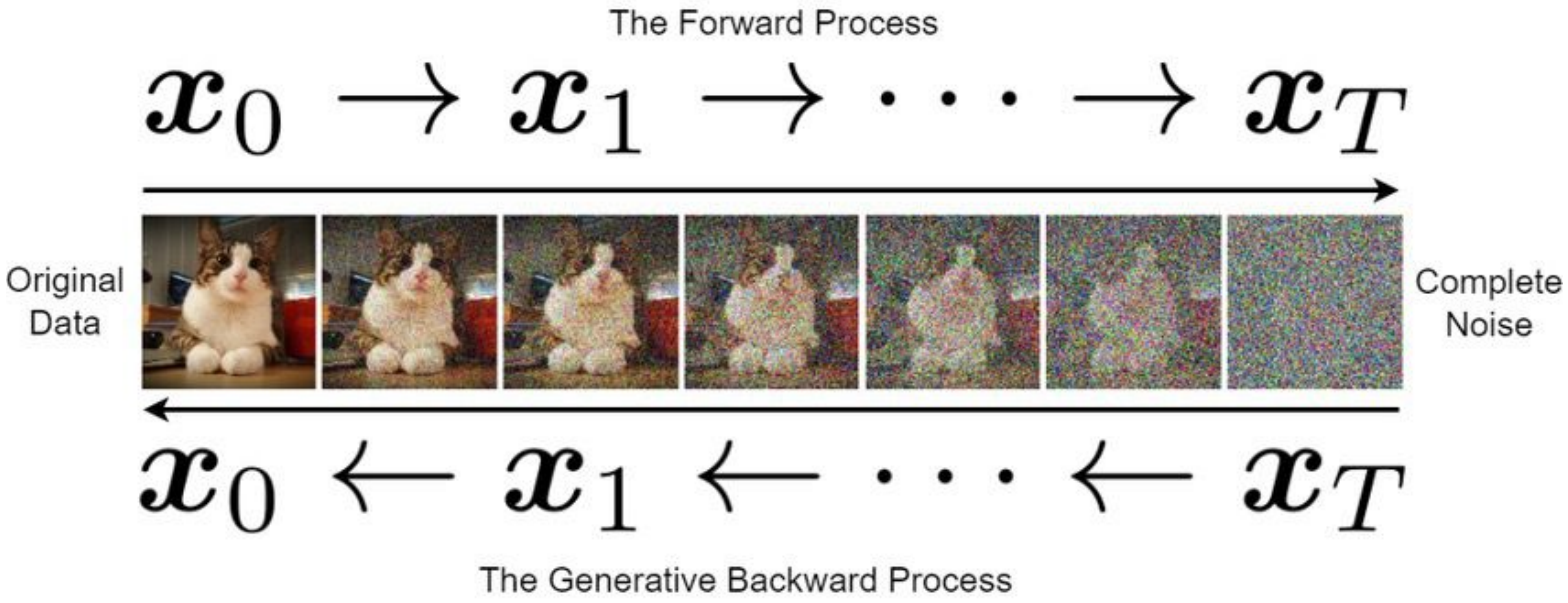
A solvable model of Generative Diffusion and the memorization phenomenon

Leonardo Bandera

Agenda

- Diffusion Sampling and problem statement
- Understanding training dynamics: the original setting
- First line of research: towards a more realistic architecture
- Second line of research: structured data

Generative Diffusion

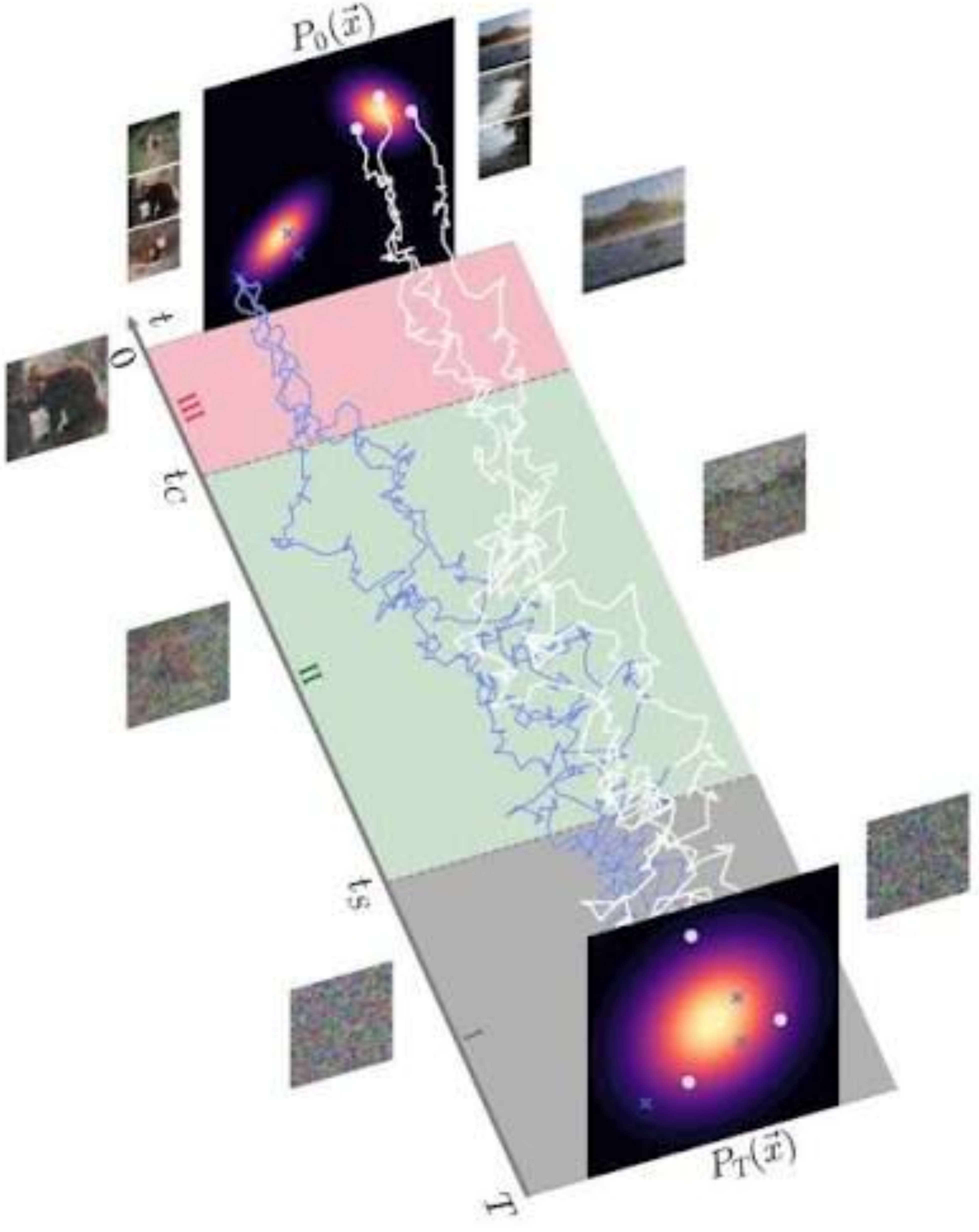

The Forward Process
$x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_T$
Original
Data
Complete
Noise
$x_0 \leftarrow x_1 \leftarrow \cdots \leftarrow x_T$
The Generative Backward Process

Score-based Diffusion

- Forward process: $\dfrac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{\eta}(t)$
- Reversed process: $\dfrac{d\boldsymbol{x}}{d\tau} = \boldsymbol{x}(\tau) + 2\nabla \log \widetilde{P}_{\tau}(\boldsymbol{x}) + \widetilde{\boldsymbol{\eta}}(\tau)$
- The score $S(\boldsymbol{x}, t) = \nabla \log P_t(\boldsymbol{x})$ is unknown… we learn it!

$$\mathscr{L}_{\lambda}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\mu=1}^{n} \mathbb{E}_{t\sim Q(t)}\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,\mathbb{I}_d)} \left\| \hat{\boldsymbol{S}}^{\boldsymbol{\theta}}(\boldsymbol{x}_t^{\mu}(\boldsymbol{\xi}), t) + \frac{\boldsymbol{\xi}}{\sqrt{\Delta_t}} \right\|^2$$

Problem: sampling with the empirical loss minimizer leads to memorization!



Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. Nature Communications, 15(1):9957, nov 2024

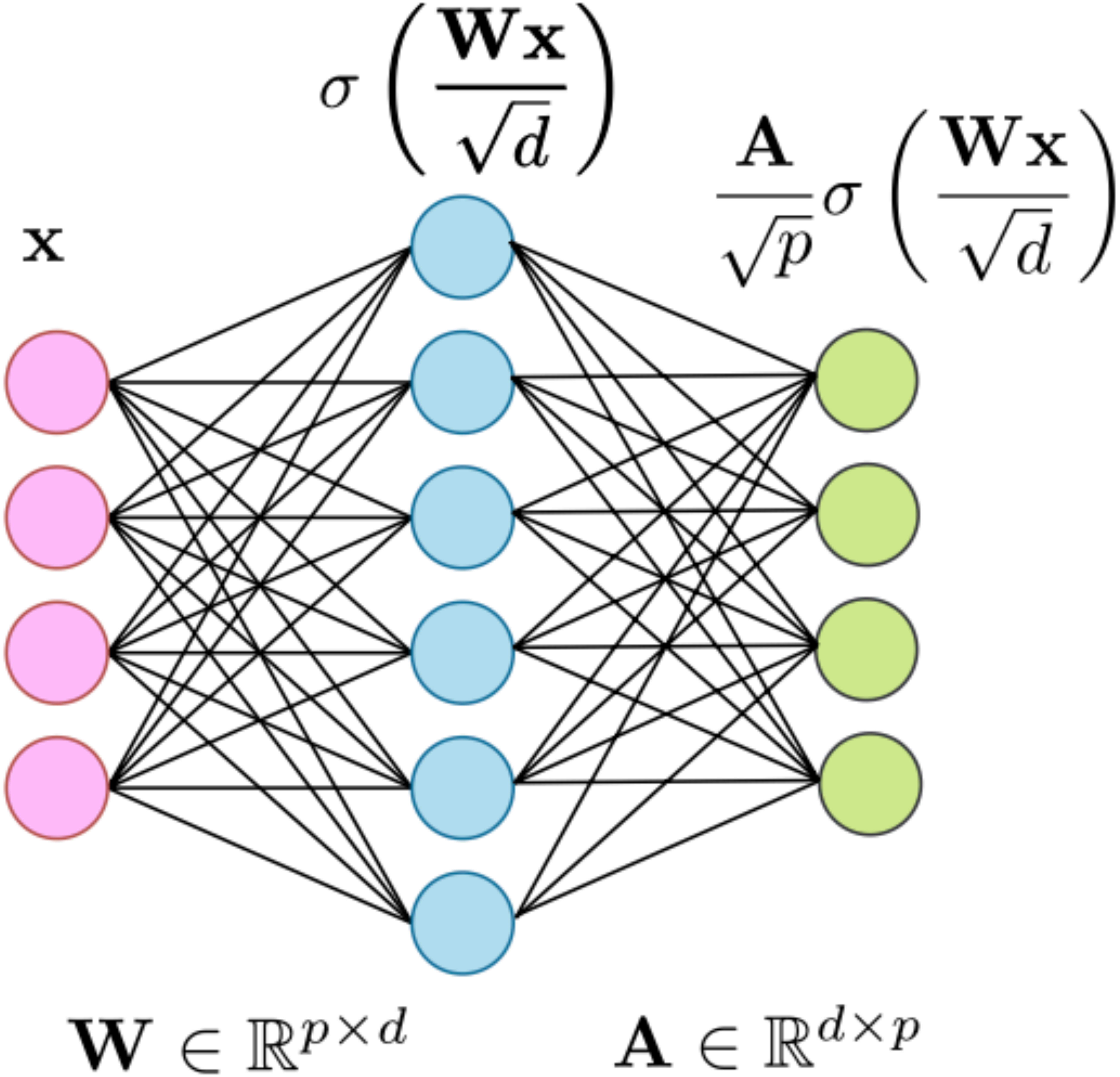
Why real diffusion systems avoid this degeneracy?

- Architectural constraints
- **Training dynamics**

The original setting

Gaussian data

One RFNN per time



Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don’t memorize: The role of implicit dynamical regularization in training. arXivpreprint arXiv:2505.17638, 2025

Gradient Flow

$$\dot{\boldsymbol{A}}(\tau) = -2\Delta_t \frac{d}{p} \boldsymbol{A}\boldsymbol{U} - \frac{2d\sqrt{\Delta_t}}{\sqrt{p}} \boldsymbol{V}^\top$$

where

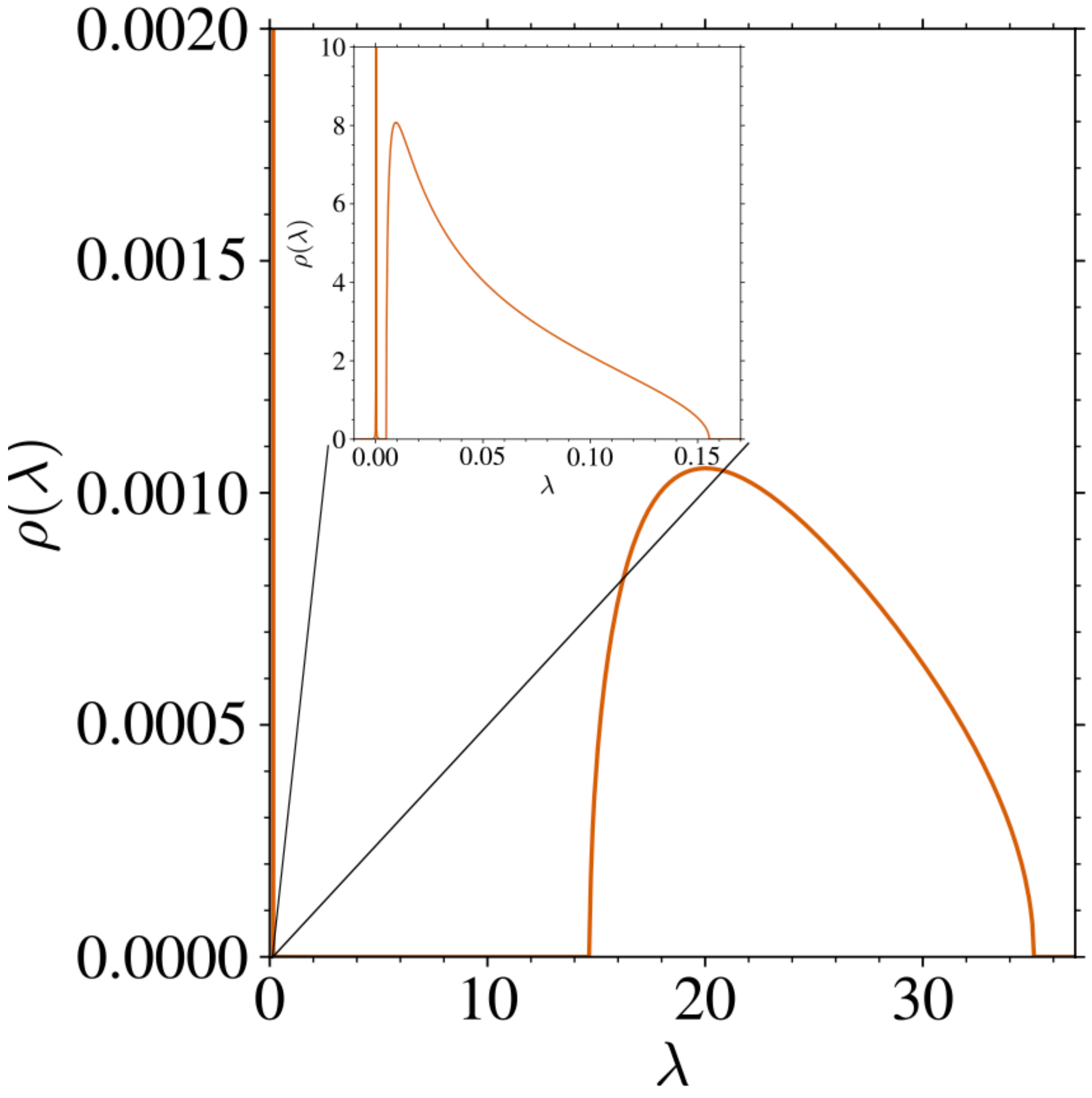
$$\boldsymbol{U} = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\xi}}[\sigma(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}})\sigma(\frac{\boldsymbol{W}\boldsymbol{x}_t^\nu(\boldsymbol{\xi})}{\sqrt{d}})^\top]$$

The timescales of the training dynamics are given by the inverse eigenvalues of the matrix $\Delta_t \frac{d}{p} \boldsymbol{U}$!

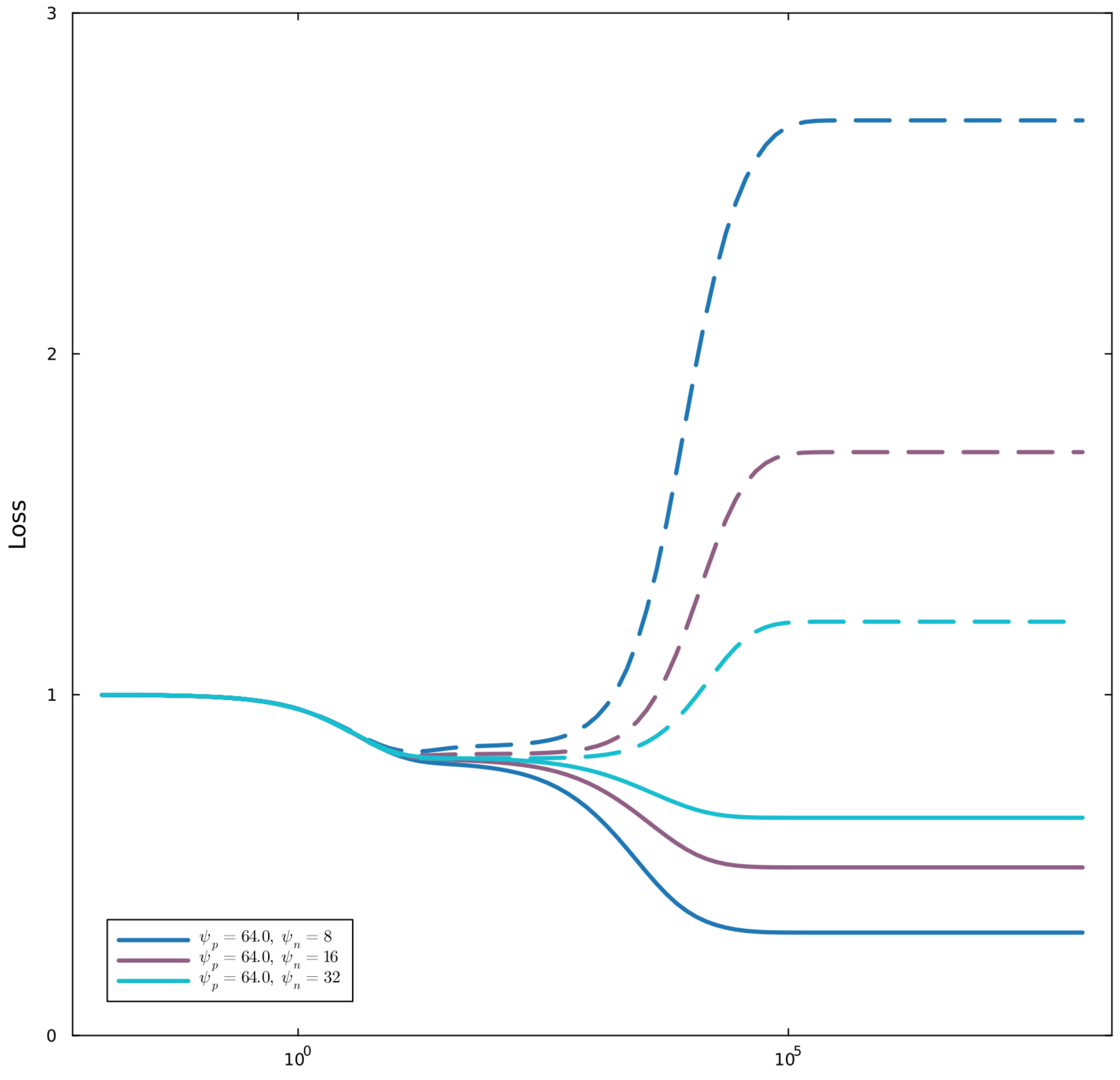
U spectrum

- Stieltjes transform
- GEP
- Replica

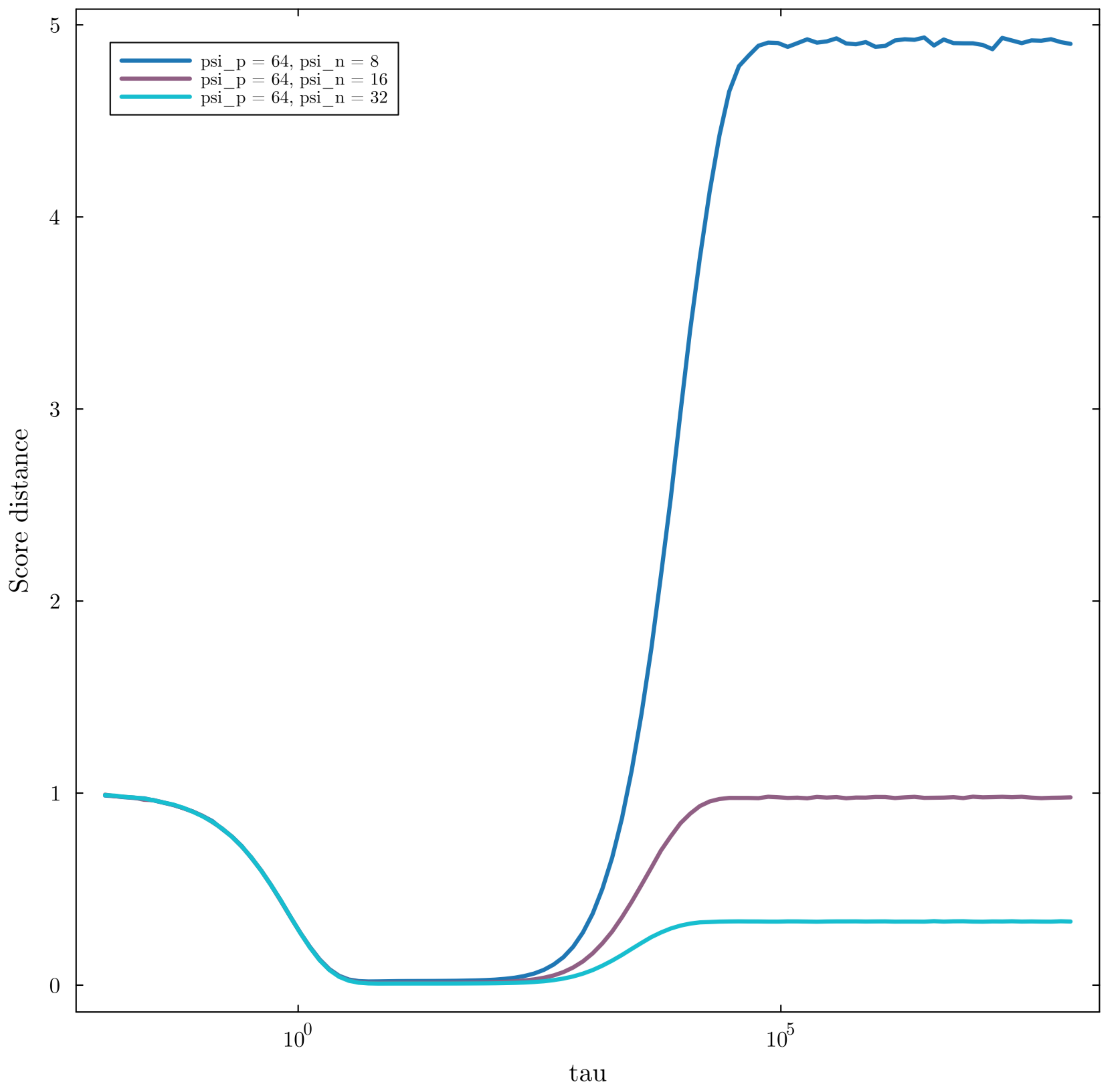
Two bulks!



What’s happening at the metrics at the timescales associated to the bulks?



Loss
0
1
2
3
10^0
10^5
ψ_p = 64.0, ψ_n = 8
ψ_p = 64.0, ψ_n = 16
ψ_p = 64.0, ψ_n = 32



Score distance
0
1
2
3
4
5
10^0
10^5
tau
psi_p = 64, psi_n = 8
psi_p = 64, psi_n = 16
psi_p = 64, psi_n = 32

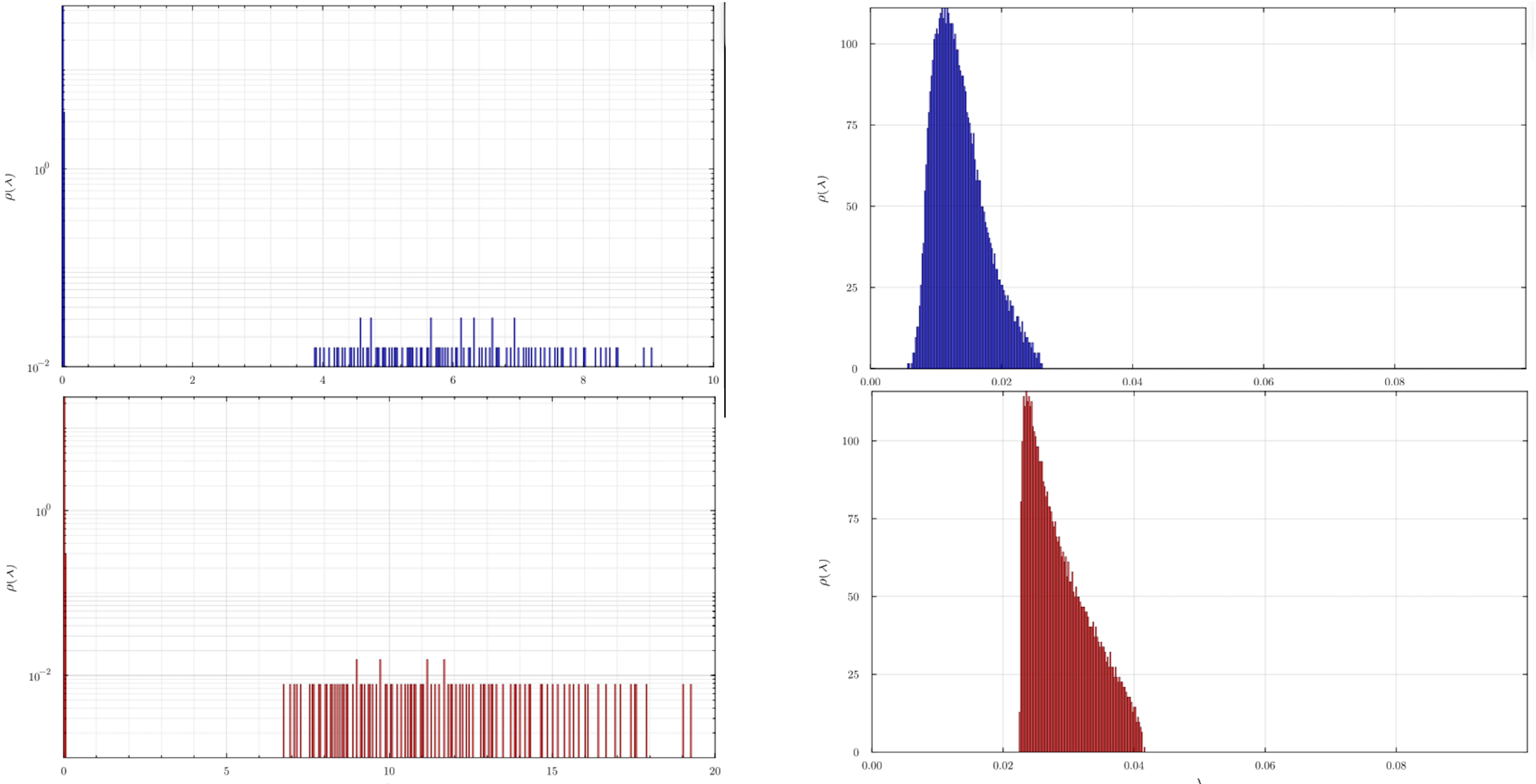
Time Integrated RFNN

$$S_{\boldsymbol{A}}(\boldsymbol{x}, t) = \alpha_t \frac{\boldsymbol{A}}{\sqrt{p}} \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}} + t\boldsymbol{b} \right) + \beta_t \boldsymbol{x}$$

$$\mathscr{L}(\boldsymbol{A}, \{\boldsymbol{x}_{\boldsymbol{i}}^{\nu}\}) = \frac{1}{n} \sum_{\nu=1}^{n} \frac{1}{d} \mathbb{E}_{\boldsymbol{\xi},t} \parallel \sqrt{\Delta_t} S_{\boldsymbol{A}}(\boldsymbol{x}_t^{\nu}(\boldsymbol{\xi}), t) + \boldsymbol{\xi} \parallel^2$$

- Same phenomenology

- Time integration brakes GEP: we cannot compute $\boldsymbol{U}$ spectrum!



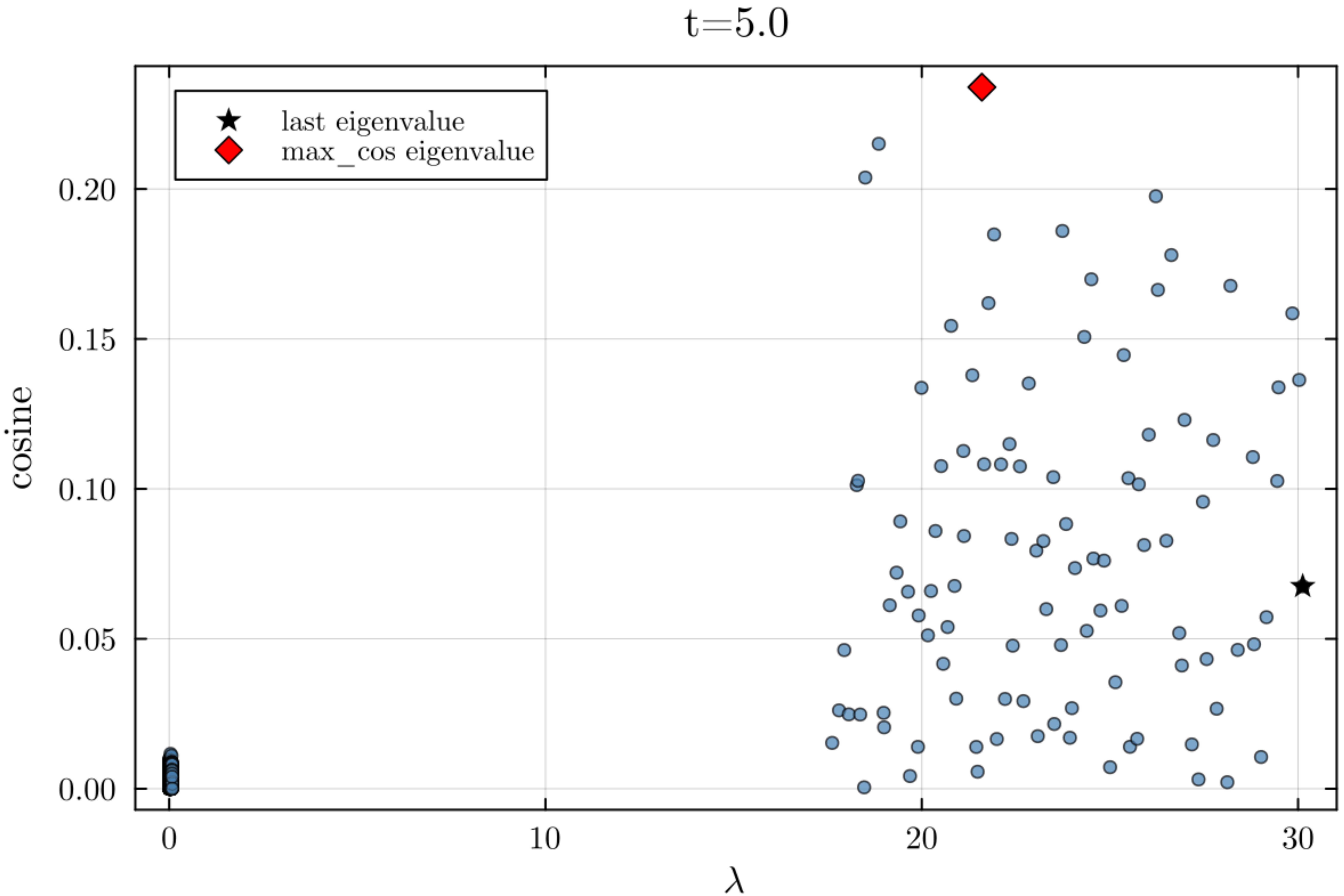
MoG Data

$$P_0 = \frac{1}{2}\mathcal{N}(\boldsymbol{m}, \sigma_{\boldsymbol{x}}^2 \mathbb{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{m}, \sigma_{\boldsymbol{x}}^2 \mathbb{I}_d)$$

or equivalently

$$\boldsymbol{x}^\mu = c^\mu \boldsymbol{m} + \sigma_{\boldsymbol{x}} \boldsymbol{z}^\mu$$

Phenomenology: U spectrum BBP transition



GEP holds!

- $\boldsymbol{U}_{GEP}$ for MoG data is $\boldsymbol{U}_{GEP}$ for gaussian data plus rank-1 terms: the bulks are the same!

- We must focus on the outlier…

Brandon Livio Annesi, Dario Bocchi, and Chiara Cammarota. Overparametrization

bends the landscape: Bbp transitions at initialization in simple neural networks.

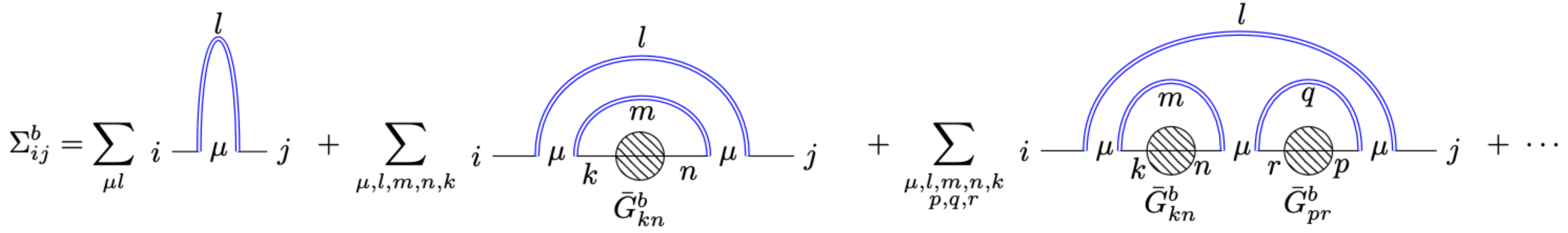
arXiv preprint arXiv:2510.18435, 2025. Submitted on 21 Oct 2025

We need to compute this:

$$\overline{G}_{ij} = \lim_{n \to 0} \mathbb{E}_x \int \prod_{a=1}^{n} d\psi^a \exp\left[-\frac{1}{2}\sum_a (\psi^a)^T (z\mathbb{I}_N - \boldsymbol{U})\psi^a\right] \psi_i^1 \psi_j^1$$

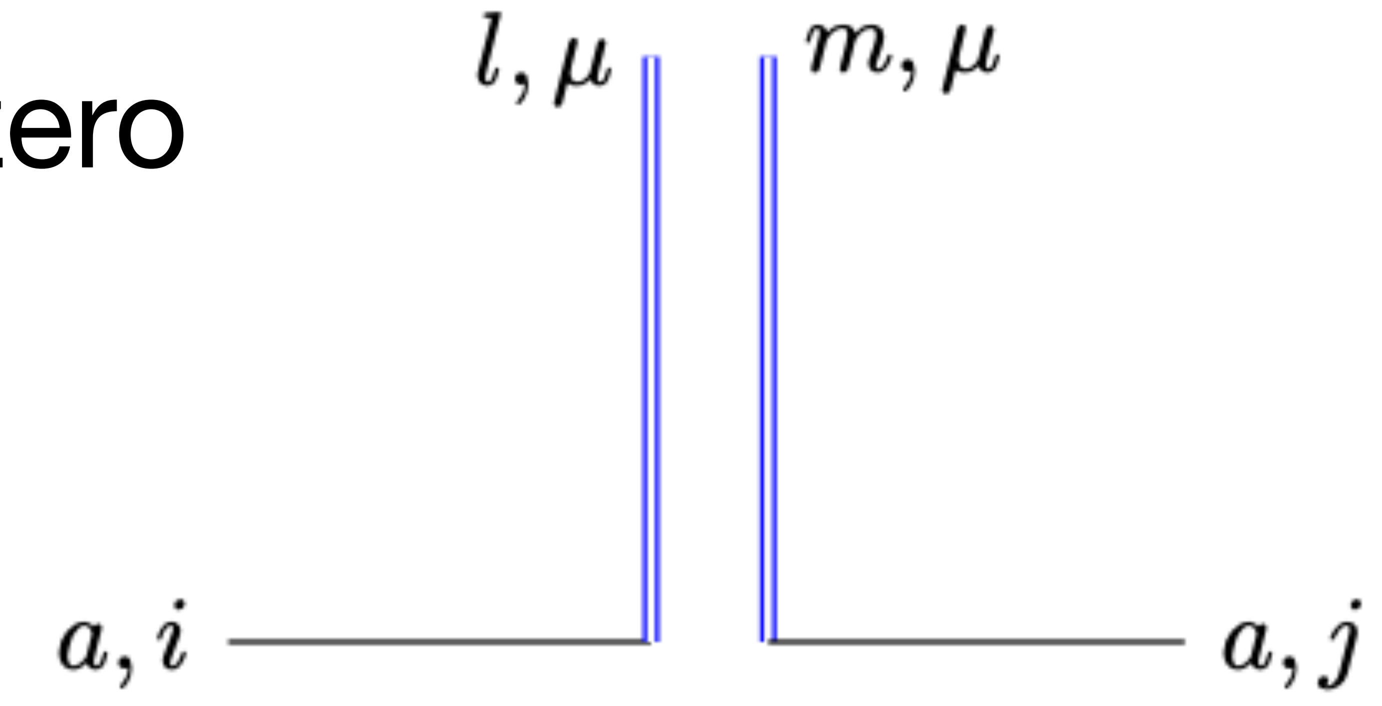
It's a matter of gaussian averages!

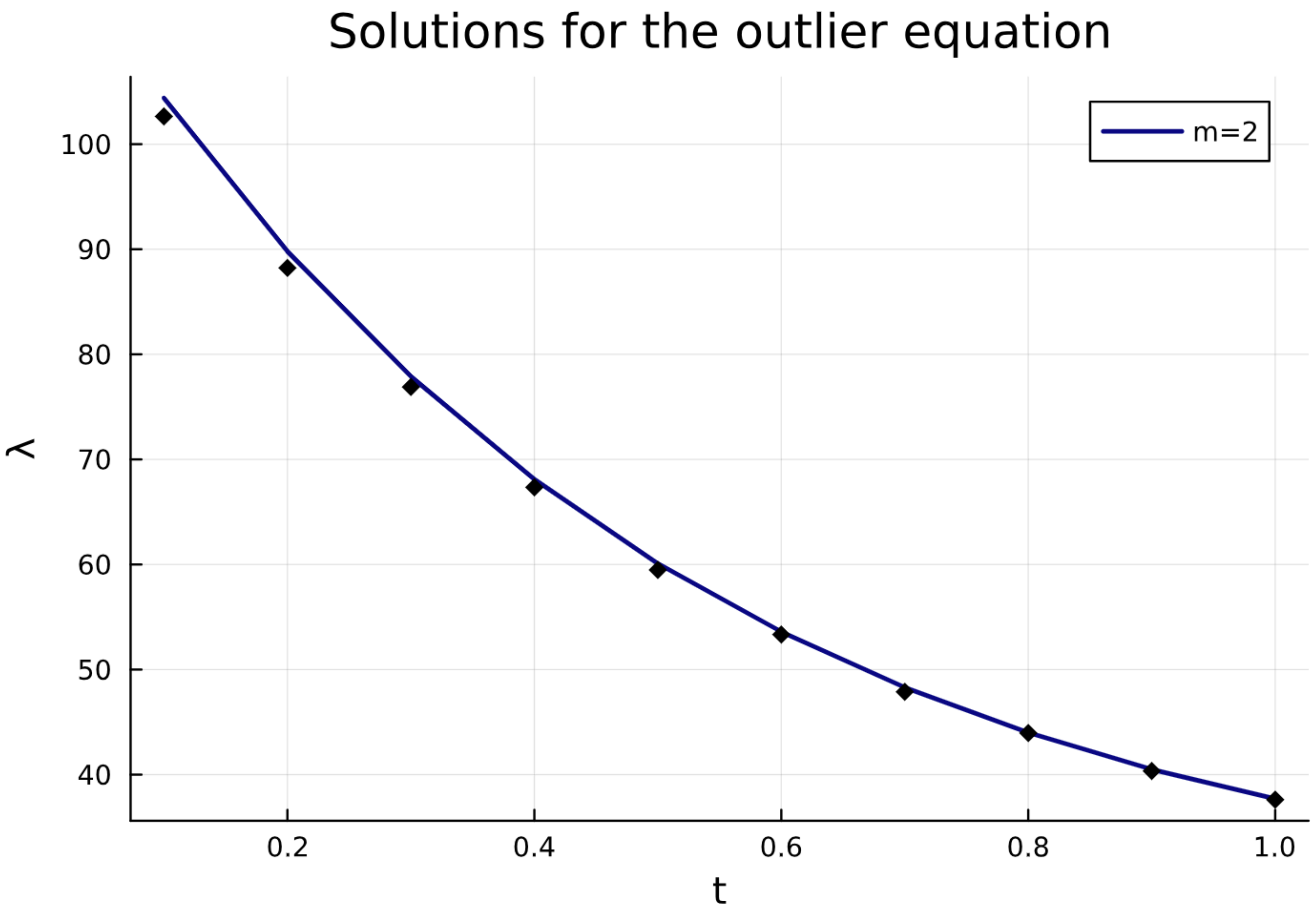
- Wick theorem to reduce averages over many fields to averages of pairs
- Feynman diagrams to keep track of all possible pairings and interactions between different gaussian fields



We get to a self-consistent equation for $\overline{G}_{ij}$:

- solving it gives the bulks
- the condition for the outlier is setting a denominator to zero





Solutions for the outlier equation
m=2
λ
t
100
90
80
70
60
50
40
0.2
0.4
0.6
0.8
1.0

This is interesting at two levels:

- *Diffusion dynamics*: what's happening at trajectories at transition time?
- *Training dynamics*: what's happening at the training metrics at the timescale associated to the outlier?

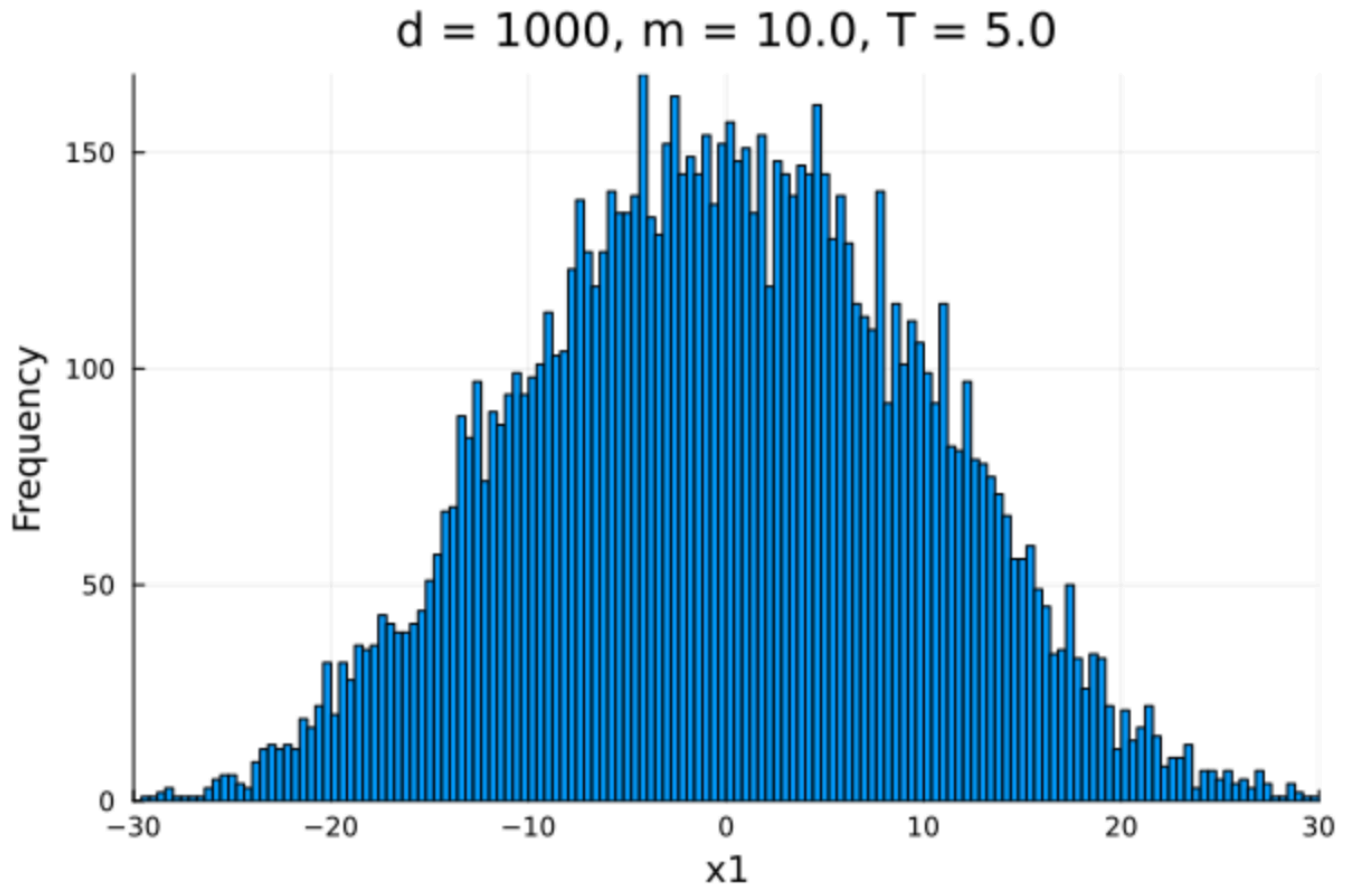
But…

… a legit doubt:

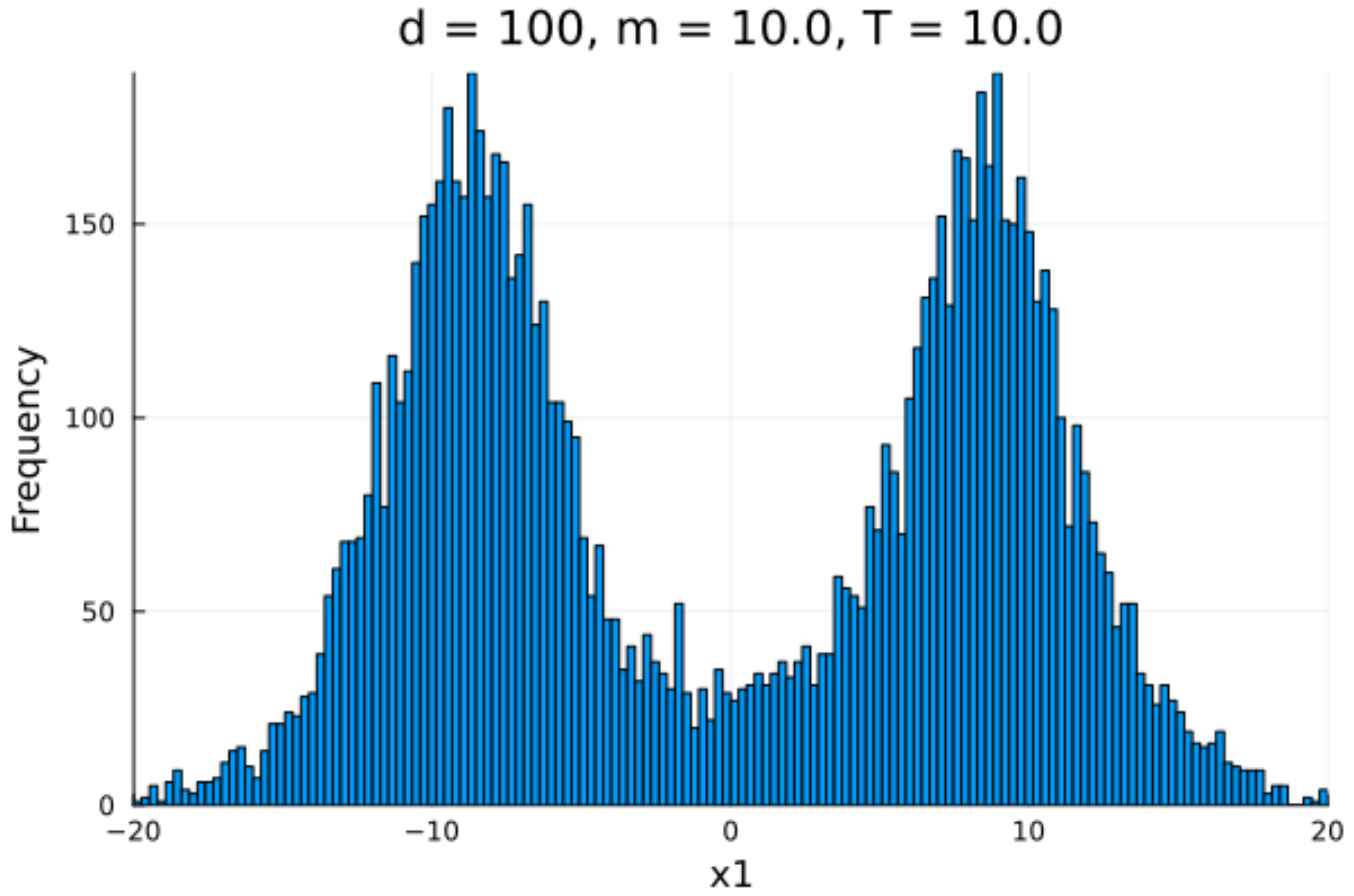
Can RFNN learn a mixture?

It depends on the Gaussians separation!

If $\| \boldsymbol{m} \| \sim \mathcal{O}(1)$ the random features score is approximately linear, so the backward process can at most produce a gaussian distribution!



If $\| \boldsymbol{m} \| \sim \mathcal{O}(\sqrt{d})$ the random features score is non linear, and we can sample effectively from a mixture.



Conclusions

The behavior of the dominant eigenvalue reveals the point (in terms of t and τ) at which the diffusion process becomes sensitive to the informative direction.

- ***Non-separated regime***: the process cannot recover the bimodal structure and diffusion collapses to a single component whose variance aligns with the signal direction

- ***Well-separated regime***: the process recovers the multimodal structure and separates the modes