The Generative Dynamics of Diffusion Models in Large Dimensions

Tony Bonnaire

École Normale Supérieure, Physics Department & Centre for Data Sciences, Paris

G. Biroli

V. De Bortoli

M. Mézard

Scan or click the QR to check the paper!

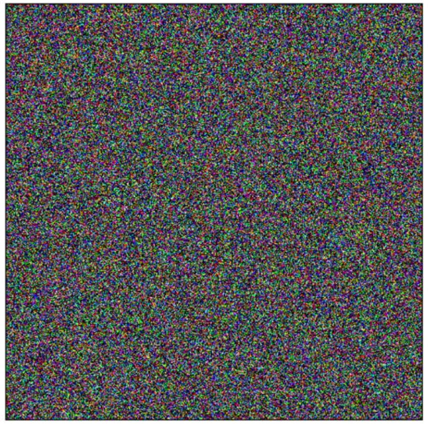
ENS | PSL

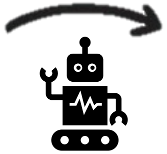
Physics Informed Machine Learning Workshop, Los Alamos, October, 18th

Talk based on arXiv:2402.18491

Generative models: a brief overview

- Goal: model the probability distribution of the data $P_0(\boldsymbol{a}), \boldsymbol{a} \in \mathbb{R}^d$
- **Sampling task**: usually relies on learning a mapping from a simple distribution to $P_0(\boldsymbol{a})$ based on *finite* training set of size n





- Several successful paths include:
  - Variational Autoencoders (VAEs) [Kingma+2013]
  - Generative Adversarial Networks (GANs) [Goodfellow+2014]
  - Normalizing flows [Tabak+2010, Rezende+2015]
  - Diffusion Models (DMs) [Sohl-Dickstein+2015, Ho+2020]

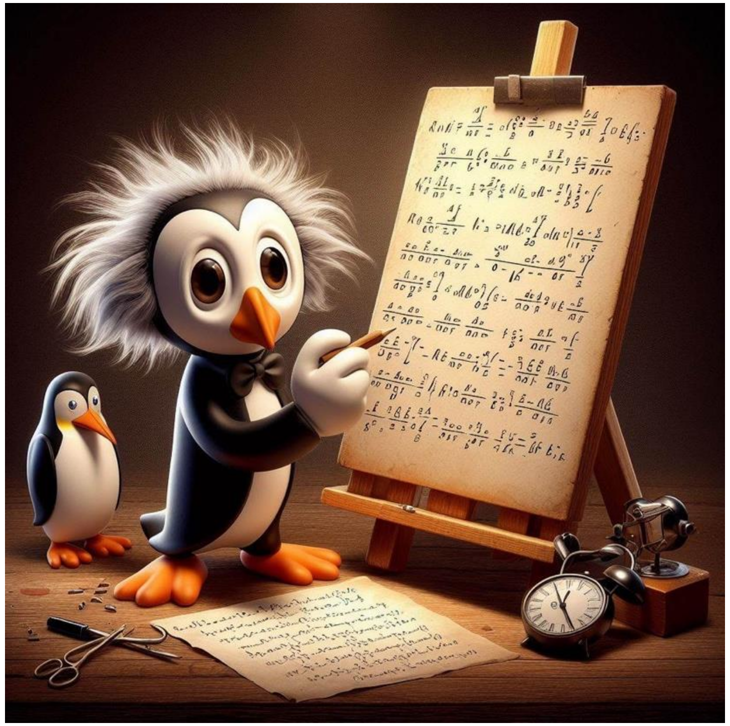
Some examples: DALL-E

"Realistic silhouette of a horse running at sunset time with a vibrant sky"

Obtained with DALL-E 3

Some examples: DALL-E

“Realistic silhouette of a horse running at sunset time with a vibrant sky”

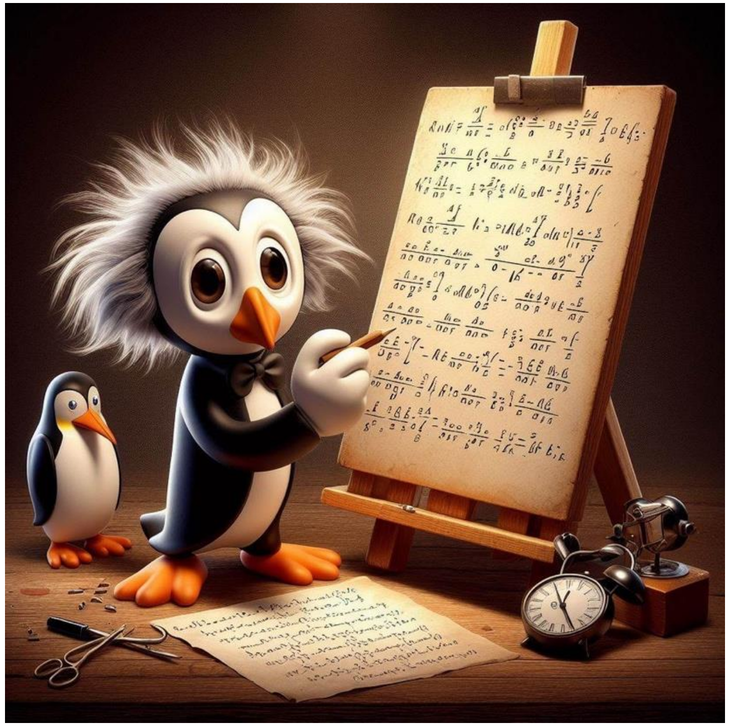


“A penguin writing down Einstein’s equations”

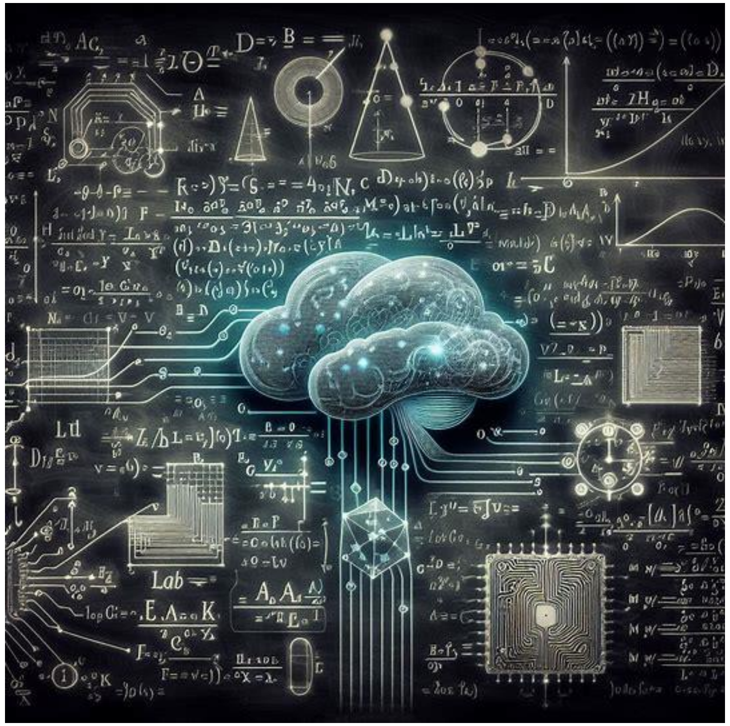
Obtained with DALL-E 3

Some examples: DALL-E

"Realistic silhouette of a horse running at sunset time with a vibrant sky"



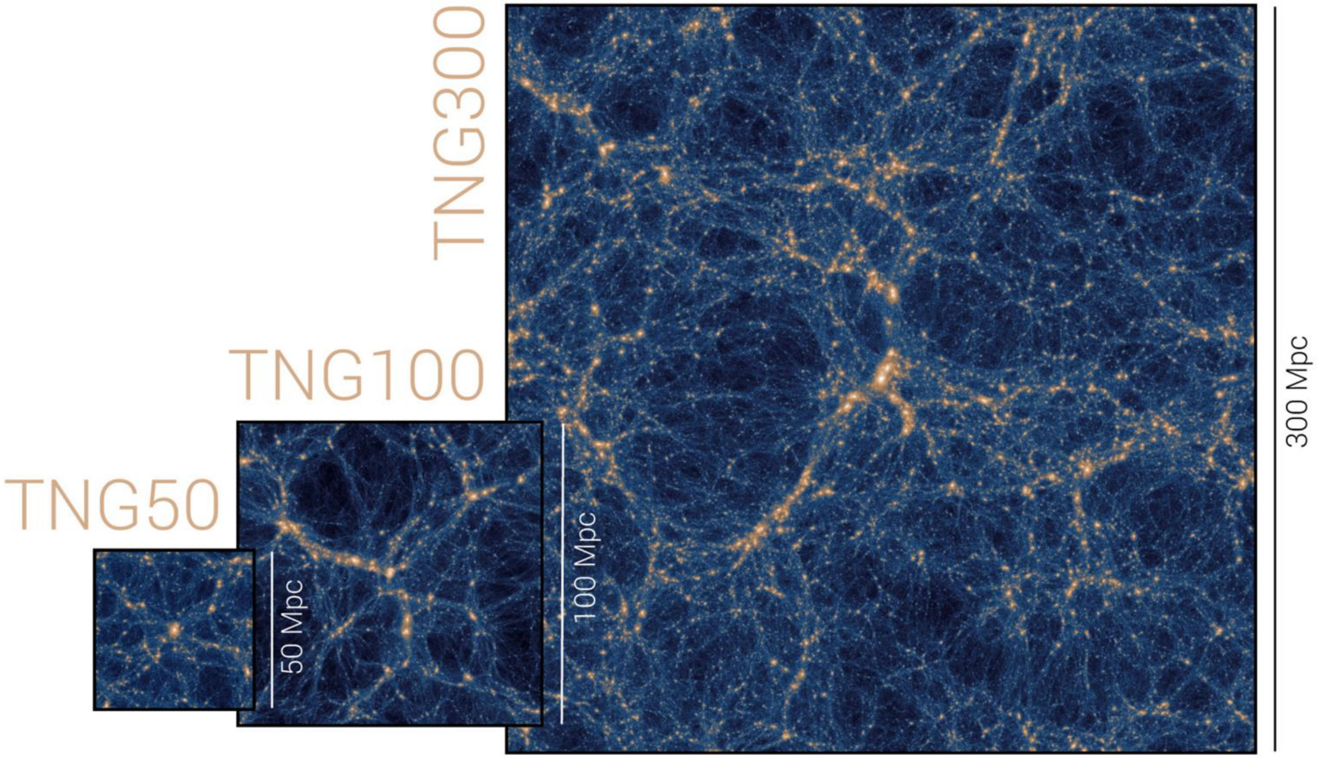
"A penguin writing down Einstein's equations"



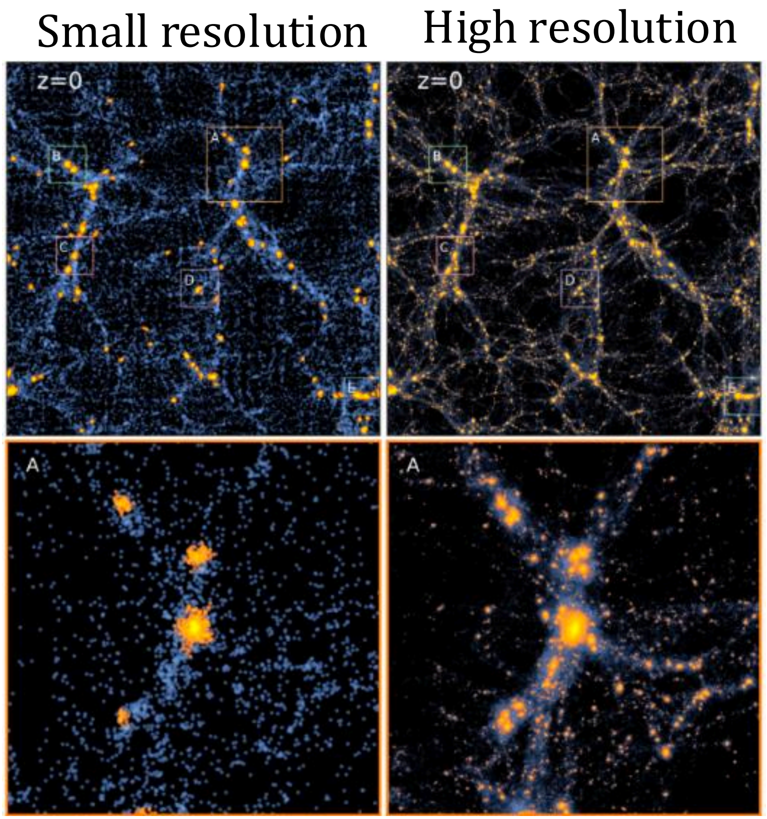
"Physics and Machine Learning"

Obtained with DALL-E 3

What about science?



Cosmological matter fields obtained from simulations [Nelson+2018]



Small-to-high resolution mapping using generative AI [Li+2021]

In science: Realistic data generation (fields, molecules, etc.) , Super-resolution, Test hypothesis

1 *forward process* 2 *backward process*

- The idea is to progressively degrade an initial datapoint $\boldsymbol{a}^\mu$ using an **Ornstein-Uhlenbeck** stochastic process

$$\mathrm{d}\boldsymbol{x} = -\boldsymbol{x}(t)\mathrm{d}t + \boldsymbol{\xi}(t)\mathrm{d}t$$

with $\xi_i(t) \sim \mathcal{N}(0,1)$, $\boldsymbol{x}(0) = \boldsymbol{a}^\mu$

- Using Ito's formula, one can express

$$\boldsymbol{x}(t) = e^{-t}\boldsymbol{a}^\mu + \sqrt{\underbrace{1 - e^{-2t}}_{\Delta_t}}\,\boldsymbol{\xi}(t), \qquad \boldsymbol{x}(0) = \boldsymbol{a}^\mu, \mu \in \{1, \dots, n\}$$

Diffusion Models: Forward process

1 *forward process* 2 *backward process*

- The idea is to progressively degrade an initial datapoint $\boldsymbol{a}^{\mu}$ using an **Ornstein-Uhlenbeck** stochastic process

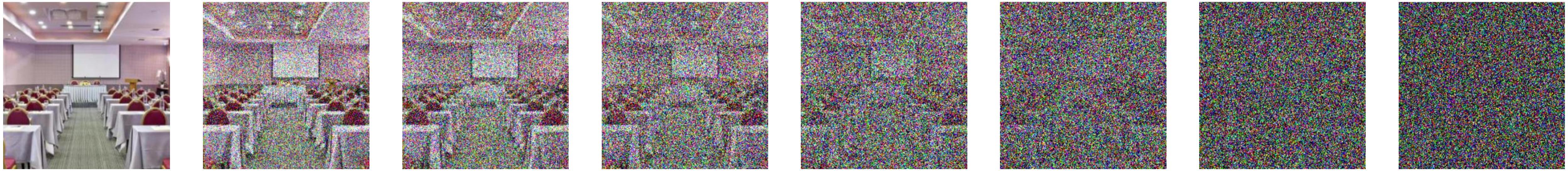
$$\mathrm{d}\boldsymbol{x} = -\boldsymbol{x}(t)\mathrm{d}t + \boldsymbol{\xi}(t)\mathrm{d}t$$

with $\xi_i(t) \sim \mathcal{N}(0,1)$, $\boldsymbol{x}(0) = \boldsymbol{a}^{\mu}$

- Using Ito's formula, one can express

$$\boldsymbol{x}(t) = e^{-t}\boldsymbol{a}^{\mu} + \sqrt{\underbrace{1 - e^{-2t}}_{\Delta_t}}\,\boldsymbol{\xi}(t), \qquad \boldsymbol{x}(0) = \boldsymbol{a}^{\mu}, \mu \in \{1, \dots, n\}$$

$t = 0.00$ $t = 0.05$ $t = 0.10$ $t = 0.20$ $t = 0.40$ $t = 0.80$ $t = 1.60$ $t = 3.20$



$\boldsymbol{a}^{\mu}$ $\mathcal{N}(\mathbf{0}, \mathbf{1})$

Forward time

Diffusion Models: Backward process

1 *forward process* 2 ***backward process***

- In the backward process, one wants to reverse the process from $\mathcal{N}(0,1)$ to $P_0(\boldsymbol{a})$
- To do so [Andersen1983], the force needed to go back is called the **score function** $F(\boldsymbol{y}, t) = \nabla \log P_t(\boldsymbol{y})$

$$\mathrm{d}\boldsymbol{y} = -[\boldsymbol{y} + 2\nabla \log P_t(\boldsymbol{y})]\mathrm{dt} + \boldsymbol{\xi}(t)\mathrm{d}t,$$

where again $\xi_i(t) \sim \mathcal{N}(0,1)$, t runs backward in time, and $\boldsymbol{y}^{(0)} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{1})$

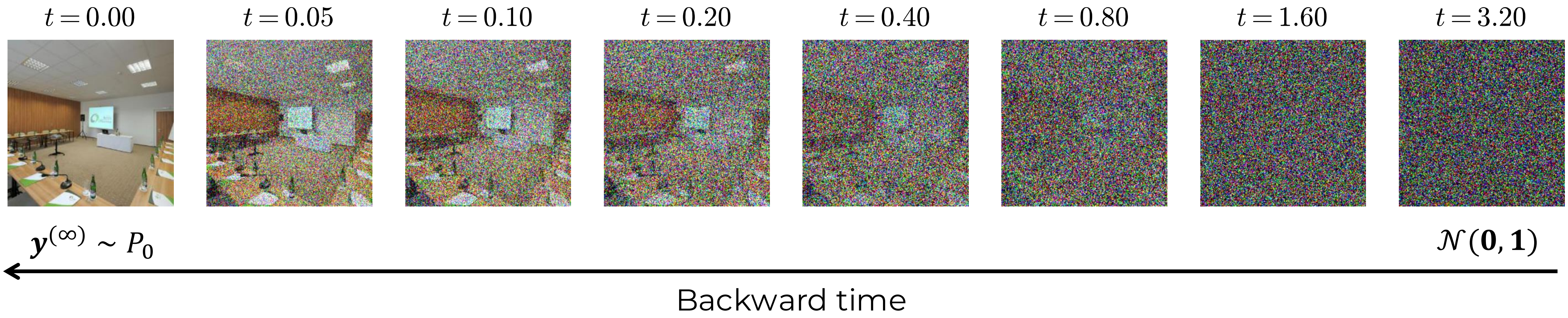
Diffusion Models: Backward process

1 *forward process* 2 ***backward process***

- In the backward process, one wants to reverse the process from $\mathcal{N}(0,1)$ to $P_0(\boldsymbol{a})$
- To do so [Andersen1983], the force needed to go back is called the **score function** $F(\boldsymbol{y}, t) = \nabla \log P_t(\boldsymbol{y})$

$$\mathrm{d}\boldsymbol{y} = -[\boldsymbol{y} + 2\nabla \log P_t(\boldsymbol{y})]\mathrm{d}t + \boldsymbol{\xi}(t)\mathrm{d}t,$$

where again $\xi_i(t) \sim \mathcal{N}(0,1)$, t runs backward in time, and $\boldsymbol{y}^{(0)} = \mathcal{N}(\mathbf{0}, \mathbf{1})$



1 *forward process* 2 ***backward process***

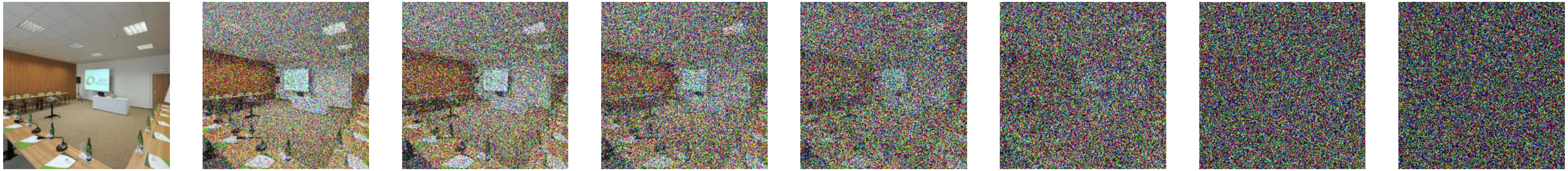
- In the backward process, one wants to reverse the process from $\mathcal{N}(0,1)$ to $P_0(\boldsymbol{a})$
- To do so [Andersen1983], the force needed to go back is called the **score function** $F(\boldsymbol{y},t) = \nabla \log P_t(\boldsymbol{y})$

$$\mathrm{d}\boldsymbol{y} = -[\boldsymbol{y} + 2\nabla \log P_t(\boldsymbol{y})]\mathrm{dt} + \boldsymbol{\xi}(t)\mathrm{d}t,$$

where again $\xi_i(t) \sim \mathcal{N}(0,1)$, t runs backward in time, and $\boldsymbol{y}^{(0)} = \mathcal{N}(\mathbf{0},\mathbf{1})$

- **Practical problem:** the score function needs to be known (and it is hard) → Use of deep networks to learn it

$t=0.00$ $t=0.05$ $t=0.10$ $t=0.20$ $t=0.40$ $t=0.80$ $t=1.60$ $t=3.20$

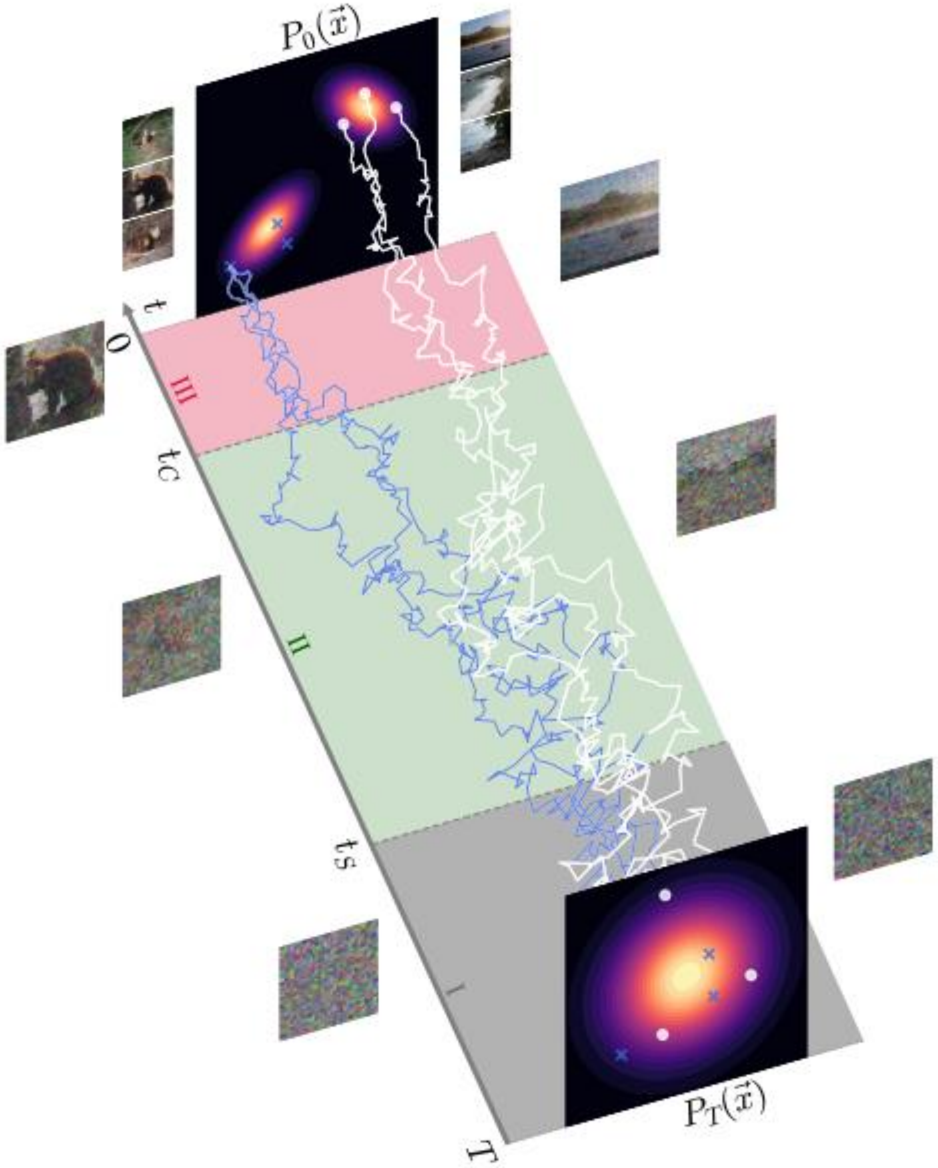


$\boldsymbol{y}^{(\infty)} \sim P_0$ $\mathcal{N}(\mathbf{0},\mathbf{1})$

Backward time

SETTINGS

- High-dimensional: $d \to +\infty$
- Large number of data: $n \to +\infty$
- Exact empirical score function hypothesis



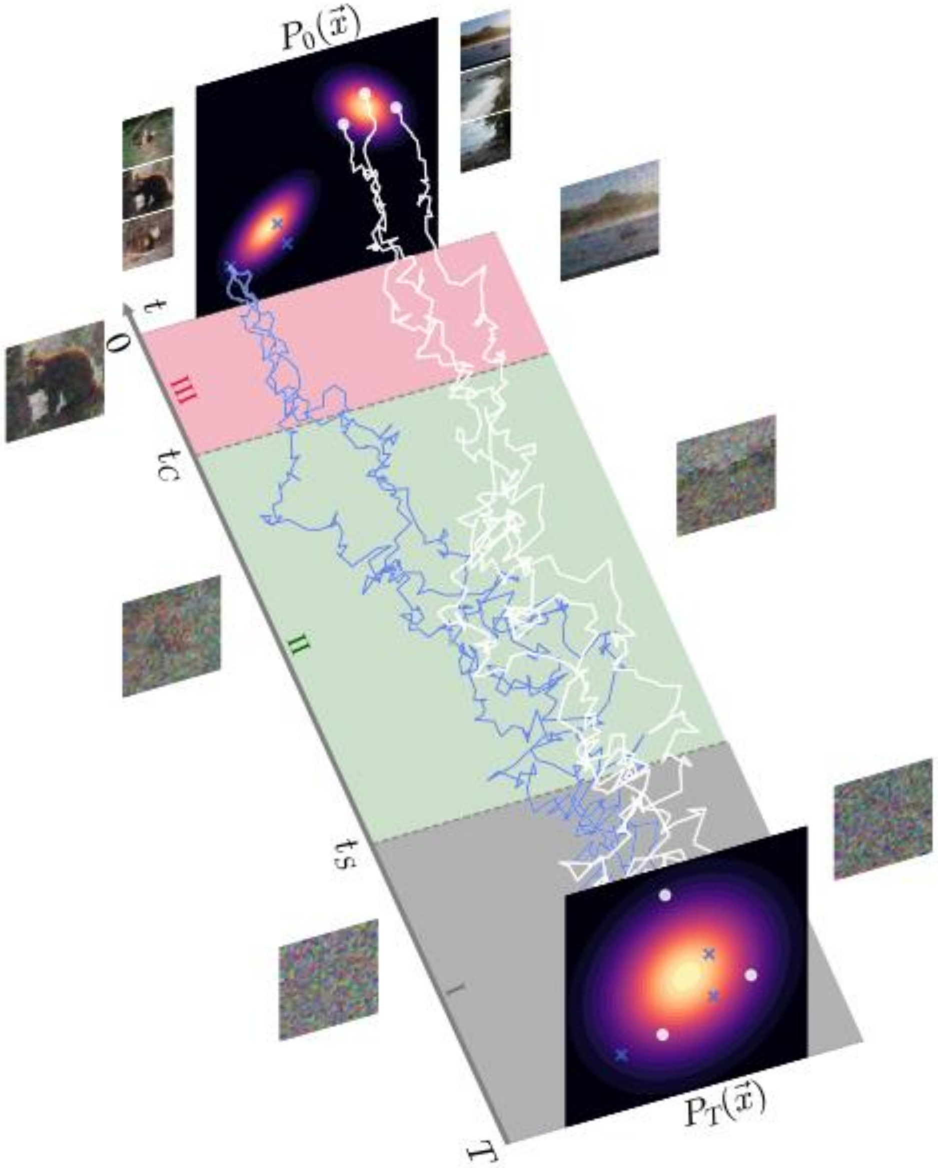
Overview of the results

SETTINGS

- High-dimensional: $d \to +\infty$
- Large number of data: $n \to +\infty$
- Exact empirical score function hypothesis

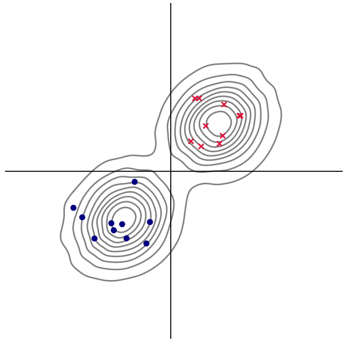
RESULTS

- Three dynamical regimes in the backward dynamics:
  - I. Random motion
  - II. Features formation
  - III. Memorization
- Characterize the timescale at which the transitions between regimes I-II and II-III occur, respectively denoted t_S and t_C



- Assume $\boldsymbol{a}$ is drawn from a (high d) Gaussian mixture model such that

$$\boldsymbol{a}^{\mu} \sim P_0 = \frac{1}{2}\mathcal{N}(\boldsymbol{m}, \boldsymbol{1}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{m}, \boldsymbol{1})$$



- At any time t in the backward (and forward) process,

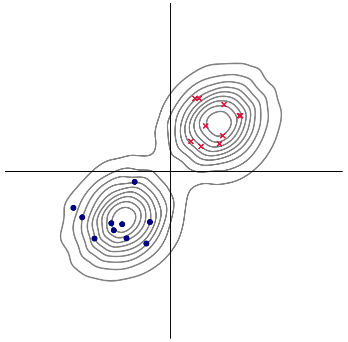
$$P_t(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{a}\, P_0(\boldsymbol{a})\gamma_t(\boldsymbol{x}, \boldsymbol{a}),$$

$$\|\boldsymbol{m}\|_2^2 = m^2 d \text{ where } m \text{ is } O(1)$$

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

with

$$\gamma_t(\boldsymbol{x}, \boldsymbol{a}) = \frac{1}{\sqrt{2\pi\Delta_t}^{\,d}} e^{-\frac{\left(\boldsymbol{x}(t) - \boldsymbol{a}e^{-t}\right)^2}{2\Delta_t}}$$



- Assume $\boldsymbol{a}$ is drawn from a (high d) Gaussian mixture model such that

$$\boldsymbol{a}^{\mu} \sim P_0 = \frac{1}{2}\mathcal{N}(\boldsymbol{m}, \mathbf{1}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{m}, \mathbf{1})$$

- At any time t in the backward (and forward) process,

$$P_t(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{a}\, P_0(\boldsymbol{a})\gamma_t(\boldsymbol{x}, \boldsymbol{a}),$$

with

$$\gamma_t(\boldsymbol{x}, \boldsymbol{a}) = \frac{1}{\sqrt{2\pi\Delta_t}^{d}} e^{-\frac{(\boldsymbol{x}(t) - \boldsymbol{a}e^{-t})^2}{2\Delta_t}}$$

$$\|\boldsymbol{m}\|_2^2 = m^2 d \text{ where } m \text{ is } O(1)$$

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

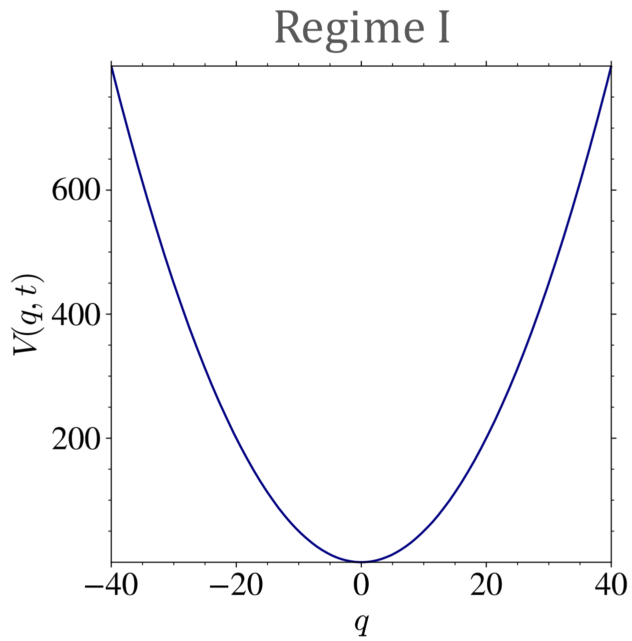
All that matters in this case is the overlap between $\boldsymbol{x}$ and $\pm\boldsymbol{m}$, $q(t) = \frac{1}{\sqrt{d}}\boldsymbol{x}(t) \cdot \boldsymbol{m}$, evolving through

$$-\mathrm{d}q = -\frac{\partial V(q,t)}{\partial q}\mathrm{d}t + \mathrm{d}\xi(t)$$

with

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

Speciation transition in GMM



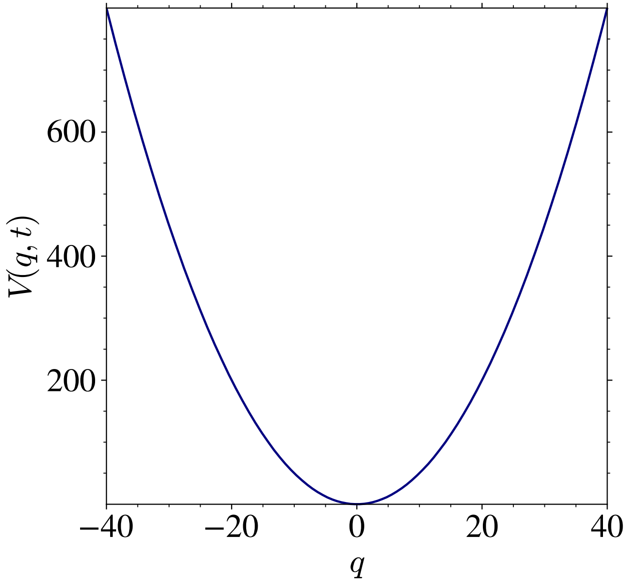
$$t \gg \frac{1}{2}\log d$$

$$-\mathrm{d}q = -\frac{\partial V(q,t)}{\partial q}\mathrm{d}t + \mathrm{d}\xi(t)$$

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

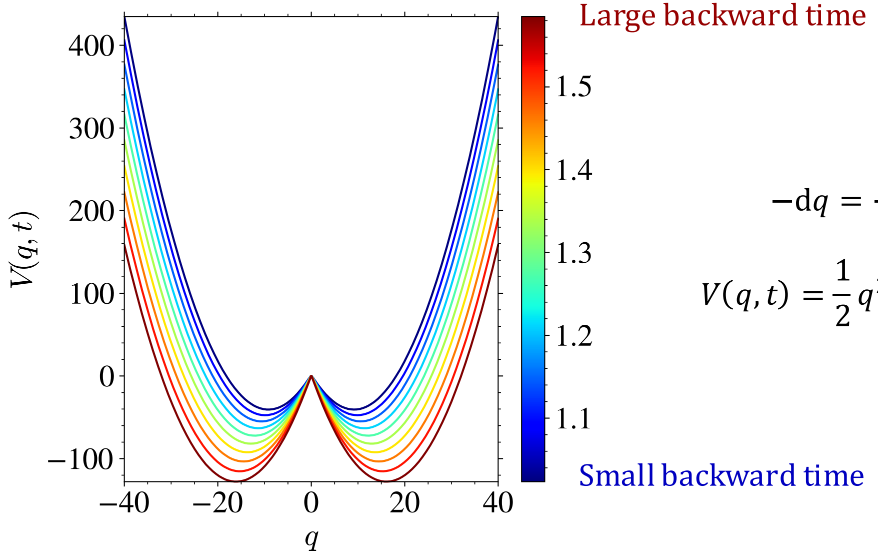
Speciation transition in GMM

Regime I



$$t \gg \frac{1}{2}\log d$$

Regime II

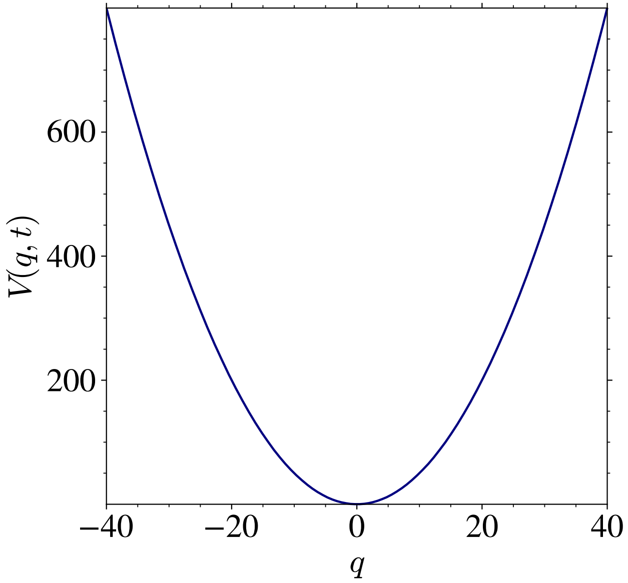


$$t \ll \frac{1}{2}\log d$$

$$-\mathrm{d}q = -\frac{\partial V(q,t)}{\partial q}\mathrm{d}t + \mathrm{d}\xi(t)$$

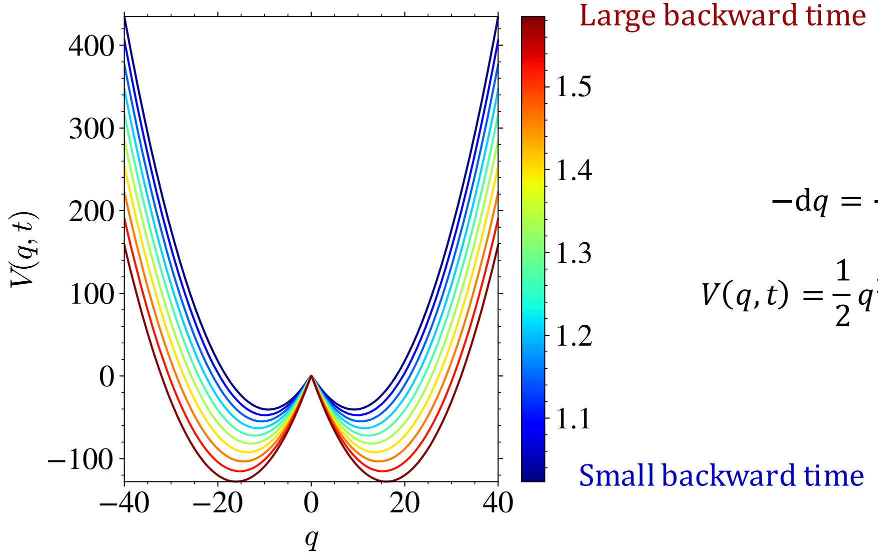
$$V(q,t) = \frac{1}{2}q^2 - 2\mu^2 \log\cosh\left(q e^{-t}\sqrt{d}\right)$$

Regime I



$$t \gg \frac{1}{2}\log d$$

Regime II



$$t \ll \frac{1}{2}\log d$$

$$-\mathrm{d}q = -\frac{\partial V(q,t)}{\partial q}\mathrm{d}t + \mathrm{d}\xi(t)$$

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

- The transition **from single to double well structure of $V(q,t)$ characterises** the first transition between a regime where the trajectory is essentially noise to a regime where the cluster has been decided

It is a transition we dubbed ***speciation*** in reference to ecology, and occurring on a timescale

$$t_S = \frac{1}{2}\log d\,.$$

Regime II and generalisation

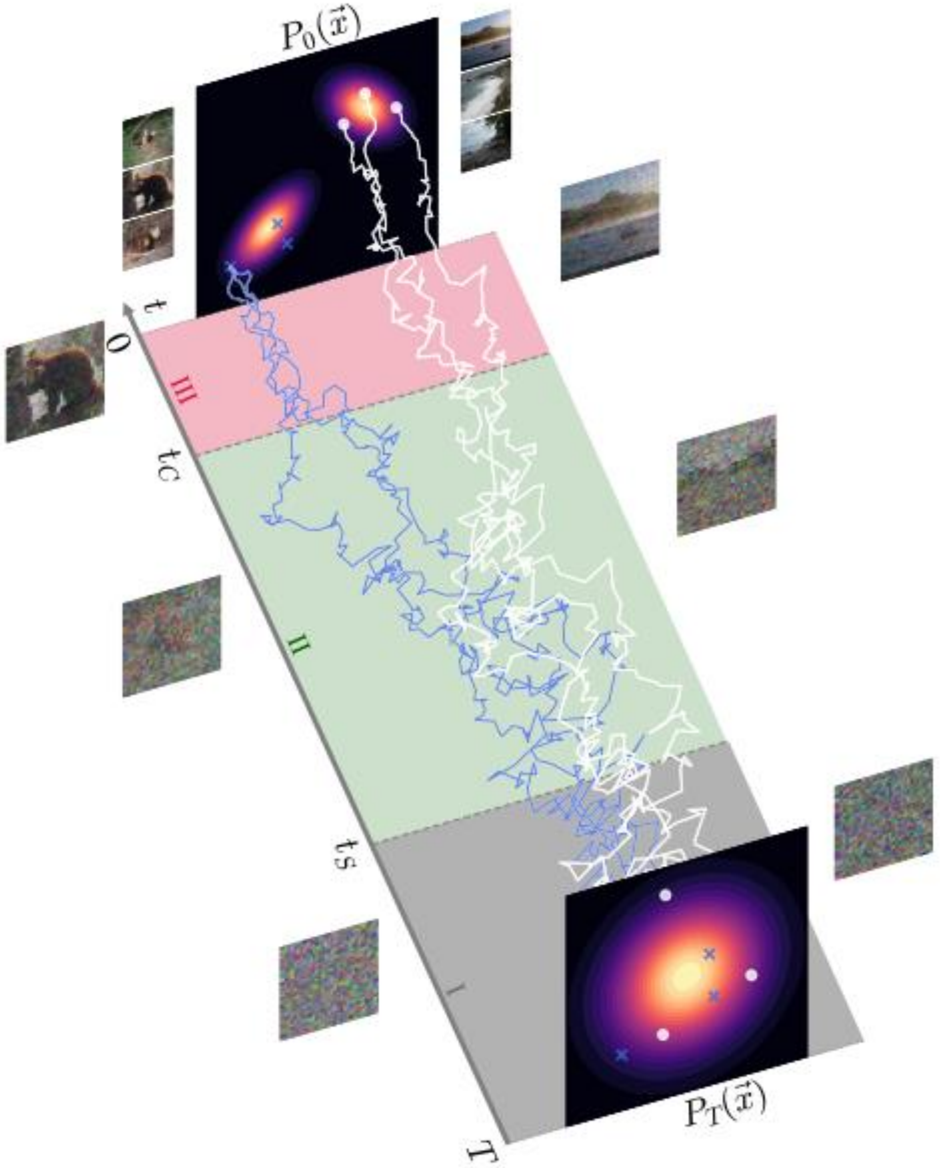
- Regime I is therefore characterised by generating pure noise fror quadratic potential

- In Regime II (i.e. when $t < t_S$), $q = \frac{\boldsymbol{x}\cdot\boldsymbol{m}}{\sqrt{d}}$ diverges to $\pm\infty$ with a sig that depends on the cluster

- The backward process is therefore the one of a single Gaussia centred on $\pm\boldsymbol{m}$

$$-\mathrm{d}\boldsymbol{x} = (-\boldsymbol{x} \pm \boldsymbol{m}e^{-t})\mathrm{d}t + d\boldsymbol{\eta}(t)$$

- In this regime, the trajectories following this equation wi generate a Gaussian $\pm\boldsymbol{m}$, independent of the training set, meanin that the backward dynamics generalises



Regime III and collapse

- In Regimes I and II, $P_t^e(\boldsymbol{x}) \approx P_t^{\mathrm{true}}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{a} P_0(\boldsymbol{a}) \gamma_t(\boldsymbol{x}, \boldsymbol{a})$
- This is no longer true in Regime III where the dynamics get attracted by one of the training point

Regime III and collapse

- In Regimes I and II, $P_t^e(\boldsymbol{x}) \approx P_t^{\text{true}}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{a} P_0(\boldsymbol{a}) \gamma_t(\boldsymbol{x}, \boldsymbol{a})$
- This is no longer true in Regime III where the dynamics get attracted by one of the training point
- Consider a noisy sample $\boldsymbol{x}$ obtained from $\boldsymbol{a}^1$. The empirical probability is hence

$$P_t^e(\boldsymbol{x}) \propto \left[e^{-\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{a}^1 e^{-t}\|^2}{2\Delta_t}} + \sum_{\mu=2}^{n} e^{E_{\text{eff}}^{\mu}(\boldsymbol{x})} \right] \qquad E_{\text{eff}}^{\mu}(\boldsymbol{x}) = -\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{a}^\mu e^{-t}\|^2}{2\Delta_t}$$

- The energy levels being independent, the second term is an instance of the *Random Energy Model,* well-studied in statistical physics of spin-glasses **and concentrates for large $\boldsymbol{n}, \boldsymbol{d}$** [Derrida+1981, Lucibello+2024]
- The goal is to know if the first or the second term dominates, **respectively leading to collapse or generalisation**

Regime III and collapse

- In Regimes I and II, $P_t^e(\boldsymbol{x}) \approx P_t^{\text{true}}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{a} P_0(\boldsymbol{a})\gamma_t(\boldsymbol{x}, \boldsymbol{a})$
- This is no longer true in **Regime III** where the dynamics get attracted by one of the training point
- Consider a noisy sample $\boldsymbol{x}$ obtained from $\boldsymbol{a}^1$. The empirical probability is hence

$$P_t^e(\boldsymbol{x}) \propto \left[e^{-\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{a}^1 e^{-t}\|^2}{2\Delta_t}} + \sum_{\mu=2}^{n} e^{E_{\text{eff}}^{\mu}(\boldsymbol{x})} \right] \qquad E_{\text{eff}}^{\mu}(\boldsymbol{x}) = -\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{a}^{\mu} e^{-t}\|^2}{2\Delta_t}$$

- The energy levels being independent, the second term is an instance of the *Random Energy Model,* well-studied in statistical physics of spin-glasses **and concentrates for large $\boldsymbol{n}, \boldsymbol{d}$** [Derrida+1981, Lucibello+2024]
- The goal is to know if the first or the second term dominates, **respectively leading to collapse or generalisation**

Using a large-deviation analysis, we find that the timescale controlling this transition is the ***collapse*** time t_C, defined as

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

Curse of dimensionality: one requires a training set of size $n \sim e^d$ examples to avoid collapse!

Generalisation of these results

SPECIATION

From the time-reversal symmetry, speciation occurs when $\Lambda e^{-2t} \approx \Delta_t$, where Λ is the largest eigenvalue of the covariance matrix, meaning

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

Generalisation of these results

SPECIATION

From the time-reversal symmetry, speciation occurs when $\Lambda e^{-2t} \approx \Delta_t$, where Λ is the largest eigenvalue of the covariance matrix, meaning

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

COLLAPSE

- Collapse is due to the empirical approximation of the probability distribution → Need to know when $P_t^e(\boldsymbol{x}) \approx P_t(\boldsymbol{x})$

Generalisation of these results

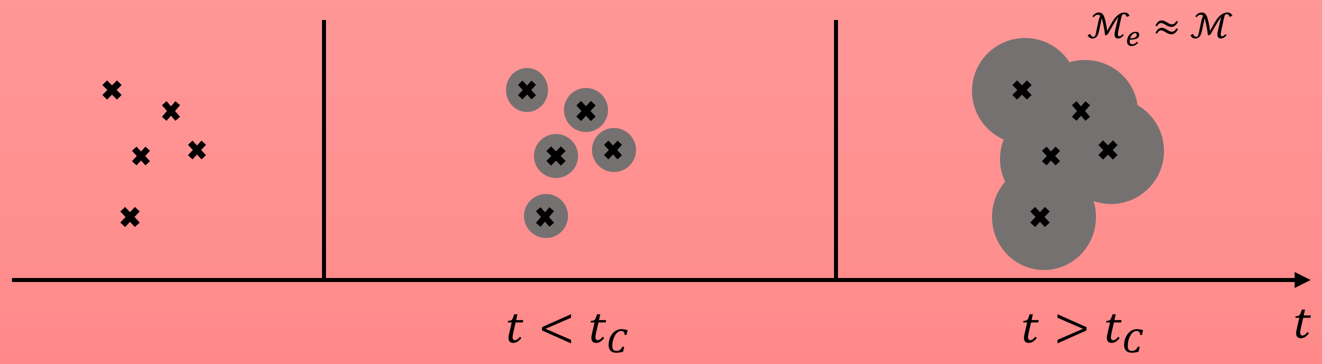
SPECIATION

From the time-reversal symmetry, speciation occurs when $\Lambda e^{-2t} \approx \Delta_t$, where Λ is the largest eigenvalue of the covariance matrix, meaning

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

COLLAPSE

- Collapse is due to the empirical approximation of the probability distribution → Need to know when $P_t^e(\boldsymbol{x}) \approx P_t(\boldsymbol{x})$



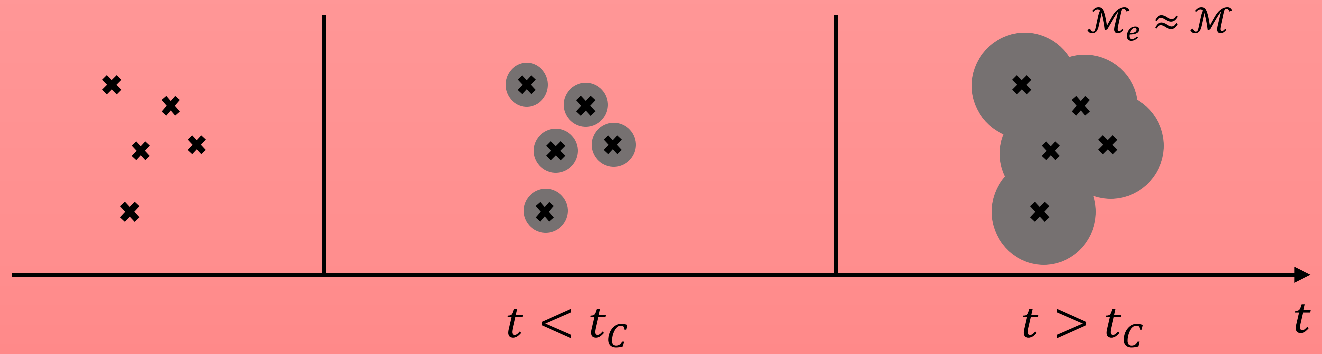
SPECIATION

From the time-reversal symmetry, speciation occurs when $\Lambda e^{-2t} \approx \Delta_t$, where Λ is the largest eigenvalue of the covariance matrix, meaning

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

COLLAPSE

- Collapse is due to the empirical approximation of the probability distribution → Need to know when $P_t^e(\boldsymbol{x}) \approx P_t(\boldsymbol{x})$



- This suggests a volume (or equivalently, entropy) argument where the collapse time is controlled by the excess entropy

$$f(t) = S_{\text{Gauss}}(t) - S(t),$$

where $S(t) = -\frac{1}{d}\int \mathrm{d}\boldsymbol{x} P_t(\boldsymbol{x}) \log P_t(\boldsymbol{x})$ is the Shannon entropy.

Generalisation of these results

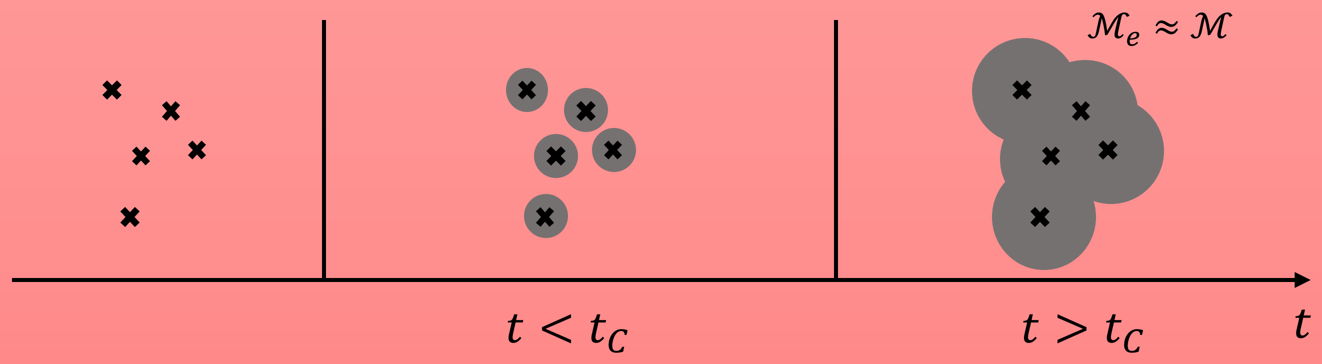
SPECIATION

From the time-reversal symmetry, speciation occurs when $\Lambda e^{-2t} \approx \Delta_t$, where Λ is the largest eigenvalue of the covariance matrix, meaning

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

COLLAPSE

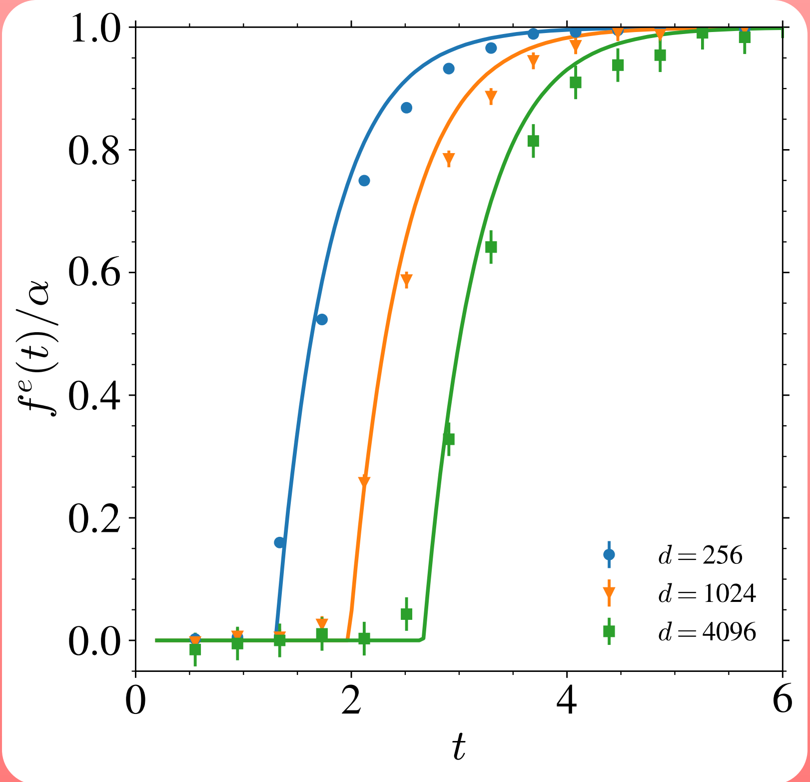
- Collapse is due to the empirical approximation of the probability distribution → Need to know when $P_t^e(\boldsymbol{x}) \approx P_t(\boldsymbol{x})$



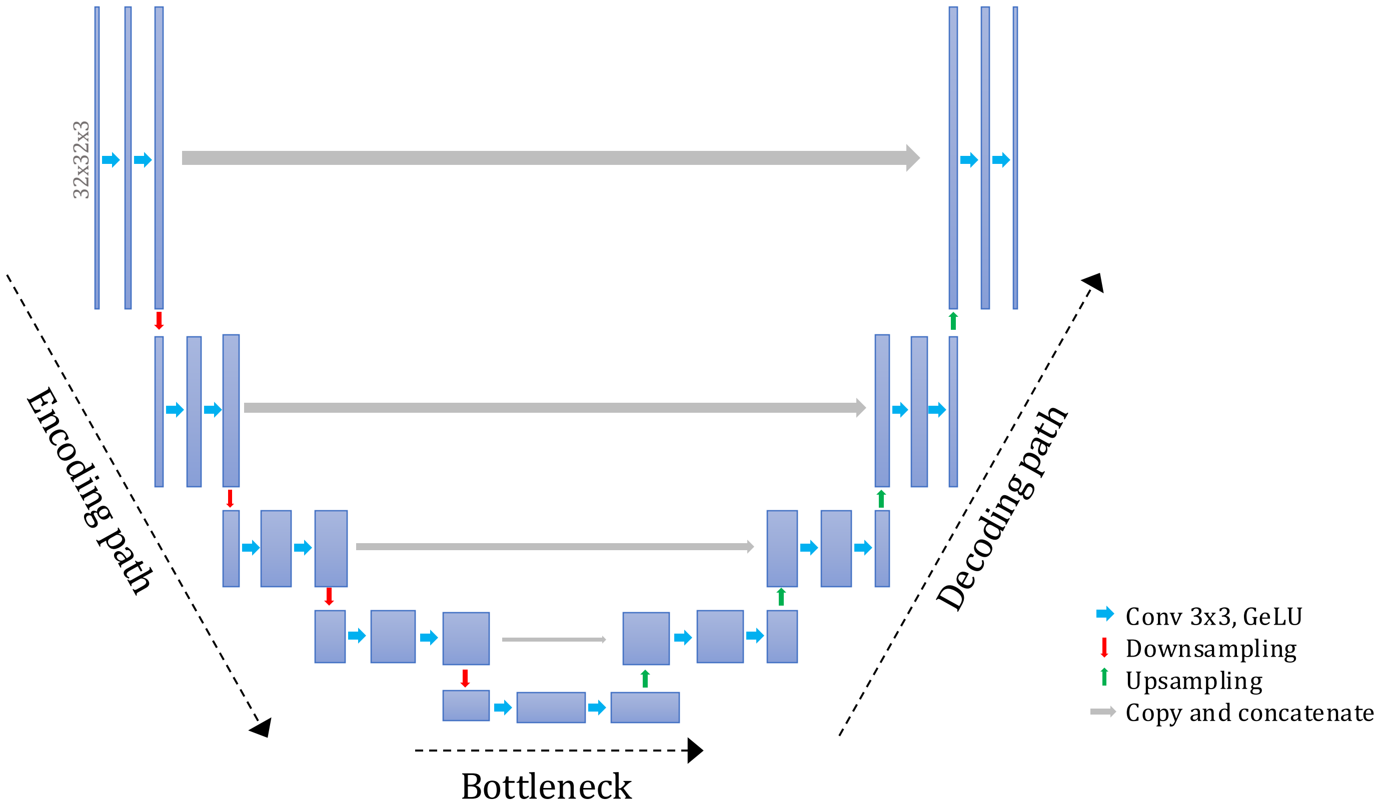
- This suggests a volume (or equivalently, entropy) argument where the collapse time is controlled by the excess entropy

$$f(t) = S_{\text{Gauss}}(t) - S(t),$$

where $S(t) = -\frac{1}{d}\int \mathrm{d}\boldsymbol{x} P_t(\boldsymbol{x}) \log P_t(\boldsymbol{x})$ is the Shannon entropy.



- We trained a Denoising Diffusion Probabilistic model (DDPM) [Ho+2020]
- The denoiser has a U-Net architecture [Ronneberger+2015] and approximates the score $F(\boldsymbol{x}, t)$

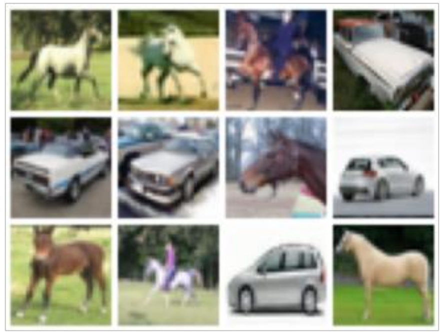


- Time is embedded through sinusoidal position embedding and added to the features of all maps
- Attention [Vaswani+2017] is applied to resolution levels two and three, resulting in a total of 25.7M parameters

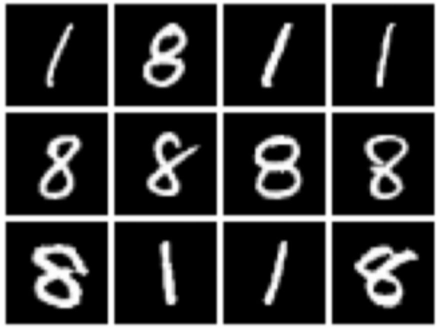
Realistic image datasets

ImageNet-16
- $n = 2000$
- L. pandas and seashores
- $d = 16 \times 16 \times 3 = 768$

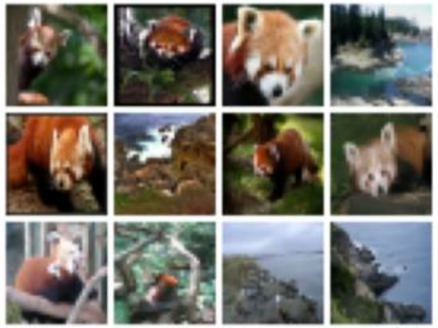
CIFAR
- $n = 3000$
- Classes horses and cars
- $N = 32 \times 32 \times 3 = 3072$



MNIST32
- $n = 10000$
- Classes 1 and 8
- $d = 32 \times 32 \times 1 = 1024$



ImageNet-32
- $n = 2000$
- L. pandas and seashores
- $d = 32 \times 32 \times 3 = 3072$



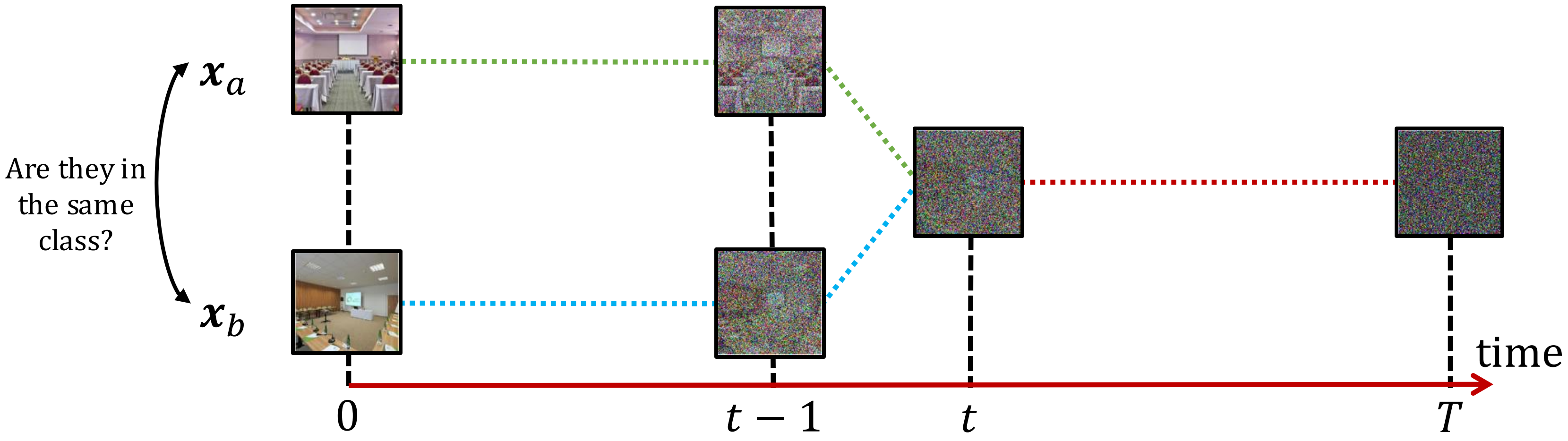
LSUN64
- $n = 40000$
- Conference and churches
- $d = 64 \times 64 \times 3 = 12288$

- All the models are trained for 350k steps
- Fixed learning rate of 10^{-4} and ADAM optimizer
- Linear scheduler for the variance as in [Ho+2020]
- Batch size of 128 except for LSUN with 64

Cloning experiment

HOW TO ANALYZE SPECIATION NUMERICALLY?

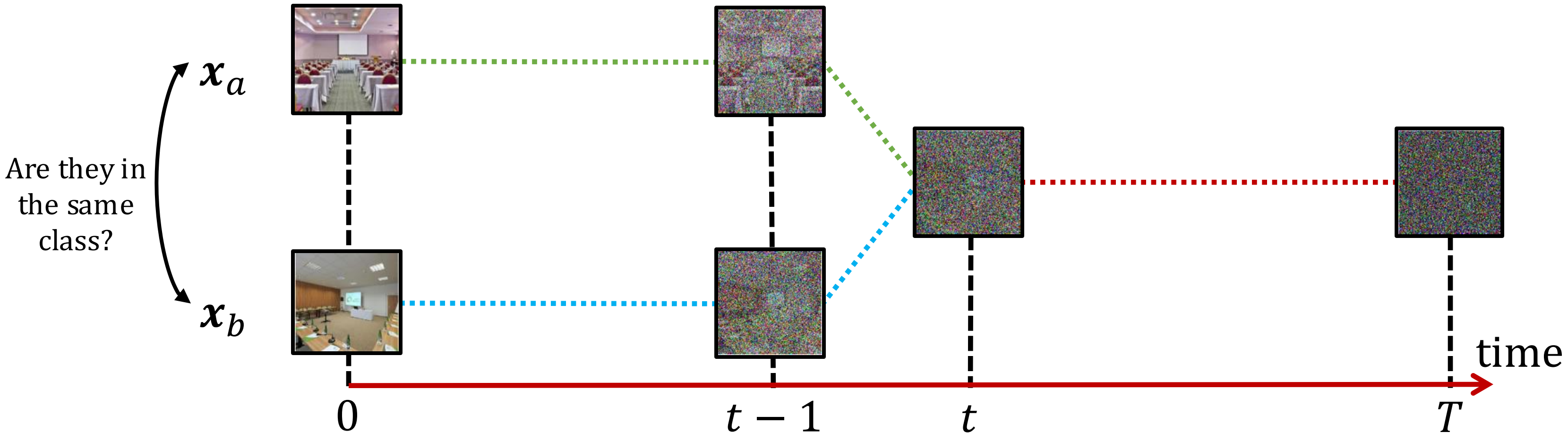
- Characterize the time at which the barrier do not allow to switch between the two classes
- **Cloning experiment**: Sample a trajectory backward in time and then clone it for $\tau < t$ to make two trajectories evolve with independent noise

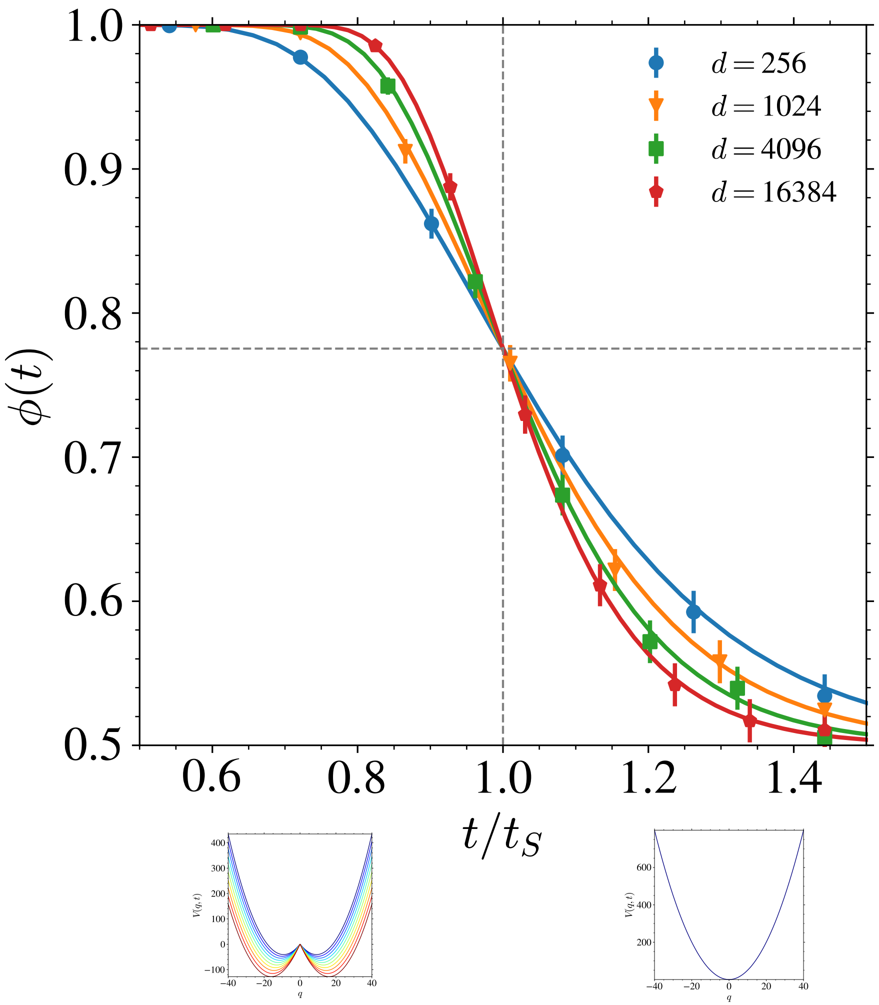


- Measure the probability $\phi(t)$ that the two clones end up in the same class

HOW TO ANALYZE SPECIATION NUMERICALLY?

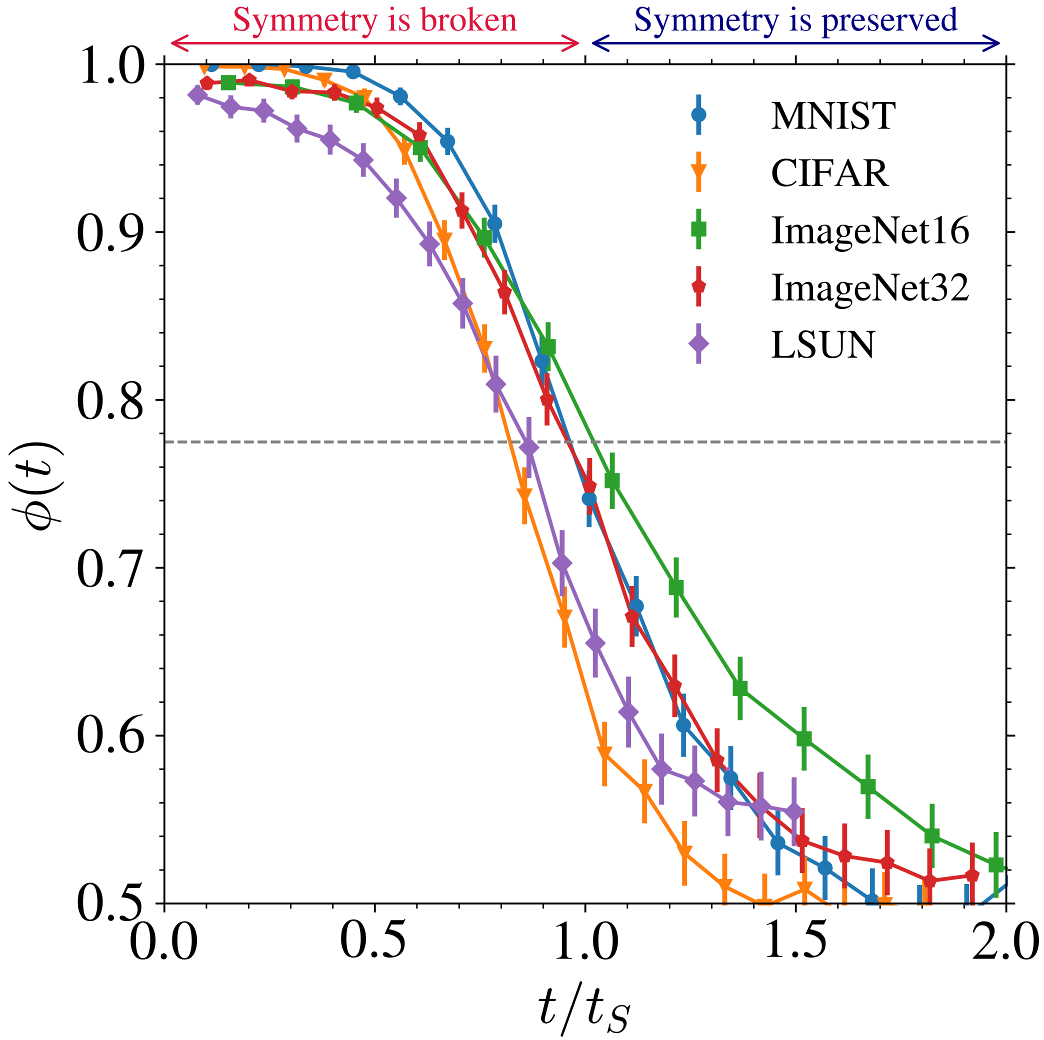
- Characterize the time at which the barrier do not allow to switch between the two classes
- **Cloning experiment**: Sample a trajectory backward in time and then clone it for $\tau < t$ to make two trajectories evolve with independent noise

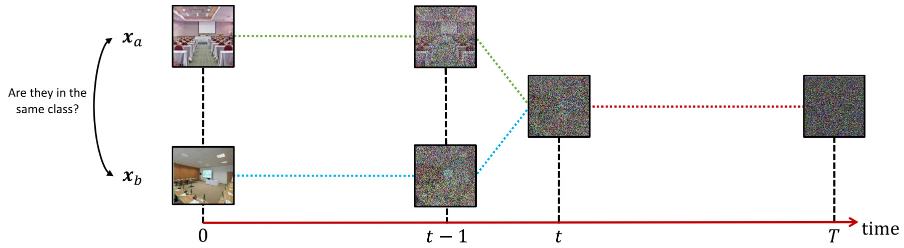




- Measure the probability $\phi(t)$ that the two clones end up in the same class

CLONING EXPERIMENT ON REALISTIC DATASETS





- $\phi(t)$ is computed using a ResNet-18 pre-trained on ImageNet and re-trained on each dataset
- The cloning time t is rescaled by the prediction

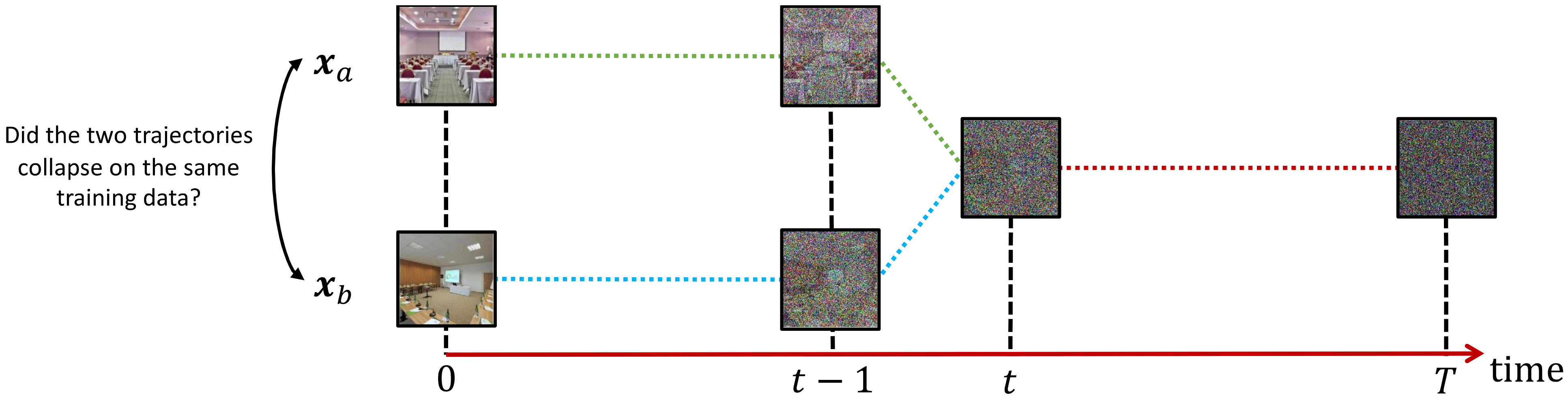
$$t_S = \frac{1}{2}\log\Lambda$$

- **Validates the speciation phenomenon in realistic datasets** and on a timescale in agreement (max 15% error) with the theoretical prediction
- See also the U-turn experiment from [Behjoo+2023]

$$f^e(t) = \underbrace{\frac{\log n}{d} + \frac{1}{2} + \frac{1}{2}\log 2\pi\Delta_t}_{S_{\text{Gauss}}(t)} + \underbrace{\frac{1}{d}\int \mathrm{d}x P_t^e(x) \log P_t^e(x)}_{-S(t)}$$

HOW TO ANALYZE COLLAPSE NUMERICALLY?

1. Cloning experiment but computing $\phi_C(t)$, the probability that the two trajectories have the same nearest neighbour at the end of the backward time



$$f^e(t) = \underbrace{\frac{\log n}{d} + \frac{1}{2} + \frac{1}{2}\log 2\pi\Delta_t}_{S_{\text{Gauss}}(t)} + \underbrace{\frac{1}{d}\int \mathrm{d}x P_t^e(x) \log P_t^e(x)}_{-S(t)}$$

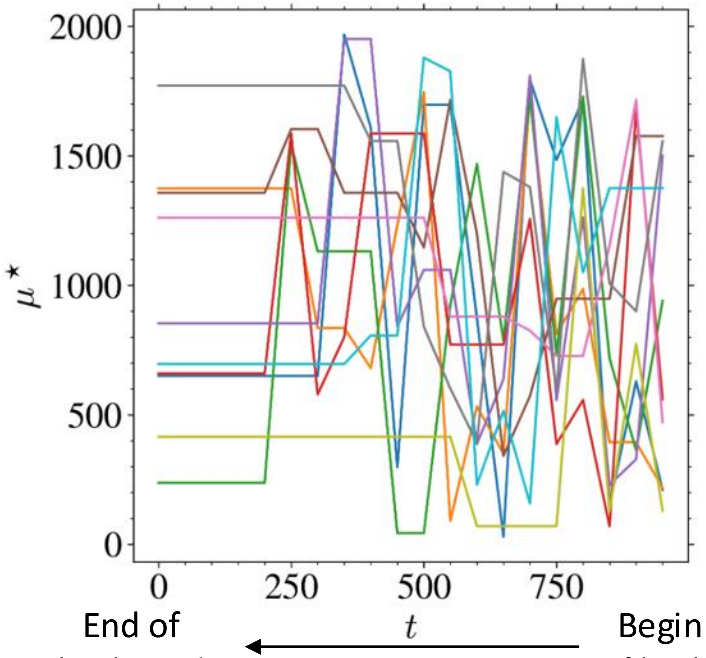
HOW TO ANALYZE COLLAPSE NUMERICALLY?

2. Time of last-changing index $\mu_\star(t)$ of closest neighbour in the training set

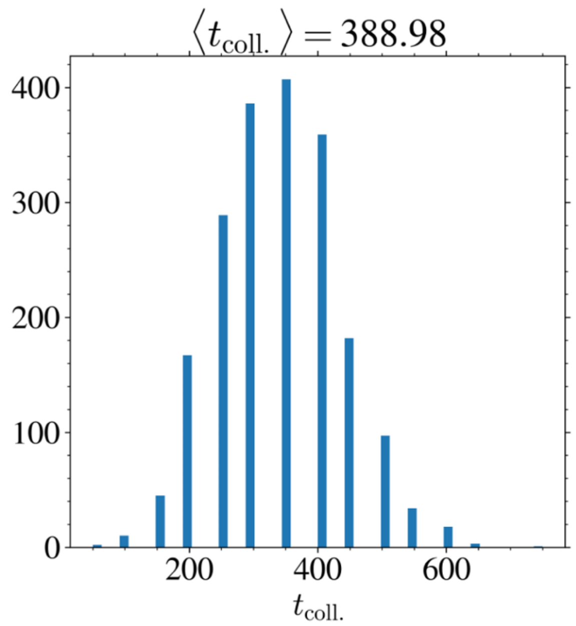
$$\mu_\star(\tilde{x}) = \text{argmin}_{\boldsymbol{a}^\mu \in X} \|\boldsymbol{a}^\mu e^{-t} - \tilde{\boldsymbol{x}}\|_2^2$$

$\boldsymbol{a}^\mu$: training image

$\tilde{\boldsymbol{x}}$: generated image



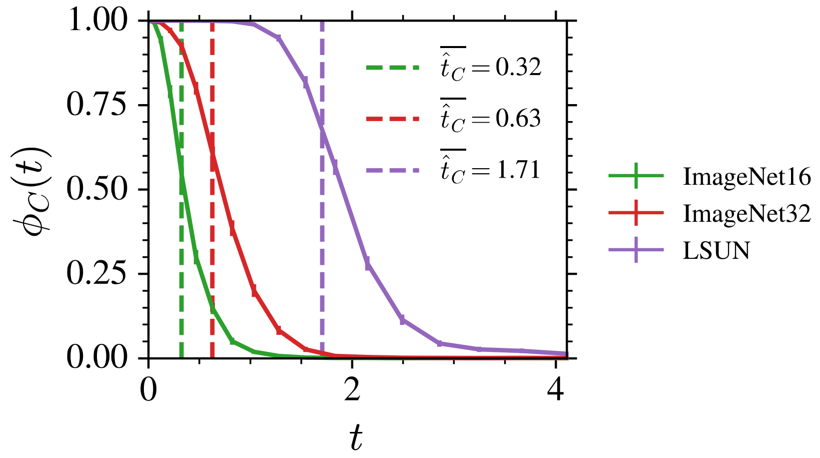
End of backward ← Beginning of backward



Collapse transition (Regimes II-III)

$$f^e(t) = \underbrace{\frac{\log n}{d} + \frac{1}{2} + \frac{1}{2}\log 2\pi\Delta_t}_{S_{\mathrm{Gauss}}(t)} + \underbrace{\frac{1}{d}\int \mathrm{d}x P_t^e(x) \log P_t^e(x)}_{-S(t)}$$

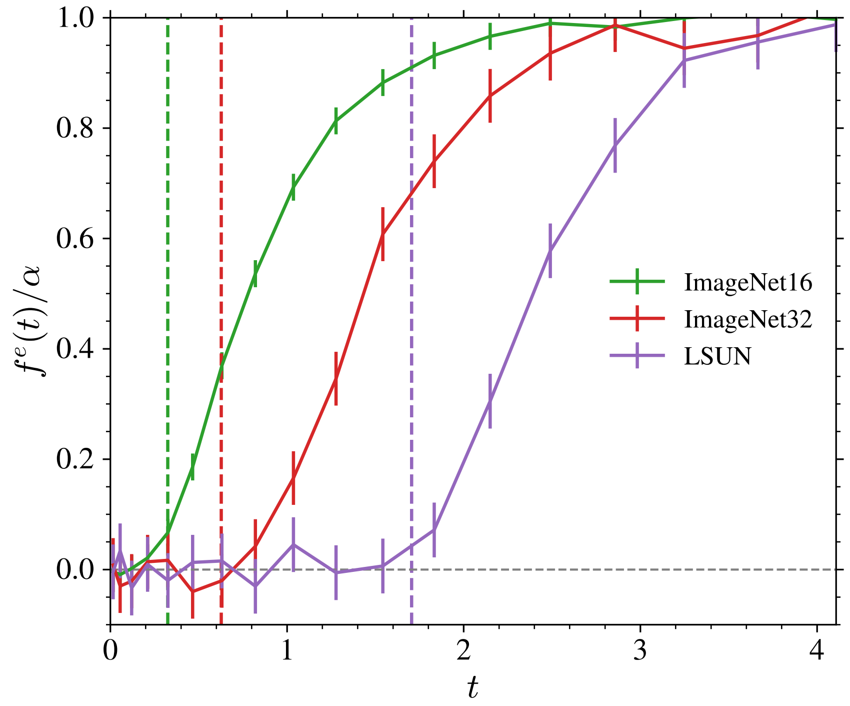
- The two estimates agree quite well on realistic datasets



Collapse transition (Regimes II-III)

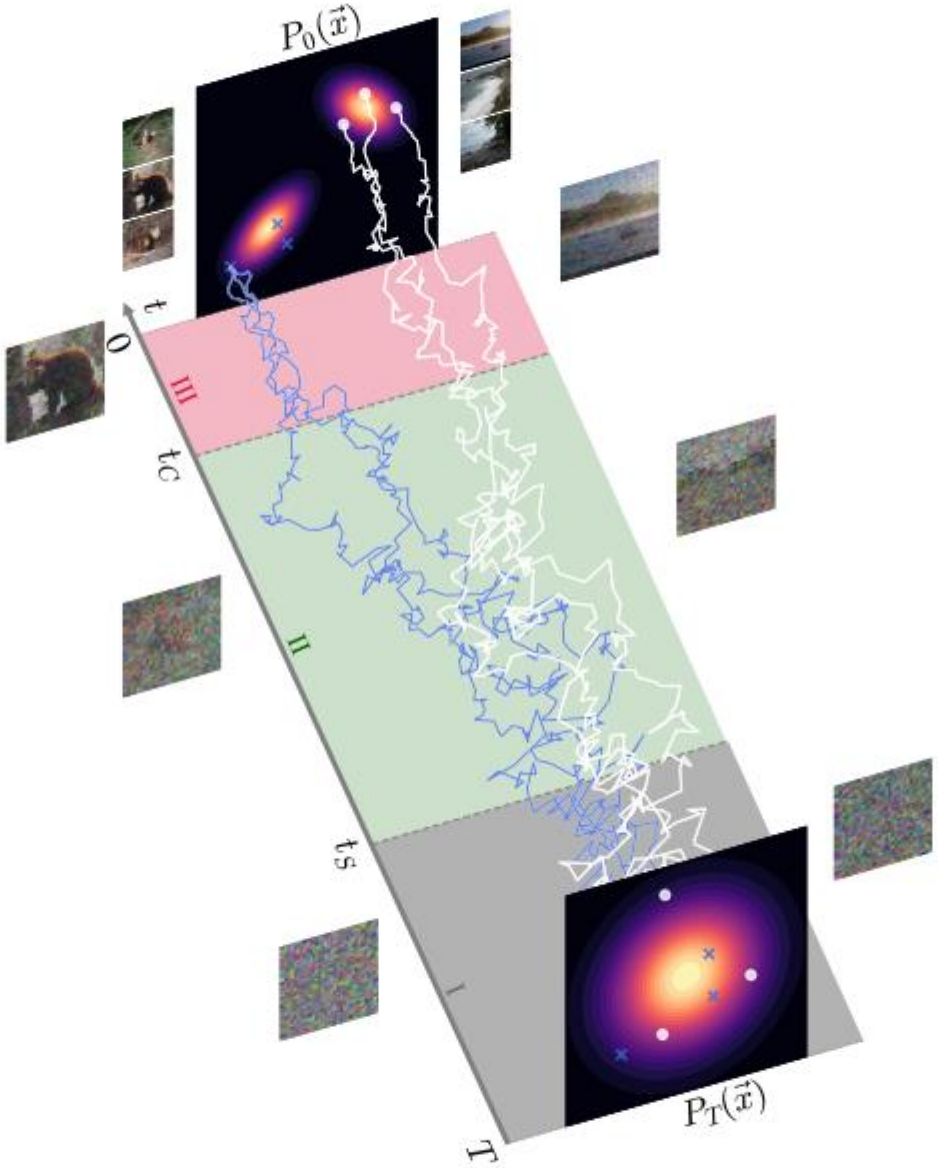
$$f^e(t) = \underbrace{\frac{\log n}{d} + \frac{1}{2} + \frac{1}{2}\log 2\pi\Delta_t}_{S_{\text{Gauss}}(t)} + \underbrace{\frac{1}{d}\int \mathrm{d}x P_t^e(x) \log P_t^e(x)}_{-S(t)}$$

- The two estimates agree quite well on realistic datasets
- They are also consistent with the time where $f^e(t)/\alpha$ cancels for all datasets, as predicted by the theory
- **Validates the collapse phenomenon in realistic datasets** and on a timescale in agreement with the theoretical prediction



SUMMARY

- Three dynamical regimes in the backward dynamics:
  - I. Random motion
  - II. Features formation
  - III. Memorization
- Transition I-II was called *speciation* and characterised by the largest eigenvalue of the data covariance.
- Transition II-III was called *collapse* and characterised by the excess entropy of the distribution.



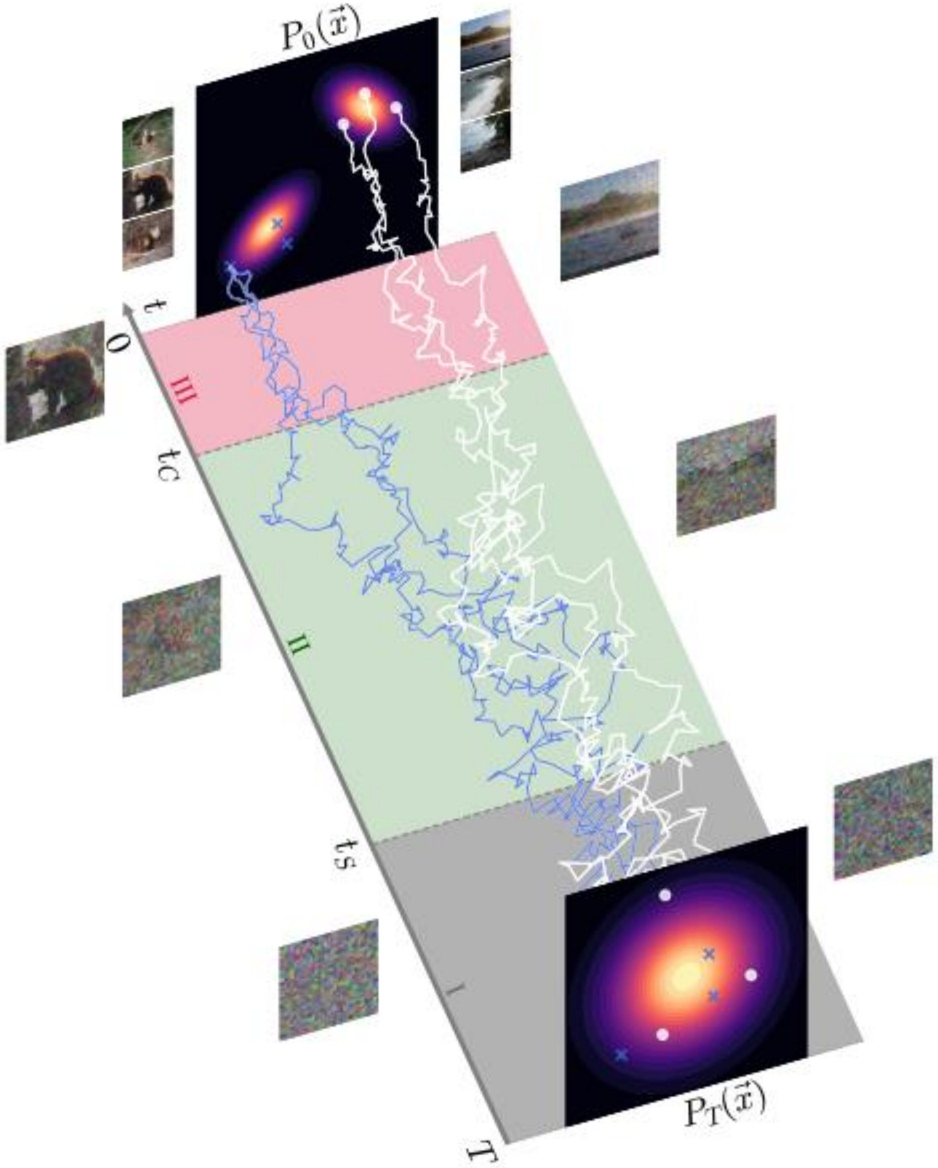
Wrapping-up and perspectives

SUMMARY

- Three dynamical regimes in the backward dynamics:
  I. Random motion
  II. Features formation
  III. Memorization
- Transition I-II was called *speciation* and characterised by the largest eigenvalue of the data covariance.
- Transition II-III was called *collapse* and characterised by the excess entropy of the distribution.

PERSPECTIVES

- Can we use the first 'noise' phase to accelerate sampling?
- How is memorization avoided in practice?
  1. What is the role of regularization and number of data?
  2. What is the role of structure in the data? Can it be studied analytically?



THANKS FOR YOUR ATTENTION!