1)$ at t>>1

$$P_t(\vec{x}) = \int d\vec{a}\; P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right) = \int d\vec{a}\; P_t(\vec{a},\vec{x})$$

$$\Delta_t = T(1-e^{-2t})$$



Diffusion models learn how to go backward in time

Generating new images from white noise (denoising)

Time-reversal

Score function provides the force field to go back in time

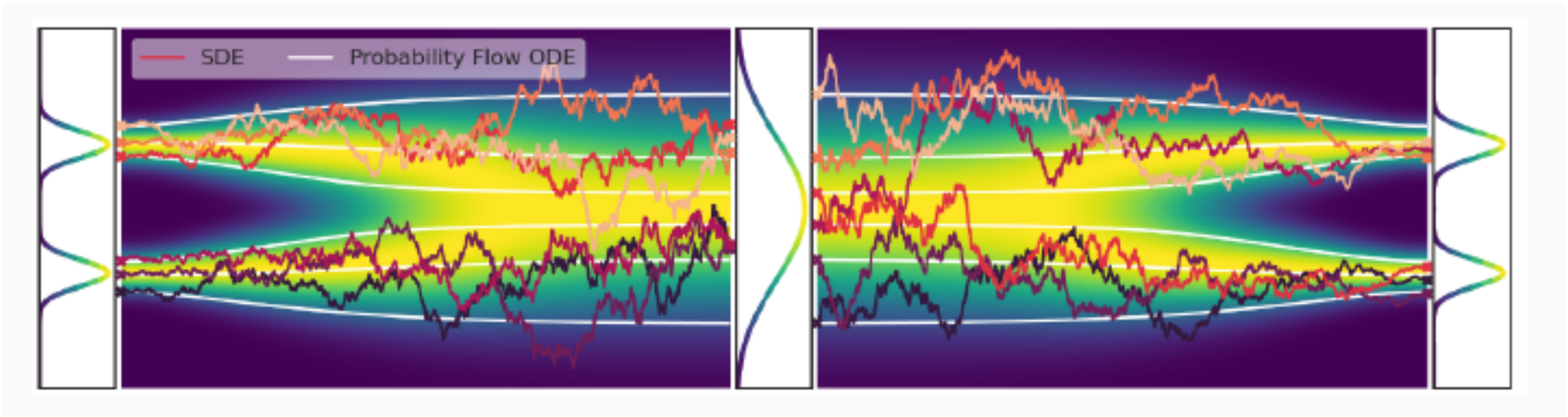
$$\mathcal{F}_i(\vec{x},t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i} \qquad\qquad -\frac{dy_i}{dt} = y_i + 2T\mathcal{F}_i(y,t) + \eta_i(t)$$

Time-reversed Langevin equation transforms iid Gaussians $\mathcal{N}(0,1)$ in new data

$$P_{Gauss}(\vec{x}) \rightarrow P_{data}(\vec{x})$$

Diffusion models learn an estimate of the score from data $\vec{\mathcal{S}}^{\theta}(\vec{x})$

and are able to reverse time (denoise)



Sohl-Dickstein et al 2015 (ideas from out of equilibrium thermodynamics)
Yang & Ermon 2019, Ho et al 2020, …
For a review see Yang et al “Diffusion Models: A Comprehensive Survey of Methods and Applications

Theory: a partial state of the art

- Convergence toward the data distribution (De Bortoli, Doucet et al. (2021, 2022))
- Generalisation and structure of data (GB, Mézard 2023; Raya, Ambrogioni 2023; Kadkhodale, Guth, Simoncelli, Mallat 2023; Schlocchi, Favero, Wyart 2024,…)
- Sample complexity and high-dimensional data (GB, M. Mézard 2023; Cui, Krzakala, Vanden-Eijnden, Zdeborova 2023; Mei 2024,…)
- Sampling and Stochastic Interpolants, Stochastic Localisation (Ghio, Dandi, Krzakala, Zdeborova 2023; Montanari et al. 2023-2024; Bach, Park, Saremi 2023; Bae, Marinari, Ricci-Tersenghi 2024,…)

Two main questions & phenomena

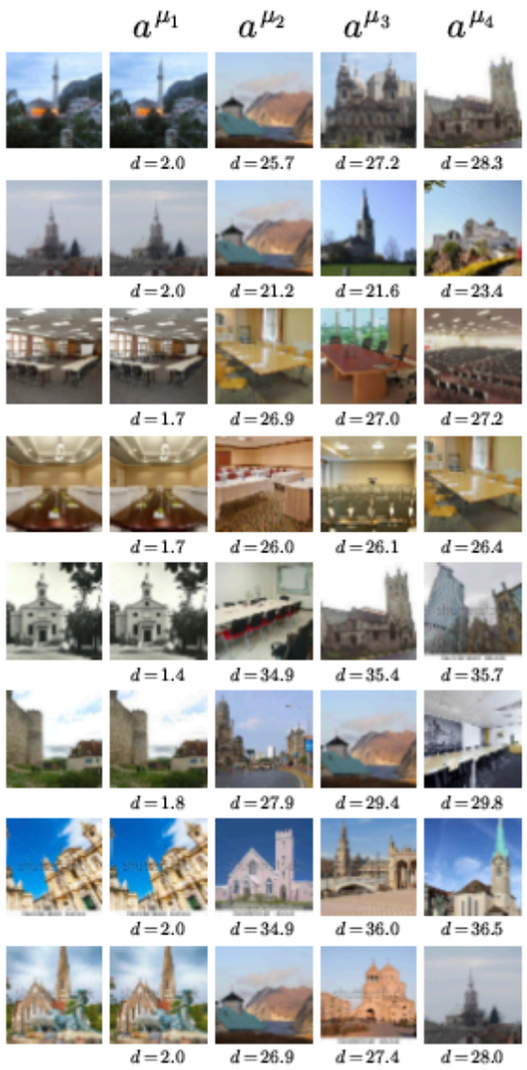
Formation on main classes/features of the data from pure noise
Speciation

?

Connections to structure of data analysis
Symmetry Breaking Phenomenon

Two phenomena

Memorization vs Generalisation
(Collapse transition)



n=200 n=40000

@TBonnaire arxiv 2402.18491
See also Kadkhodale, Guth, Simoncelli, Mallat 2023

Relevant for copyright problems and differential privacy

Glass Transition

Theoretical Study

- Exact solution of simple high-dimensional models: Gaussian mixture and Curie-Weiss model of ferromagnetism (models of two classes)

$$\vec{a} \xrightarrow{p} \mathcal{N}(\vec{v}, I)$$
$$\vec{a} \xrightarrow{1-p} \mathcal{N}(-\vec{v}, I)$$
$$|\vec{v}|^2 = \mu d$$

$$P_0(\vec{a}) = \frac{1}{Z} e^{\frac{\beta J}{d} \sum_{i,j} a_i a_j + \frac{h}{d} \sum_i a_i}$$
$$\beta > \beta_c \qquad a_i = \pm 1$$

- High dimensional limit: the dimension and the number of data are very large, in the analysis they go to infinity.

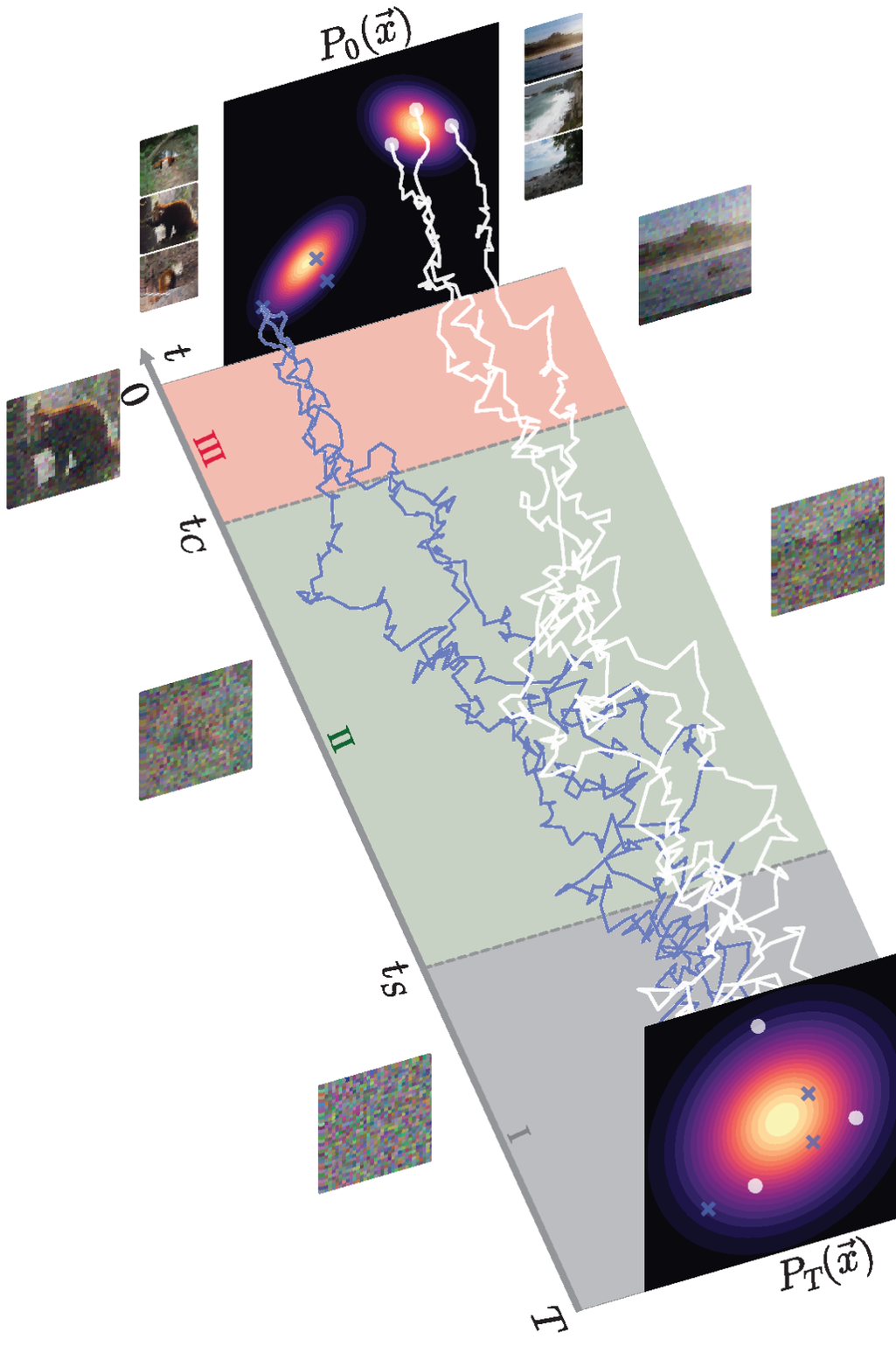
- Exact empirical score hypothesis

$$\mathcal{F}_i(\vec{x}, t) = \frac{\partial \log P_t(\vec{x})}{\partial x_i}$$

Empirical distribution

$$P_t(\vec{x}) = \int d\vec{a}\; P_0^e(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{1}{2}\frac{(\vec{x} - \vec{a}e^{-t})^2}{\Delta_t}\right)$$

Methods: Dynamical Mean-Field Theory, Spin-Glass Theory, Out-of-Equilibrium Dynamics



Three Dynamical Regimes

Two transitions

Speciation and memorization

Regime I and Speciation Transition

Gaussian Mixture

$$\vec{a} \xrightarrow{1/2} \mathcal{N}(\vec{v}, I)$$
$$\vec{a} \xrightarrow{1/2} \mathcal{N}(-\vec{v}, I)$$
$$|\vec{v}|^2 = \mu d$$

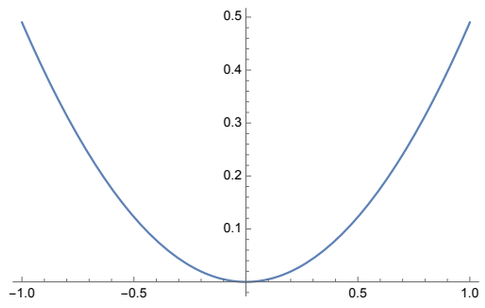
Symmetry breaking of the dynamical trajectories

Beginning of the backward process $\quad \mu(t) = \dfrac{\vec{x}\cdot\vec{v}}{\sqrt{d}} \sim O(1)$

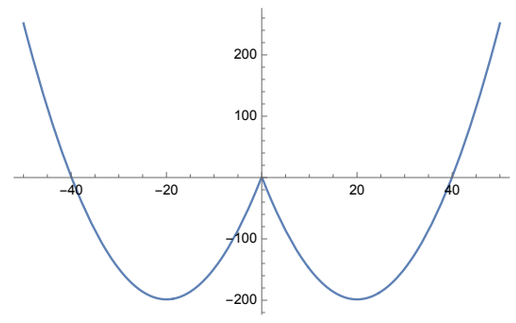
$$\frac{d\mu}{dt} = -\frac{dV}{d\mu} + \eta(t)$$

$$V(q,t) = \frac{1}{2}q^2 - 2\mu^2 \log\cosh\left(qe^{-t}\sqrt{d}\right)$$

$\sqrt{d}e^{-t} \ll 1$



$\sqrt{d}e^{-t} \gg 1$

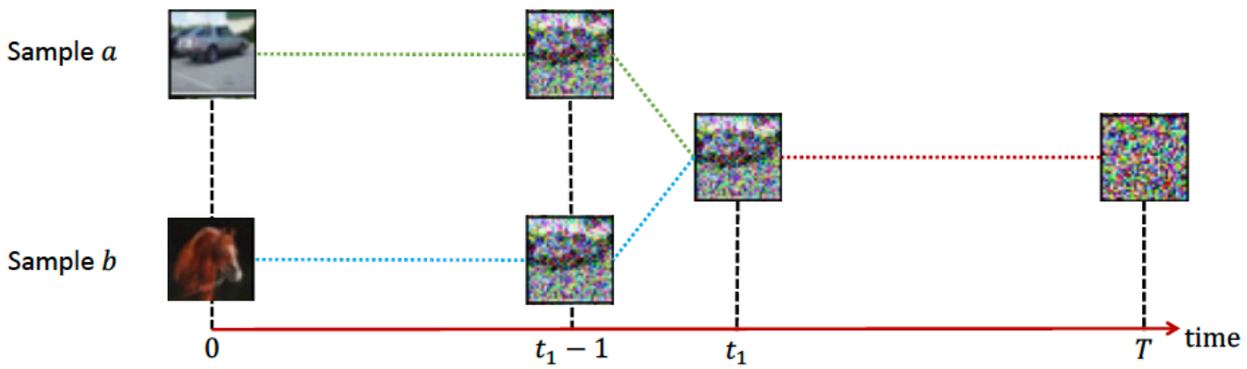


Symmetry breaking $\longrightarrow$ $t_S \simeq \dfrac{1}{2}\log d$

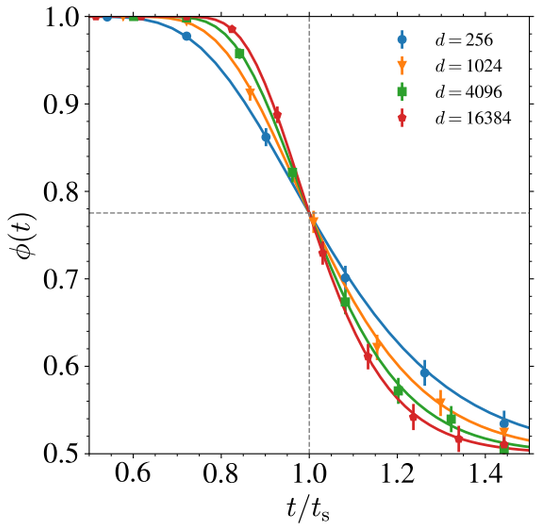
GB, Mézard 2023
Raya and Ambrogioni 2023

Regime I and Speciation Transition

Cloning trajectories



Probability of ending in the same class



$$\phi(t) = f(\sqrt{d}e^{-t})$$

Weights of classes are fixed in regime I, “high-frequency” details later

Extension to more general cases: analysis of time-dependent correlation matrix, and large-time perturbative expansion of the score (“Landau Theory”)

Λ Largest principal component of the correlation matrix of the data

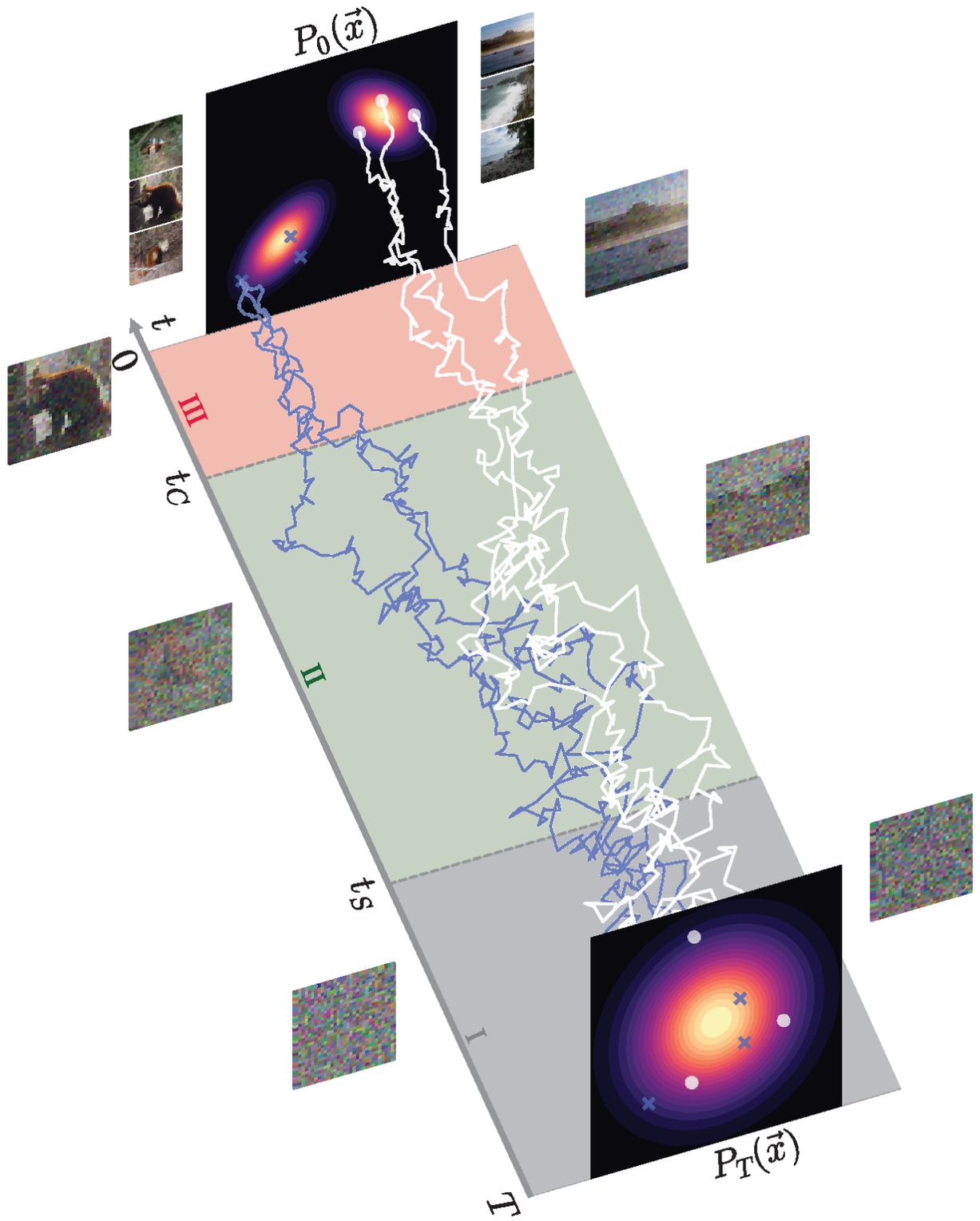
“Speciation” or “Symmetry breaking” of the backward process at $t_S = \frac{1}{2}\log\Lambda$

Bonnaire, De Bortoli, GB, Mézard 2024

If there are hierarchical classes then several transitions

See Schlocchi, Favero, Wyart 2024

Three Dynamical Regimes



Regime II & III and Memorization

In Regimes I and II: $$P_t(\vec{x}) \simeq P_t^{true}(\vec{x}) = \int d\vec{a}\, P_0(\vec{a}) \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}e^{-t})^2}{\Delta_t}\right)$$

In Regimes III: memorisation and collapse on the training set $$P_t(\vec{x}) \neq P_t^{true}(\vec{x})$$

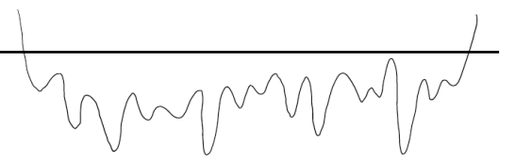
Analysis of the high-d distribution in the large n and d limit at fixed $\alpha = \log(n)/d$

$$P_t(\vec{x}) = \frac{1}{n}\sum_{\mu=1}^{n} \frac{1}{\sqrt{2\pi\Delta_t}^d} \exp\left(-\frac{1}{2}\frac{(\vec{x}-\vec{a}_\mu e^{-t})^2}{\Delta_t}\right) = \sum_{\mu=1}^{n} e^{-E^\mu_{eff}(\vec{x})}$$

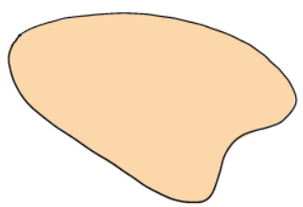
Random Energy Model (Spin-Glass theory): transition from a regime of concentration to the average (generalisation), to a fluctuating regime correlated to the data

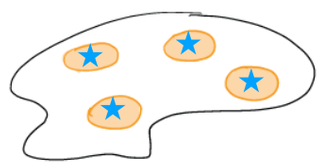
See Hopfield with exp patterns Lucibello Mézard 2023

$t_1 \gg 1$ Memorization: a (Glass) Transition $t_2 \ll 1$

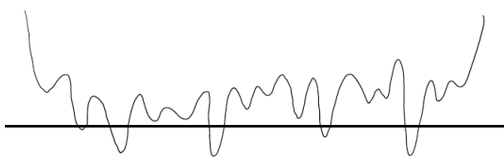


$P_{t_1}(\vec{x})$

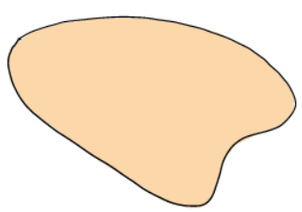


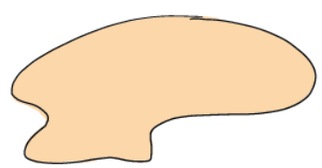


$P_{t_2}(\vec{x})$



$P^{true}_{t_1}(\vec{x})$





$P^{true}_{t_2}(\vec{x})$

Regime II & III and Memorization

Transition: volume argument & direct computation for simple models (eg GMs)

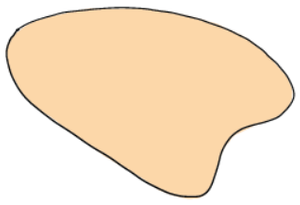
Volumes in high-dimensions are given by the entropies

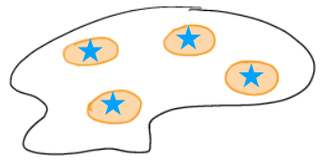
$$S(t) = -\int d\vec{x} P_t(\vec{x}) \log P_t(\vec{x})$$

$$S_{Gauss}(t) = -\int d\vec{x} G(\vec{x}, t) \log G(\vec{x}, t)$$

$$V_{\mathcal{S}} = n e^{S_{Gauss}}$$

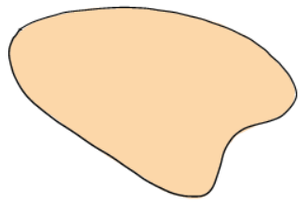
$\mathcal{S}$ $P_{t_1}(\vec{x})$

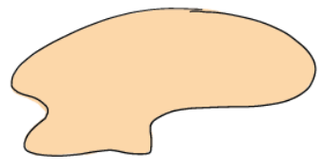




$P_{t_2}(\vec{x})$

$\mathcal{S}_{true}$ $P^{true}_{t_1}(\vec{x})$





$P^{true}_{t_2}(\vec{x})$

$$V_{\mathcal{S}_{true}} = e^{S}$$

Transition at $V_{\mathcal{S}_{true}} = V_{\mathcal{S}}$

Collapse transition at $f(t) = \frac{1}{d}\left(\log(n) + S_{Gauss}(t) - S(t)\right) = 0$

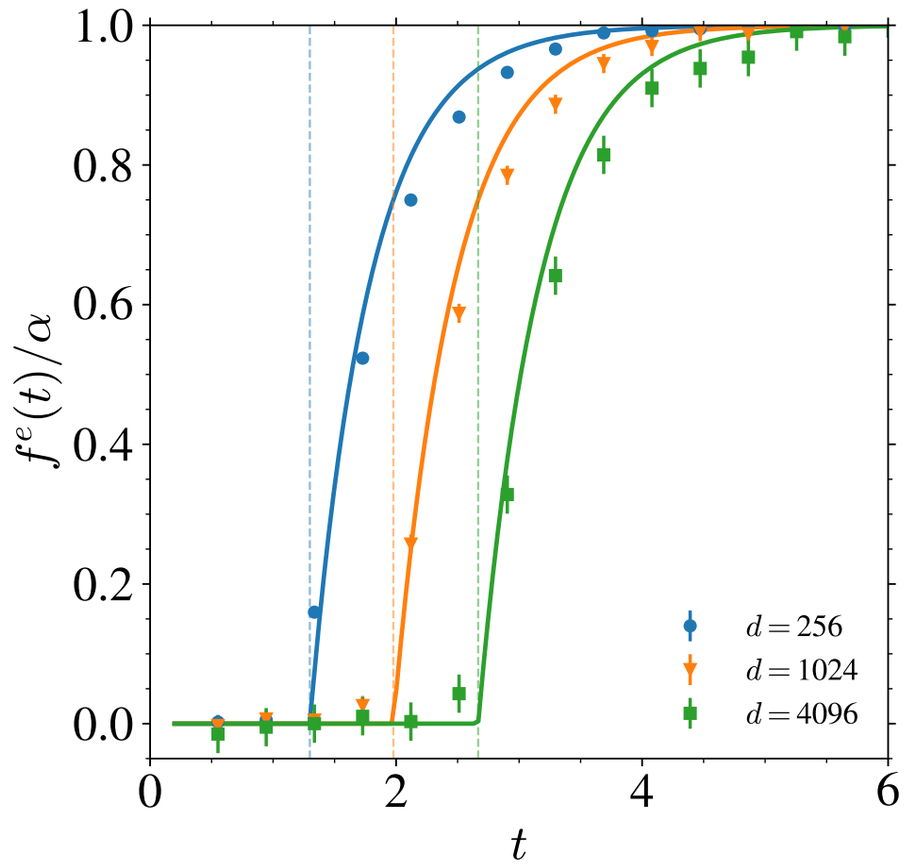
Regime II & III and Memorization

Collapse transition at $f(t) = \frac{1}{d}\left(\log(n) + S_{Gauss}(t) - S(t)\right) = 0$

$S(t) = -\int d\vec{x} P_t(\vec{x}) \log P_t(\vec{x})$ Entropy at time t $S_{Gauss}(t) = -\int d\vec{x} G(\vec{x}, t) \log G(\vec{x}, t)$

Gaussian mixture model:
Theory vs numerics

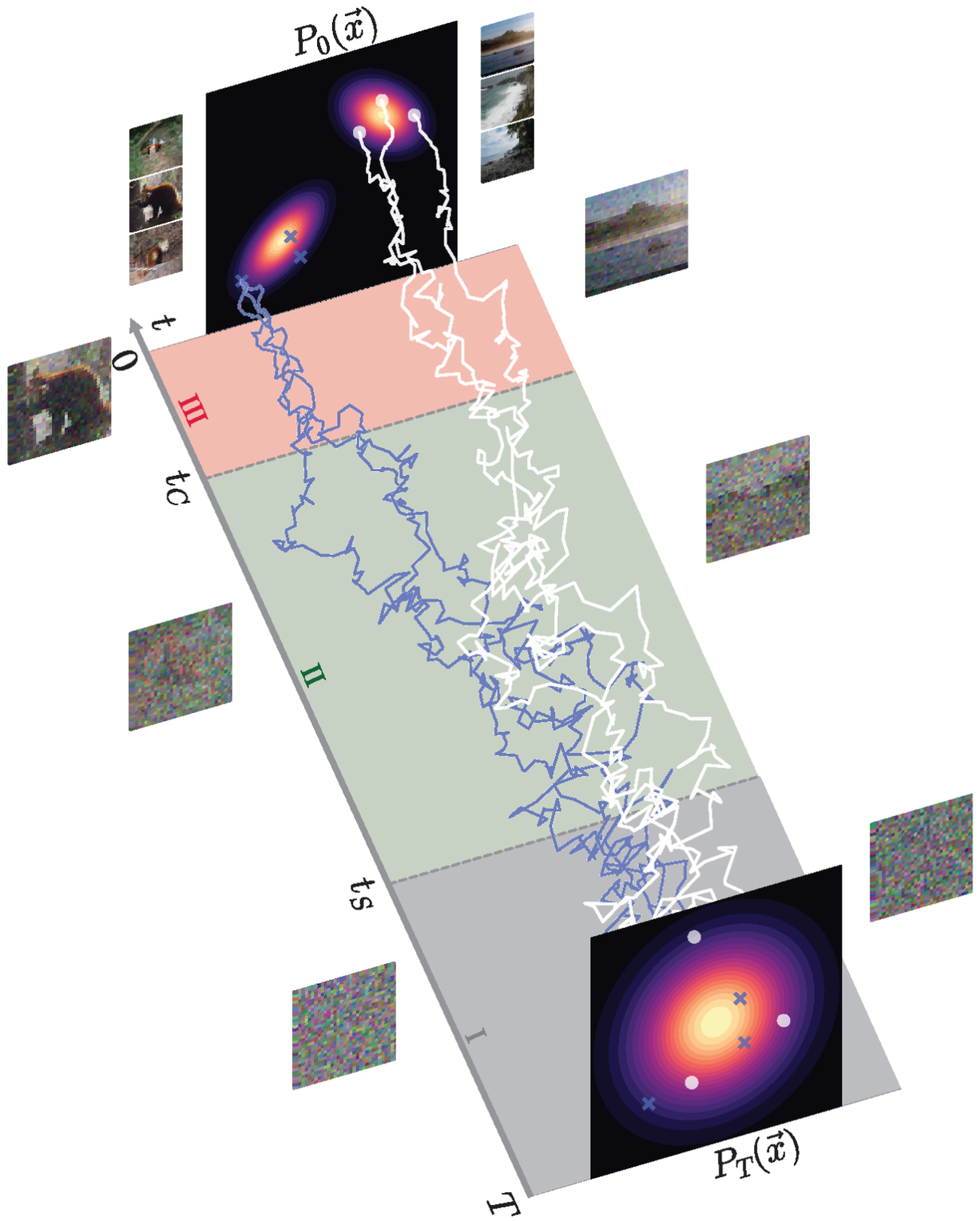
$$t_C = \frac{1}{2} \log\left(1 + \frac{\sigma^2}{n^{2/d} - 1}\right)$$



Without regularisation DMs are cursed -> memorisation

$n = 20000$

Three Dynamical Regimes



Application to Real Images

Model

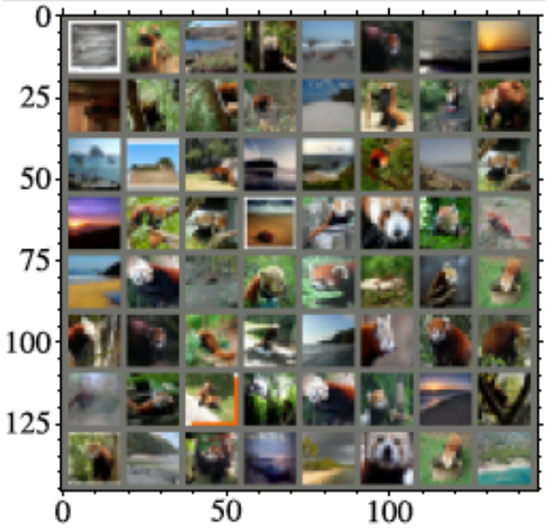
Similar model of Ho et al 2020
U-Net, 4 resolution maps
with 2 convolutional blocks
Dropout rate 0.1
25.7 millions parameters

Training

Adam optimizer
LR 10^{-4}
Multiplied by 0.98 every 50 epochs

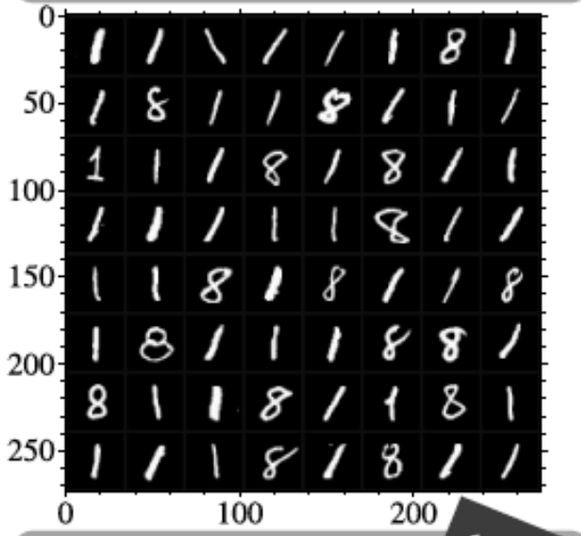
Imagenet16 — 500k steps

- 2000 samples
- L. pandas and seashores
- $N = 16 \times 16 \times 3 = 768$



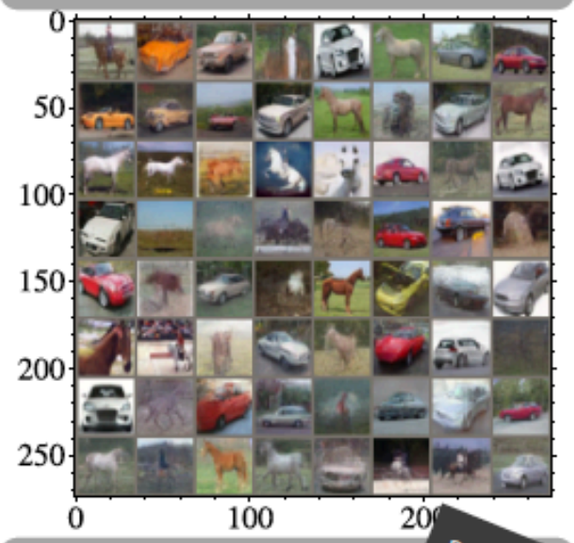
MNIST32 — 100k steps

- 10000 samples
- Classes 1 and 8
- $N = 32 \times 32 \times 1 = 1024$



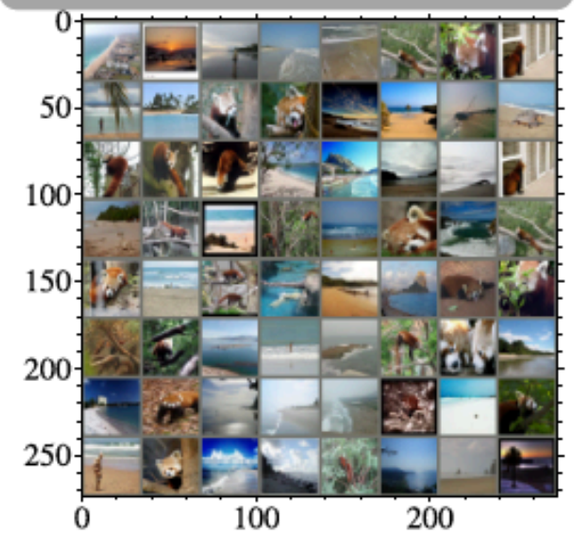
CIFAR2 — 100k steps

- 6000 samples
- Classes horses and cars
- $N = 32 \times 32 \times 3 = 3072$



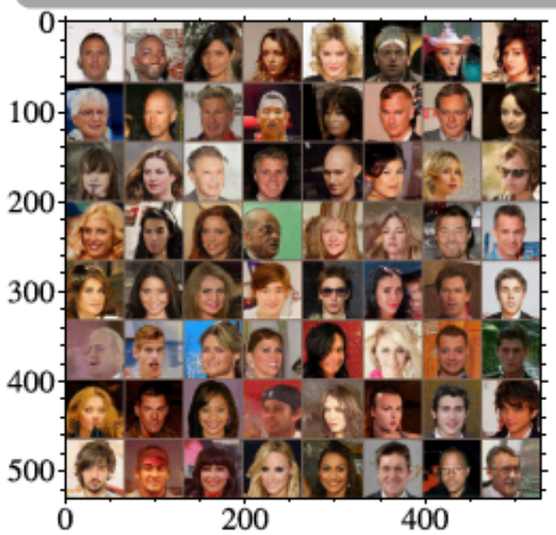
Imagenet32 — 500k steps

- 2000 samples
- L. pandas and seashores
- $N = 32 \times 32 \times 3 = 3072$



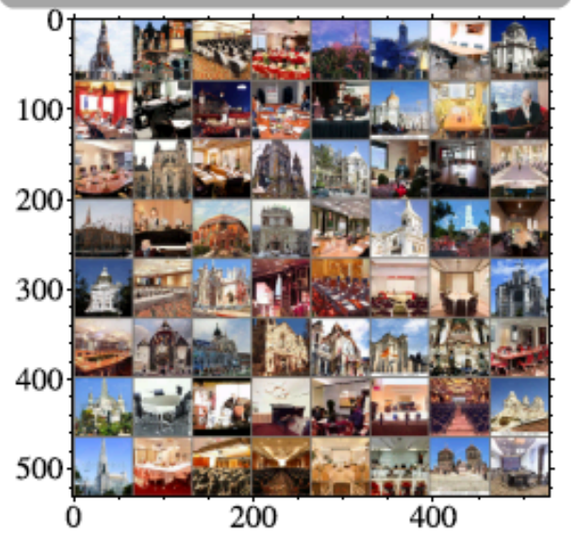
CelebA64 — 130k steps

- 40000 samples
- Classes males and females
- $N = 64 \times 64 \times 3 = 12288$

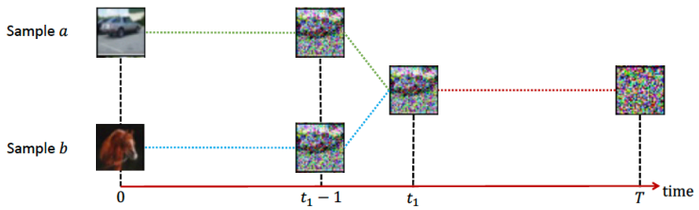


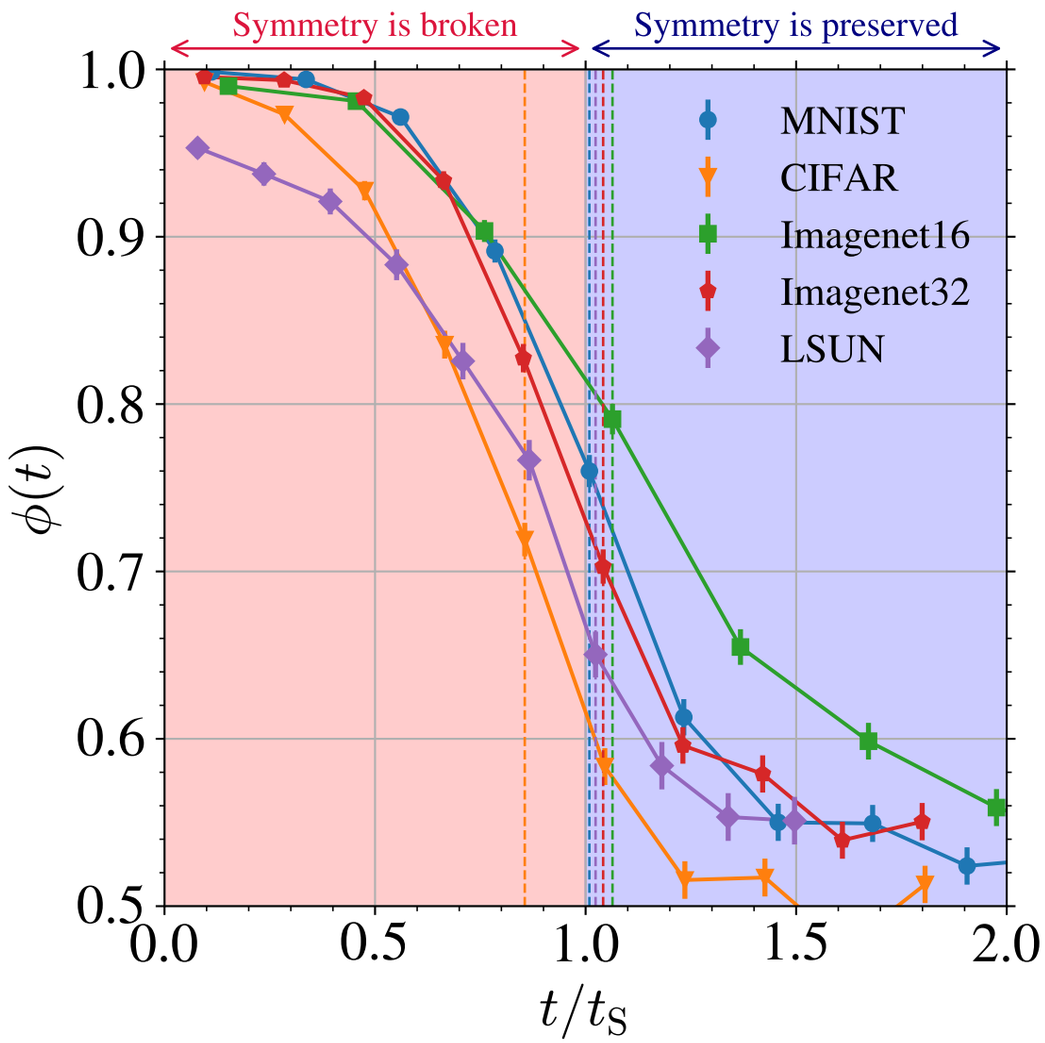
LSUN64 — 310k steps

- 40000 samples
- Conference and churches
- $N = 64 \times 64 \times 3 = 12288$



Speciation Transition in Real Images



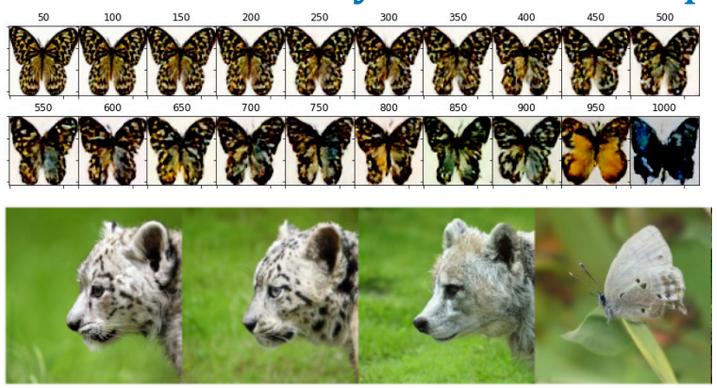


Probability 2 clones in the same class

$$t_S = \frac{1}{2}\log\Lambda$$

Confirm the speciation phenomenon & good estimation of the speciation time

Also observed numerically in U-turns experiments



Behjoo et al 2023
Schlocchi et al 2024

Collapse Transition in Real Images

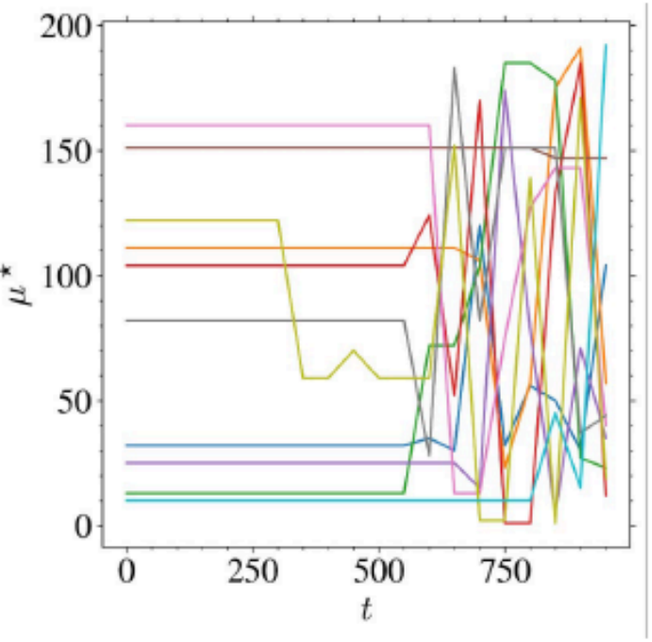
- The model has collapsed: ***all*** the generated images are from the dataset
- To estimate the collapse time, we have a look at two related quantities:

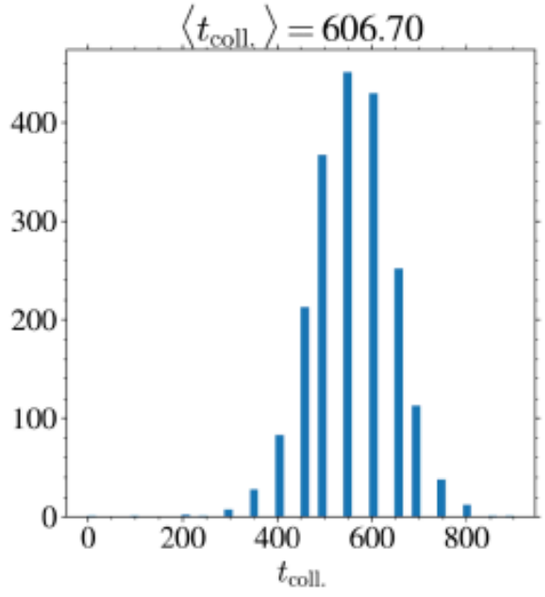
CelebA64 *130k steps*
- 200 samples
- Classes males and females

$$\mu^\star(\tilde{x}) = \text{argmin}_{x_\mu \in X} \left\| x_\mu \sqrt{\bar{\alpha}} - \tilde{x} \right\|_2^2$$

$$d^\star(\tilde{x}) = \min_{x_\mu \in X} \left\| x_\mu \sqrt{\bar{\alpha}} - \tilde{x} \right\|_2^2$$

x_μ: training image
$\tilde{x}$: generated image





→ Estimate of the average collapse time
to be compared
to the theoretical prediction

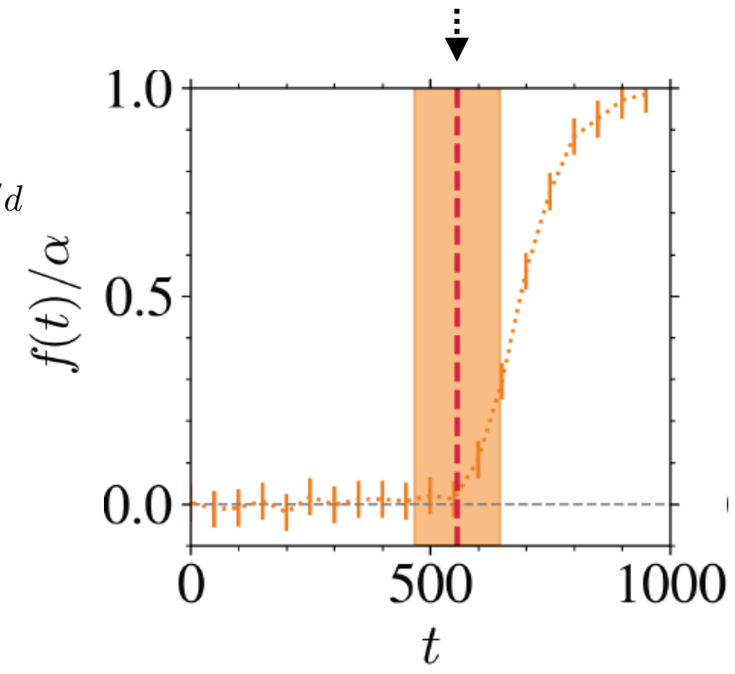
Collapse Transition in Real Images

Theory: collapse transition at

$$f(t) = (\log n)/d + \frac{1}{2} + \frac{1}{2}\log(2\pi T(1 - e^{-2t})) - S(t)/d = 0$$

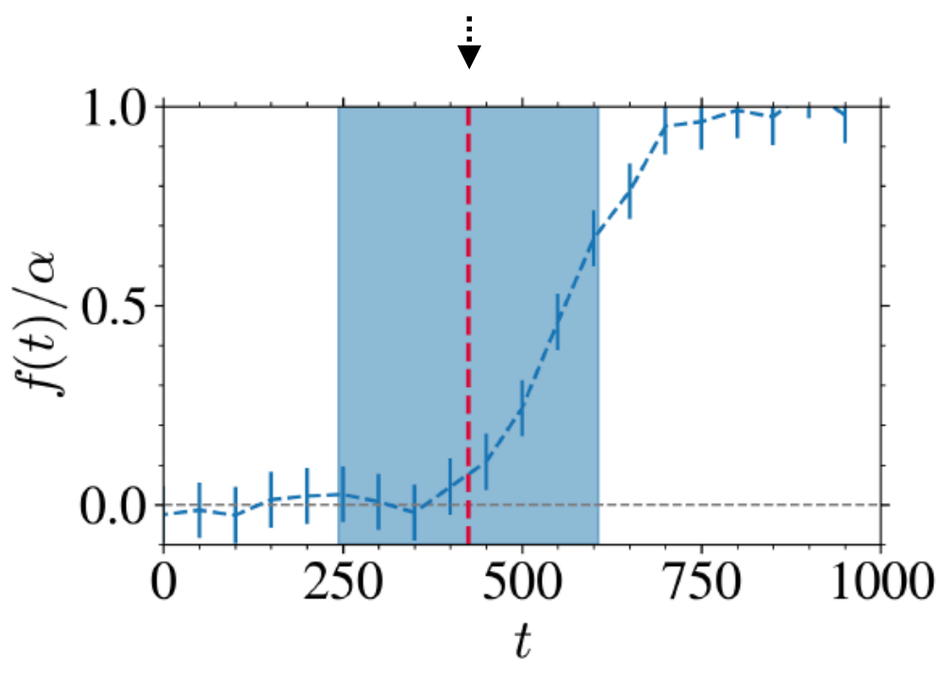
Estimate of the average collapse time

celebA64
n=200
d=12288
$\alpha = (\log n)/d$



Estimate of the average collapse time

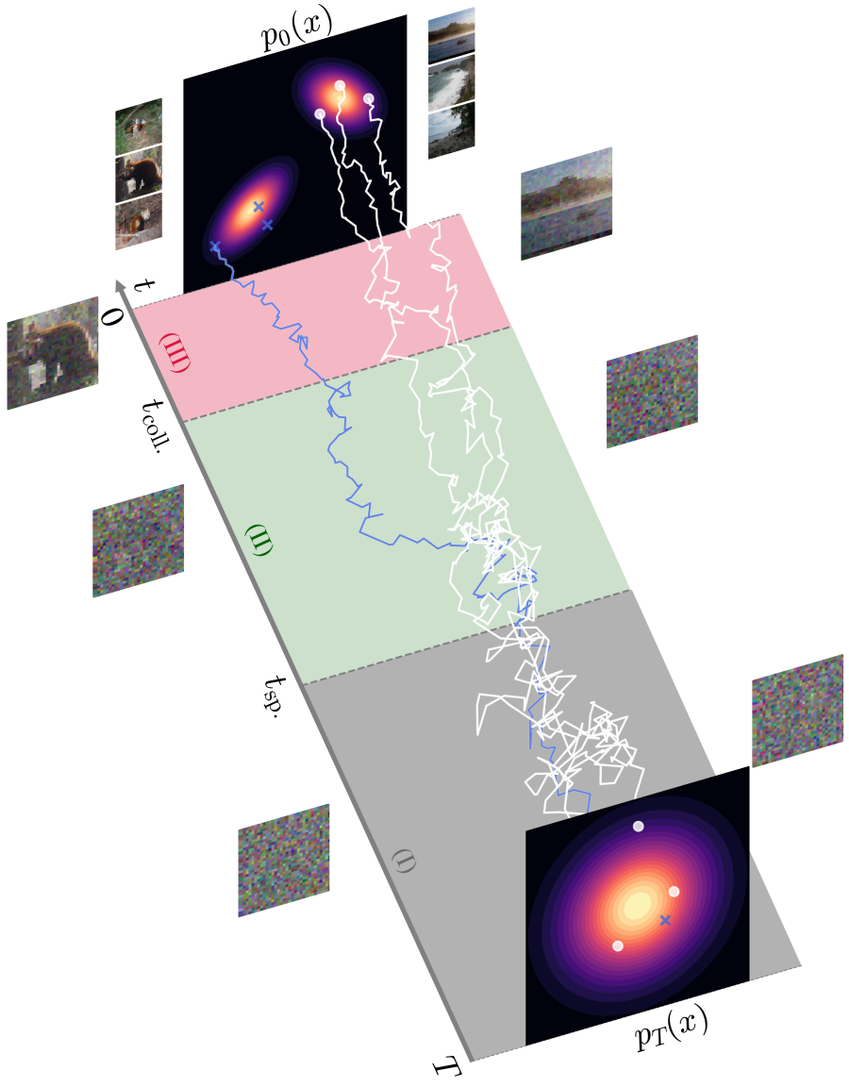
Imagenet 32
n=200
d=3072
$\alpha = (\log n)/d$



Confirm the collapse phenomenon

Good estimation of the collapse time

Conclusion



Three dynamical regimes: theory, criteria & numerical confirmation

- Regime I: where trajectories commit (weights of classes are generated)
- Regime III: where memorisation takes place - general numerical criterion for collapse based on entropy
- Not having an accurate score has different consequences in different time regimes: bad in regime I, but could be good in III

Perspectives & ongoing works:

- Guidelines for applications: time-dependent regularisation, guidance (theory of classifier free guidance),…
- Theory: regularisation, models of score, structure of data and memorisation/generalisation.

GB, Mézard JSTAT 2023; GB, Bonnaire, De Bortoli, Mézard ArXiv 2402.18491 to appear in Nature Commun.;
GB, Mézard ArXiv 2408.05807 (relation with kernel density estimation)