Memorization as Generalization
in
Physics-inspired Generative Models

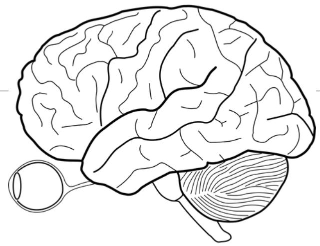
Enrico Ventura
Department of Computing Sciences, Bocconi

Bocconi

LPTMC – Sorbonne
February 4th 2025

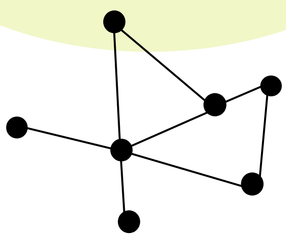
Research themes

Physical Models of Learning



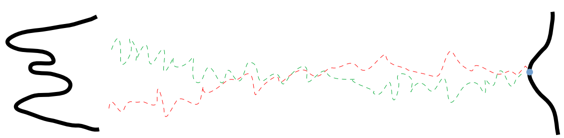
Biologically inspired learning systems

Interacting neurons with a complex dynamics



Artificial learning systems

Physical model with no biological constraints



Generalization

Learning the distribution of the data and not the data.

Memorization

Learning the data.

Outline of the Presentation

The concepts of Memorization and Generalization can be unified inside the same "thermodynamic" picture:

1. Biologically inspired learning systems.

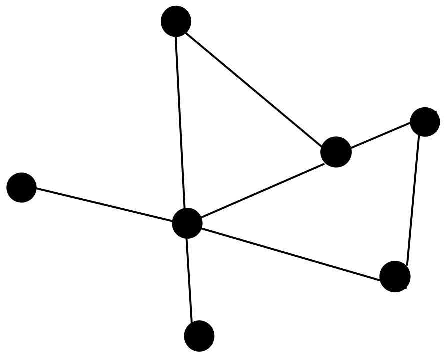
2. Diffusion Models with structured data.

3. Future perspectives.

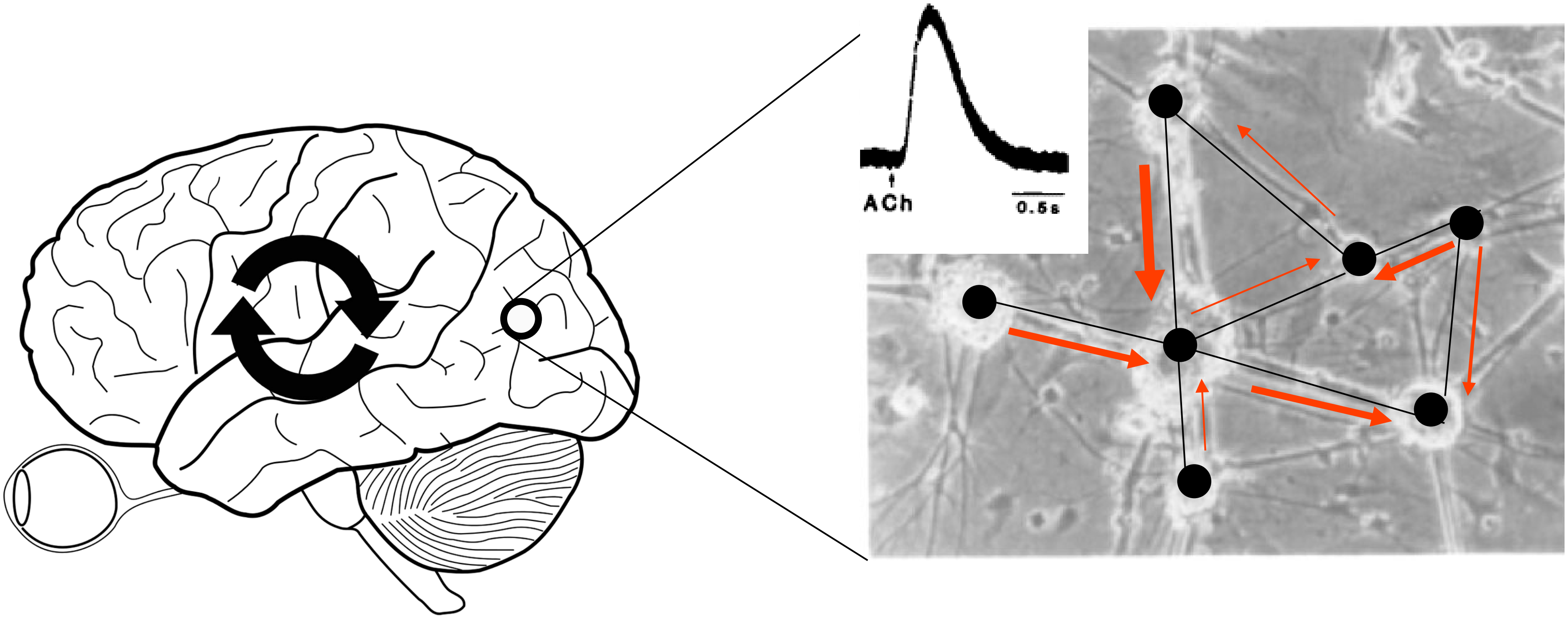
Biologically inspired learning systems

or

Recurrent Neural Networks

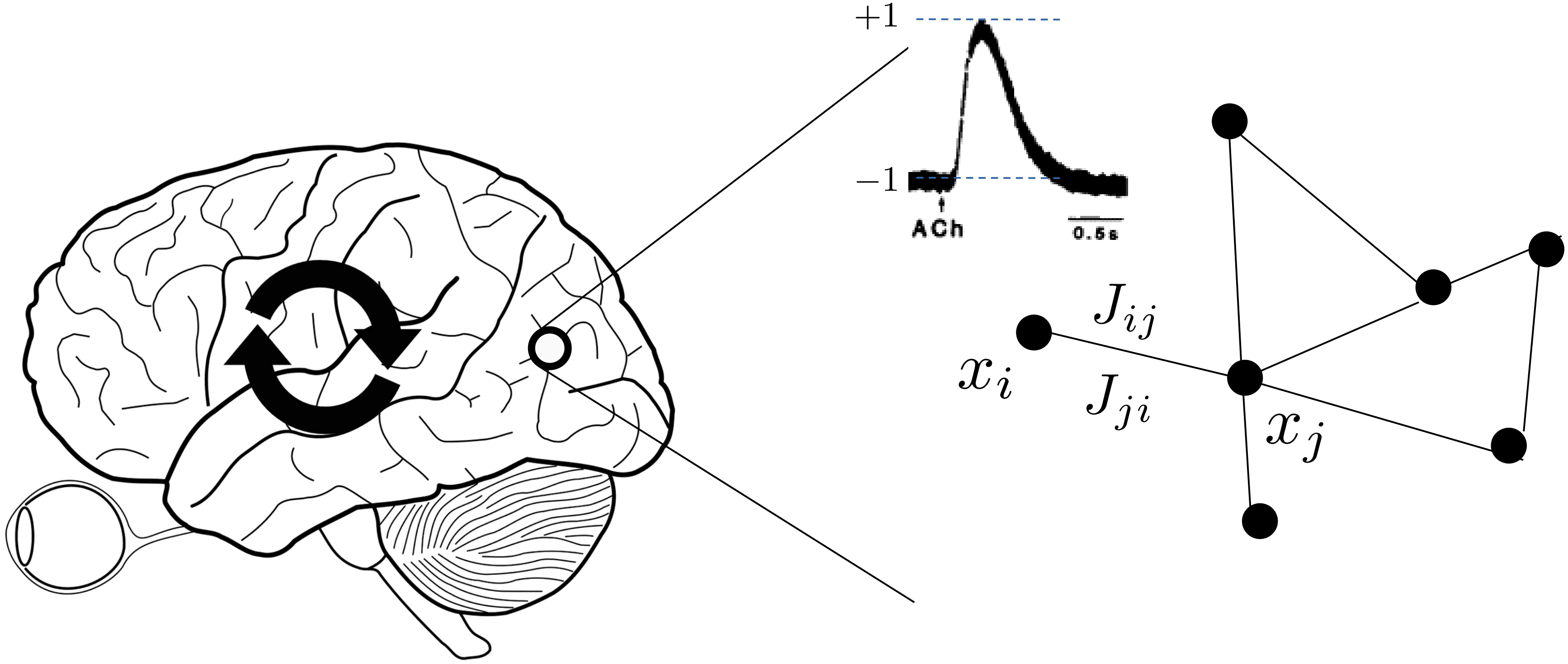


Recurrent Neural Networks



[Wood (1993).]

Recurrent Neural Networks



Complex (thermo)dynamics

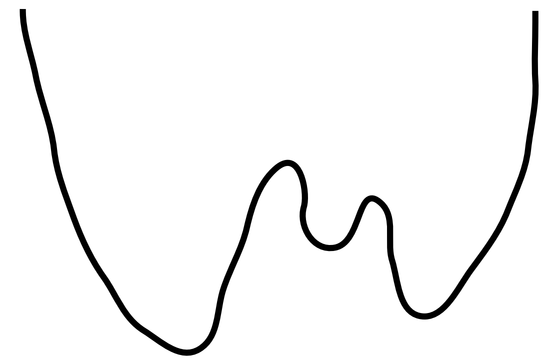
$$E(\boldsymbol{x}|J) = -\sum_{i,j} x_i J_{ij} x_j$$

$$\begin{cases} \boldsymbol{x} \in \{-1,+1\}^N & \text{neural activity} \\ \boldsymbol{J} \in \mathbb{R}^{N\times N} & \text{disorder/frustration} \end{cases}$$

[Pitts & McCulloch (1943).]

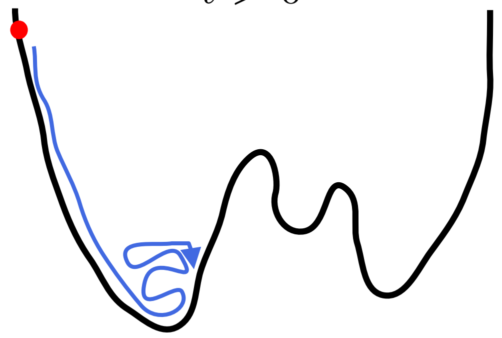
Recurrent Neural Networks

Matrix J_{ij} defines an energy landscape $E(\boldsymbol{x}|J)$

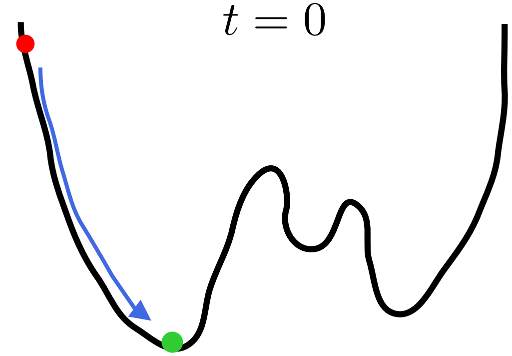


Assuming $J_{ij} = J_{ji}$ the pdf $P_t(\boldsymbol{x}|J) = \frac{1}{Z_t} e^{-\frac{1}{t} E(\boldsymbol{x}|J)}$ samples states from the landscape.

diffuse:
$t > 0$

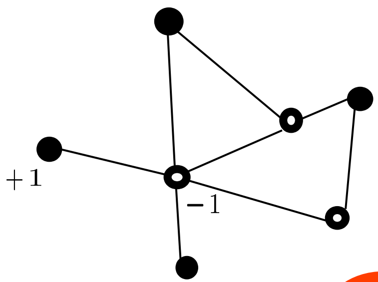


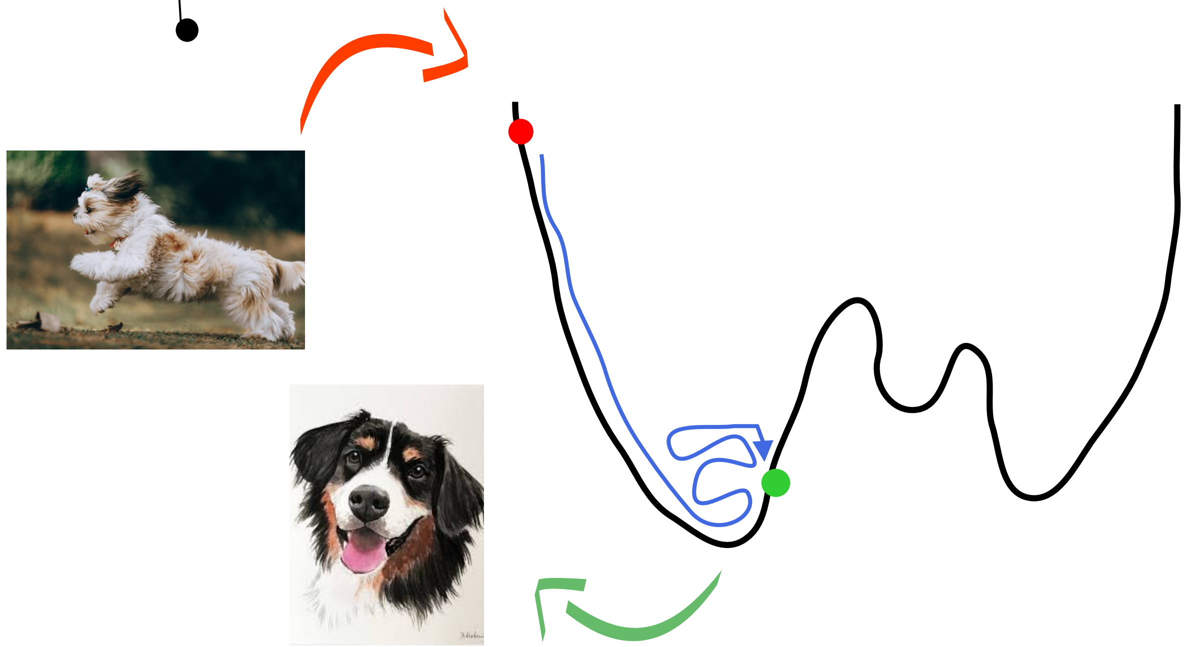
quench:
$t = 0$

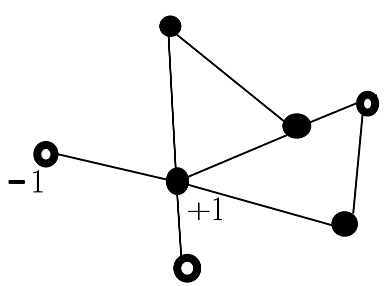


[Peretto (1984)]

Recurrent Neural Networks





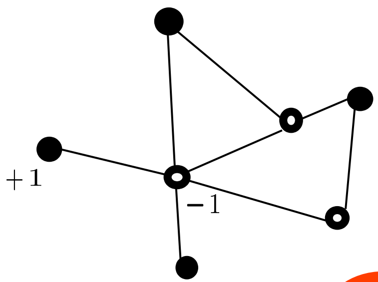


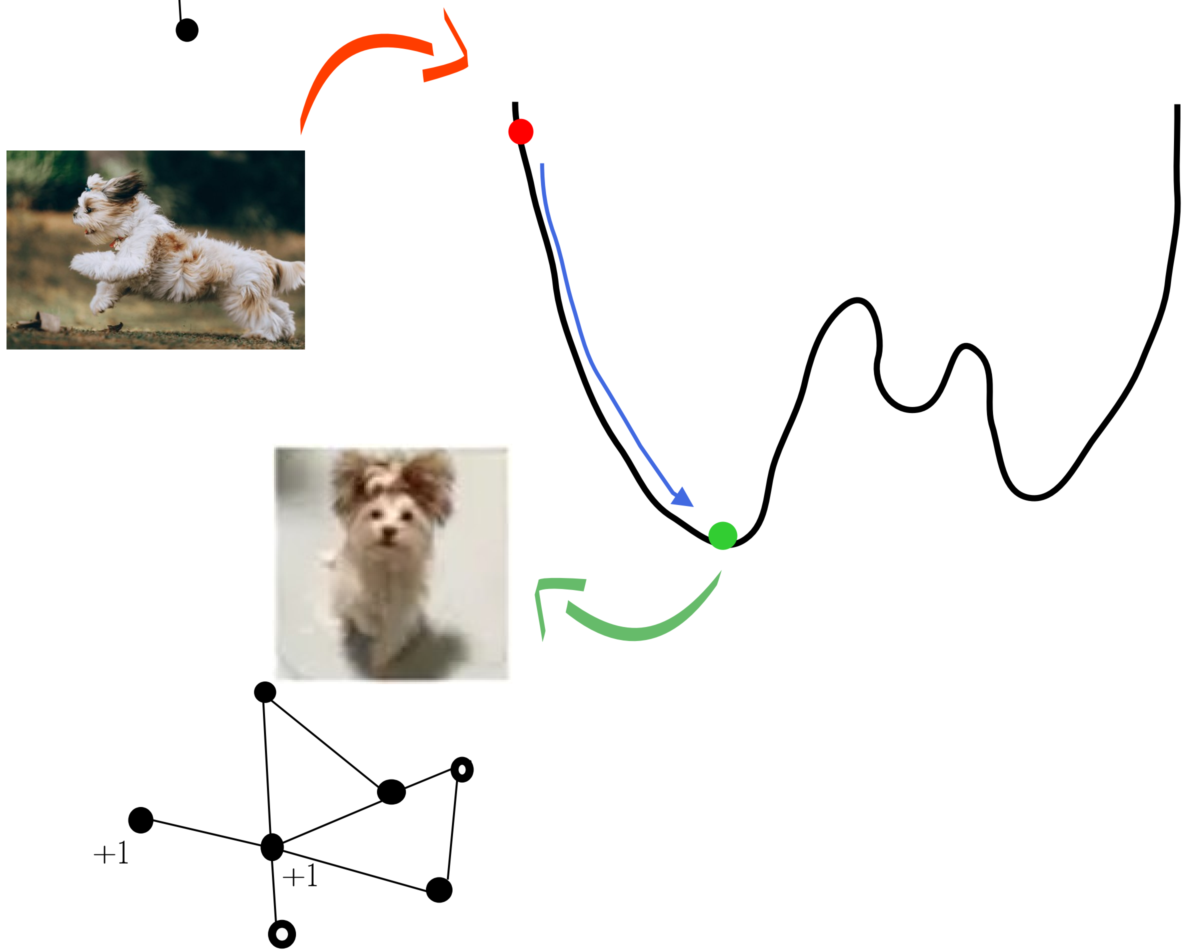
Generalization:

The **dynamics** samples **“new”** examples.

[Hinton et al. (1985)]

Recurrent Neural Networks





Generalization:

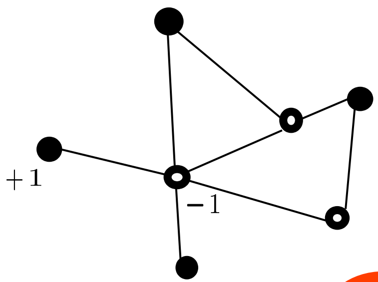
The **dynamics** samples **“new”** examples.

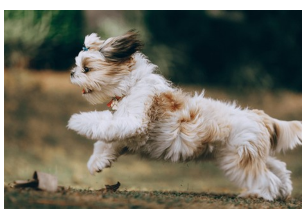
Memorization:

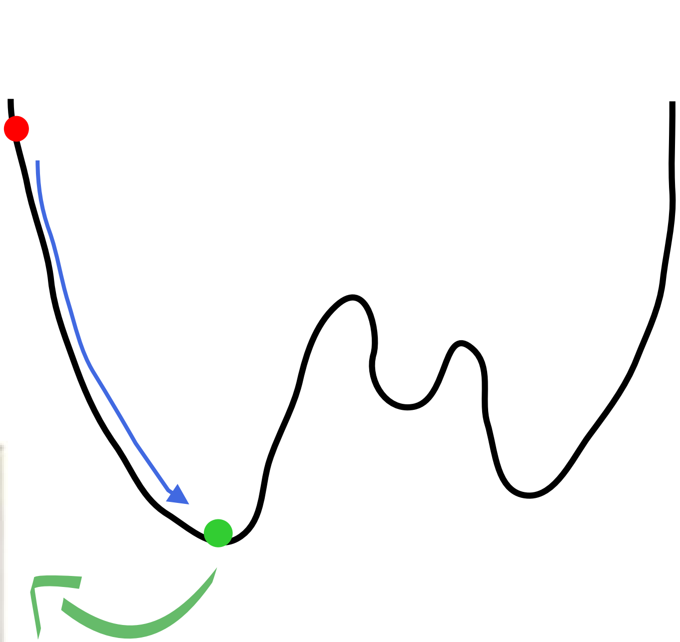
The **dynamics** retrieves **“known”** examples.

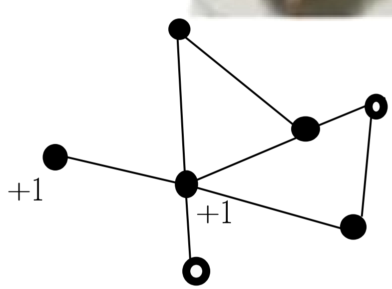
[Hopfield (1982), Gardner et al. (1988-1989), Amit (1990)]

Recurrent Neural Networks









Generalization:

The **dynamics** samples **“new”** examples.

Memorization:

The **dynamics** retrieves **“known”** examples.

How do we find J_{ij} **?**

Recurrent Neural Networks

We can derive the optimal J_{ij} for both generalizing and memorizing using **one single learning algorithm.**

[Ventura et al. (2022), Ventura & Benedetti (2024), Ventura et al. (2024)]

Moment-matching algorithms:

$$P_t(\boldsymbol{x}|J) \approx P_{data}(\boldsymbol{x})$$

$$\dot{J}_{ij} = \langle x_i x_j \rangle_{data} - \langle x_i x_j \rangle_t$$ **Generalization**

t

$$\dot{J}_{ij} = \langle x_i x_j \rangle_{data} - \langle x_i x_j \rangle_{t=0}$$ **Memorization**

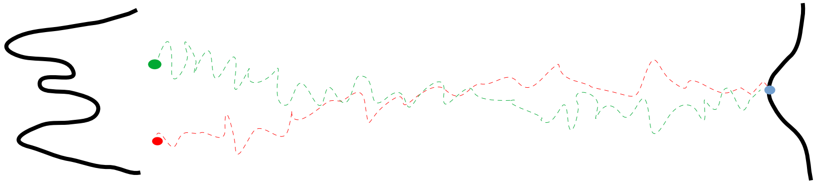
$t = 0$

We just need to change the **"temperature of learning"**.

Artificial
learning systems

or

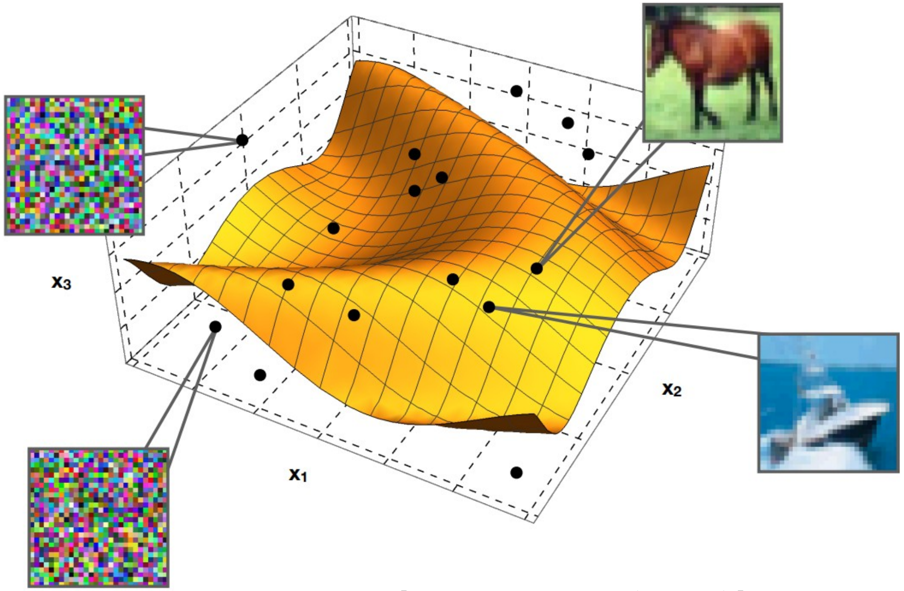
Diffusion Models with Structured Data



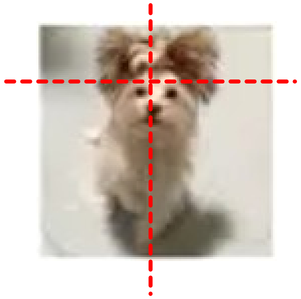
The Manifold Hypothesis

Data live in a **space of dimension** N.

Data contain symmetries and correlations.



[Goldt et al. (2019)]

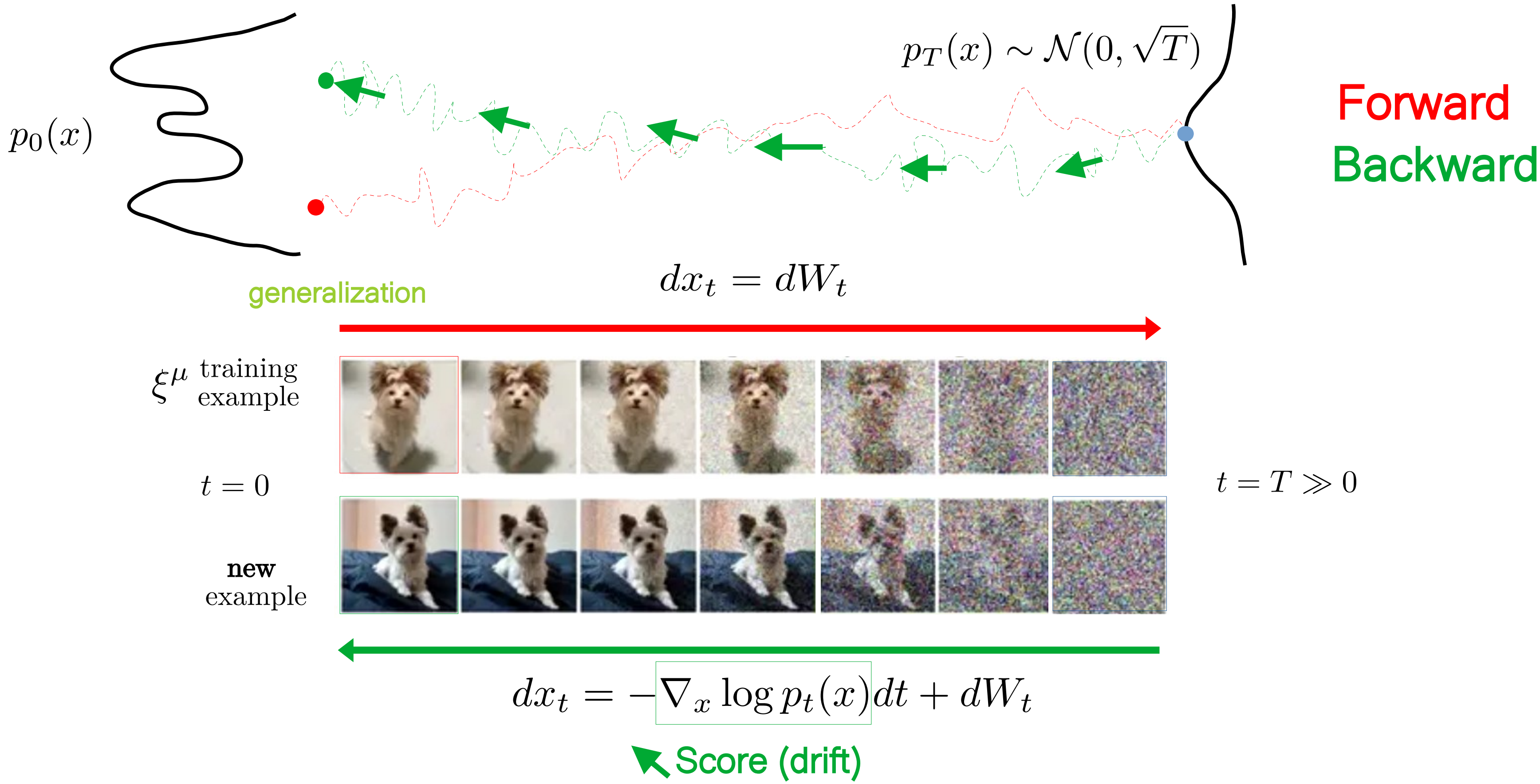


Manifold Hypothesis
[Peyré (2009), Fefferman et al. (2016)]

Data live on a hidden low-dimensional manifold.

How does this affect learning?

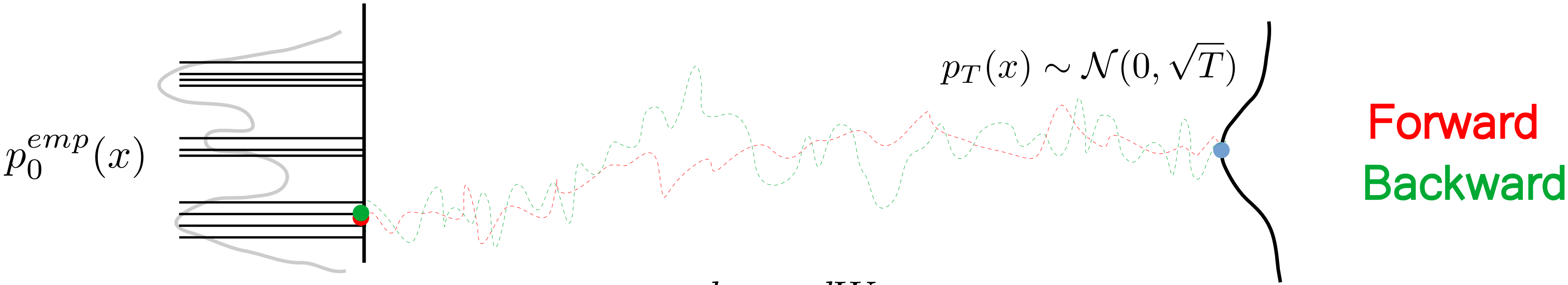
Diffusion Models



$p_0(x)$ is **not known** in real applications $\rightarrow$ I don't know the exact Score $\rightarrow$ I use **Machine Learning**.

[Anderson (1982), Sohl-Dickstein et al. (2015), Yang et al. (2024)]

Diffusion Models



memorization

$$dx_t = dW_t$$

$t = 0$

$t = T \gg 0$

$$dx_t = -\boxed{\nabla_x \log p_t(x)}dt + dW_t$$

$$p_0^{emp}(x) = \frac{1}{P} \sum_{\mu=1}^{P} \delta\left(x - \xi^\mu\right) \quad \text{implies an exact pdf} \quad p_t^{emp}(x) = \frac{1}{P\sqrt{2\pi t}^N} \sum_{\mu=1}^{P} e^{-\frac{1}{2t}(x-\xi^\mu)^2}$$

The **empirical** diffusion model **memorizes** the training examples.

[Biroli et al. (2024), Raya et al. (2024)]

Diffusion Models

Questions:

1. How is **memorization** affected by the structure of the data?

[Ventura et al. (2024), Achilli et al. (in preparation)]

2. Does the *empirical* diffusion model display any **generalization**?
   How is this property affected by the structure of the data?

[Ventura et al. (2025) accepted to ICLR'25, Achilli et al. (in preparation)]

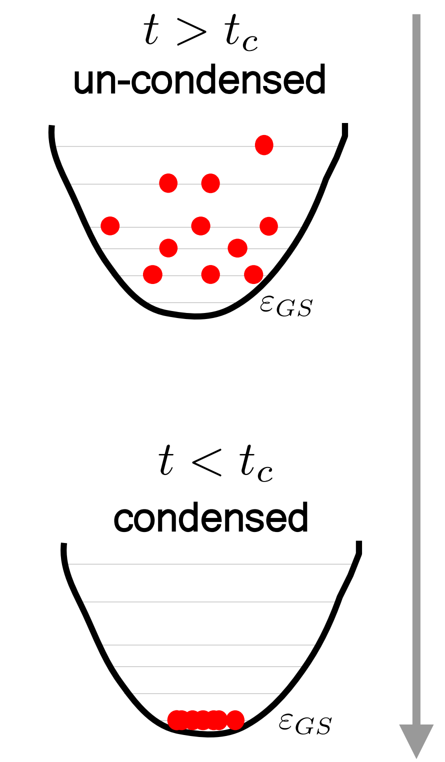
REM formalism for Diffusion Models

Diffusion Models can be mapped into **Random Energy Models (REM)**.

System with N degrees of freedom can assume $P = e^{\alpha N}$ **energy levels** $\{\varepsilon^\mu\}_{\mu=1}^{P}$ and $\varepsilon^\mu \sim p_\varepsilon$ i.i.d.

$$Z_t = \sum_{\mu=1}^{P} e^{\frac{N}{t}\varepsilon^\mu}$$ **partition function**

$$\phi_\alpha(t) = \lim_{N\to\infty} \frac{t}{N}\mathbb{E}_\varepsilon \log \sum_\mu e^{\frac{N}{t}\varepsilon^\mu}$$ **free-energy function**



[Derrida (1981), Biroli et al. (2024), Biroli & Mézard (2024), Lucibello & Mézard (2024)]

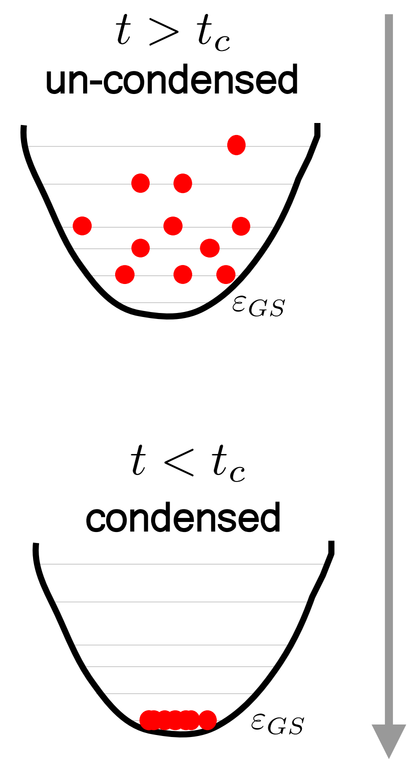
REM formalism for Diffusion Models

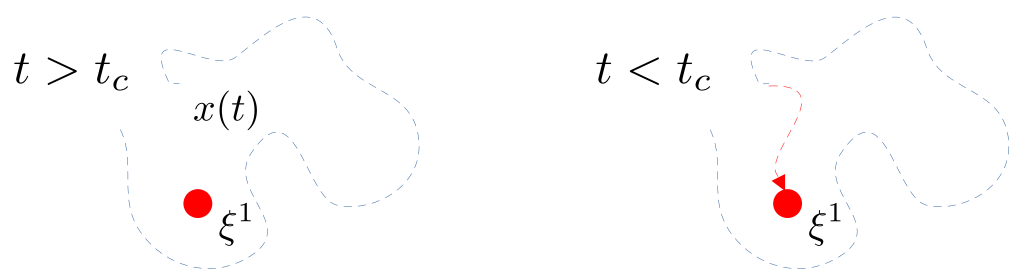
Diffusion Models can be mapped into **Random Energy Models (REM)**.

System with N degrees of freedom can assume $P = e^{\alpha N}$ **energy levels** $\{\varepsilon^\mu\}_{\mu=1}^P$ and $\varepsilon^\mu \sim p_\varepsilon$ i.i.d.

$$Z_t = \sum_{\mu=1}^{P} e^{\frac{N}{t}\varepsilon^\mu} \qquad \textbf{partition function}$$

$$\phi_\alpha(t) = \lim_{N\to\infty} \frac{t}{N} \mathbb{E}_\varepsilon \log \sum_\mu e^{\frac{N}{t}\varepsilon^\mu} \qquad \textbf{free-energy function}$$

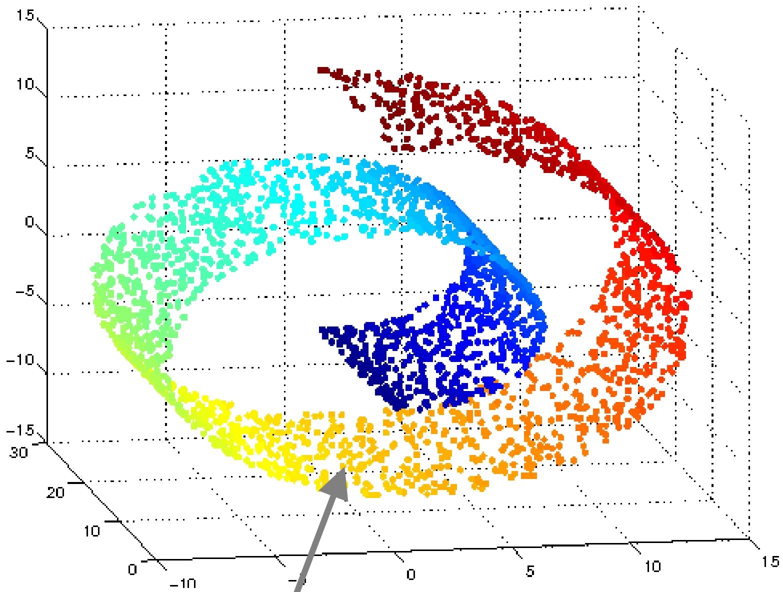




$p_t^{emp} \approx \phi_\alpha(t)$ and **condensation in REM is memorization.**

[Derrida (1981), Biroli et al. (2024), Biroli & Mézard (2024), Lucibello & Mézard (2024)]

Modeling the Manifold Hypothesis



$$p_0(x) = \int Dz\delta\left(x - \sigma(Fz)\right)$$

<u>The Hidden Manifold "recipe":</u>

$z^\mu \in \mathbb{R}^D \quad z_i^\mu \sim \mathcal{N}(0,1)$ **latent**

$$\downarrow$$

$$F \in \mathbb{R}^{N\times D} \quad F_{ij} \sim \mathcal{N}(0, 1/\sqrt{D})$$

$$Fz^\mu \in \mathbb{R}^N$$

$$\downarrow$$

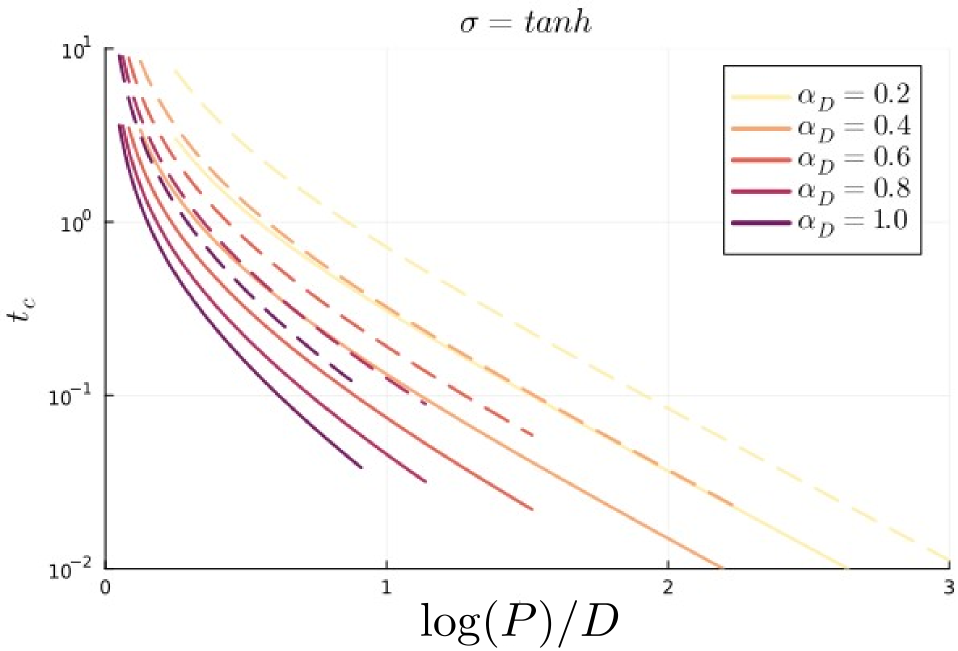
$\boxed{\xi^\mu = \sigma\left(Fz^\mu\right)}$ **visible**

[Mei & Montanari (2019), Goldt et al. (2019), Gerace et al. (2020)]

Diffusion Models under the Manifold Hypothesis

We use REM formalism and compute the condensation/memorization time for structured data.

$$t_c = \mathcal{O}\left(e^{-\frac{2\log P}{D}}\right)$$

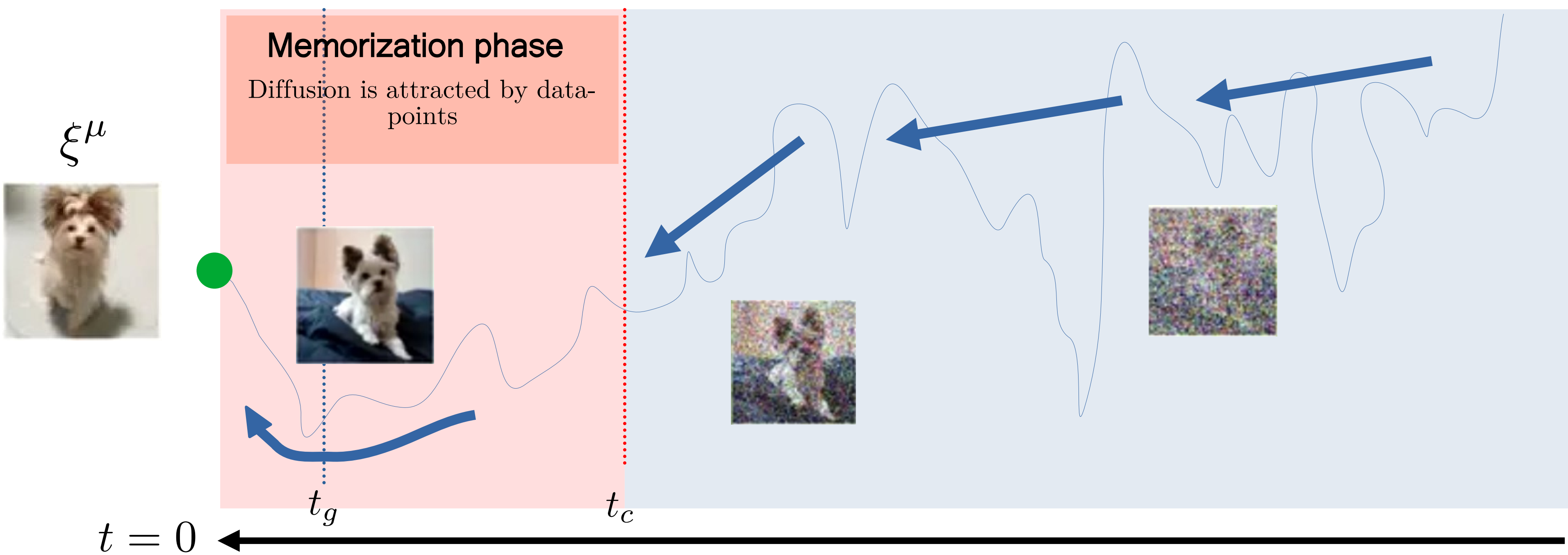


The model benefits from data structure because **memorization is "delayed".**

Diffusion Models under the Manifold Hypothesis

We study generalization also through a **REM approach.**

$t_g = \mathrm{argmin}_t D_{KL}\left(p_0 \| p_t^{emp}\right) = \mathrm{argmin}_t \left[\Phi_t\right]$ with Φ_t a **REM free-energy** function to minimize.



Generalization occurs while memorizing.

Take-home messages

1. In both recurrent neural networks and diffusion models, we can pass from memorization to generalization by changing the **“temperature of learning”**.

2. Structure helps learning in Diffusion Models.
(It is not clear if this holds in recurrent neural networks [Negri et al. (2023)]).

Future Projects: short-term

Biologically-inspired learning systems

Using moment-matching algorithms to solve a non rotationally-invariant extensive-rank **matrix factorization** problem [J. Barbier et al. (2024)] .

[E. Ventura (in preparation)]

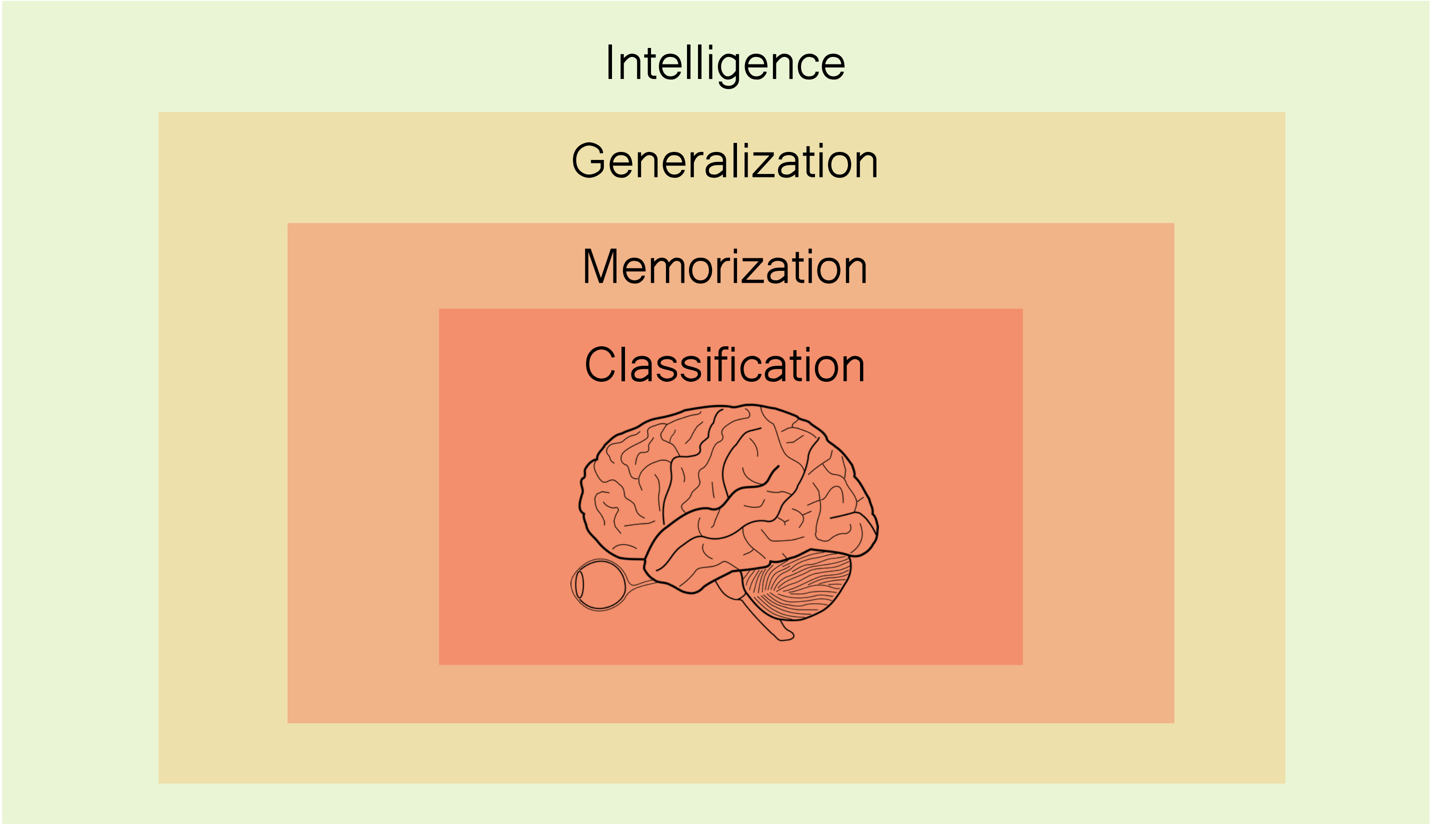
Artificial learning systems

Studying the way **artificial neural-networks fit the data manifold** during the backward stochastic process via Approximate Message Passing tools.

[Work in collaboration with B. Achilli, M. Mézard and C. Lucibello.]

Future Projects: long-term

Unifying the concepts of generalization, memorization and classification in learning systems inside a statistical physics framework.



Thank you!

Back–Up Slides

STUDIES

(physics, specialized in statistical mechanics)

RESEARCH

Bachelor's in Physics
(La Sapienza)

Dissertation:
"Ising Model and Numerical Simulations"

Supervised by G. Parisi.

Master in Th. Physics
(La Sapienza + Erasmus)

Master Thesis:
"Memory Storage and Retrieval in Sparsely Connected Balanced Networks"

Supervised by
G. Mongillo and G. Ruocco.

PhD in Physics (Cotutelle)
(La Sapienza & ENS-PSL)

PhD Thesis:
"Demolition and Reinforcement of Memories in Spin-Glass like Neural Networks"

Supervised by F. Zamponi (Ex-ENS) and G. Ruocco (La Sapienza).

Post-Doc
(Bocconi University)

Supervised by C. Lucibello.

2015 2018 2020 2023

ENS

STUDIES

(physics, specialized in statistical mechanics)

RESEARCH

Bachelor's in Physics
(La Sapienza)

Dissertation:
"Ising Model and
Numerical Simulations"

Supervised by G. Parisi.

Master in Th. Physics
(La Sapienza + Erasmus)

Master Thesis:
"Memory Storage and Retrieval
in Sparsely Connected
Balanced Networks"

Supervised by
G. Mongillo and G. Ruocco .

PhD in Physics (Cotutelle)
(La Sapienza & ENS-PSL)

Teaching:

Mission D'Einsegnement de l'ENS
E-OGS examination
(Gendermerie Nationale).

(29 hours).

Post-Doc
(Bocconi University)

Teaching:

TA to the course:
"Complex Systems
and Physical Models"

(23 hours).

2015

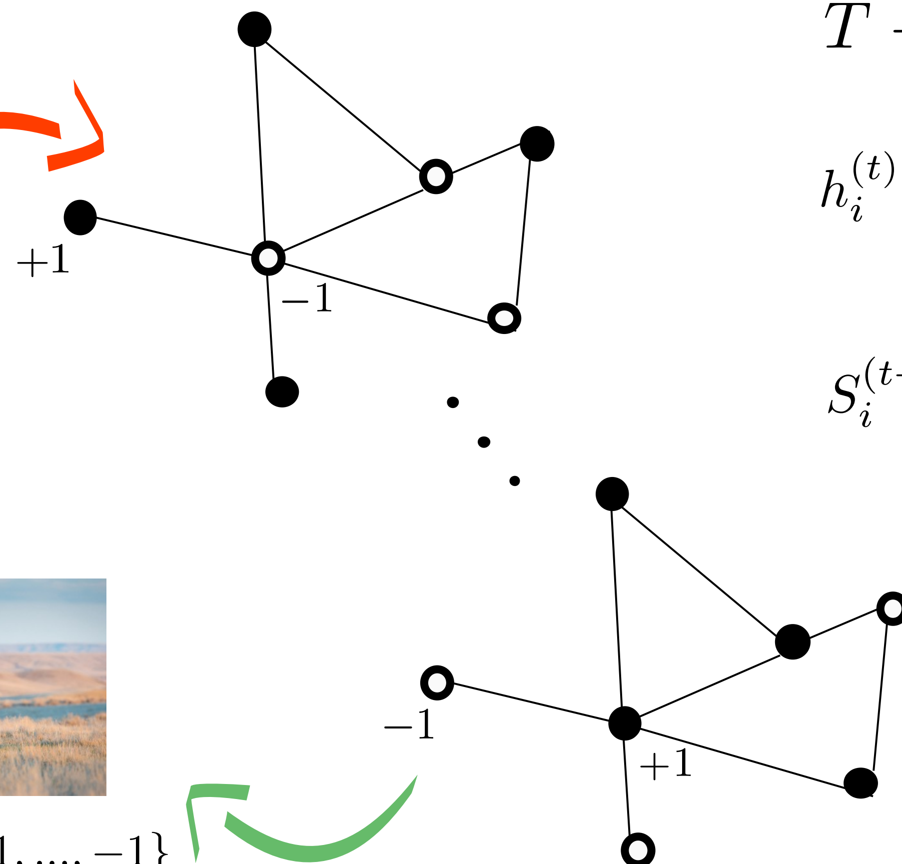
2018

2020

2023

Recurrent Neural Networks

$\vec{S}_{in} = \{+1, -1, ..., +1\}$



$\vec{S}_{out} = \{+1, +1, ..., -1\}$

$$T \to 0$$

$$h_i^{(t)} = \sum_{j \neq i} J_{ij} S_j^{(t)}$$

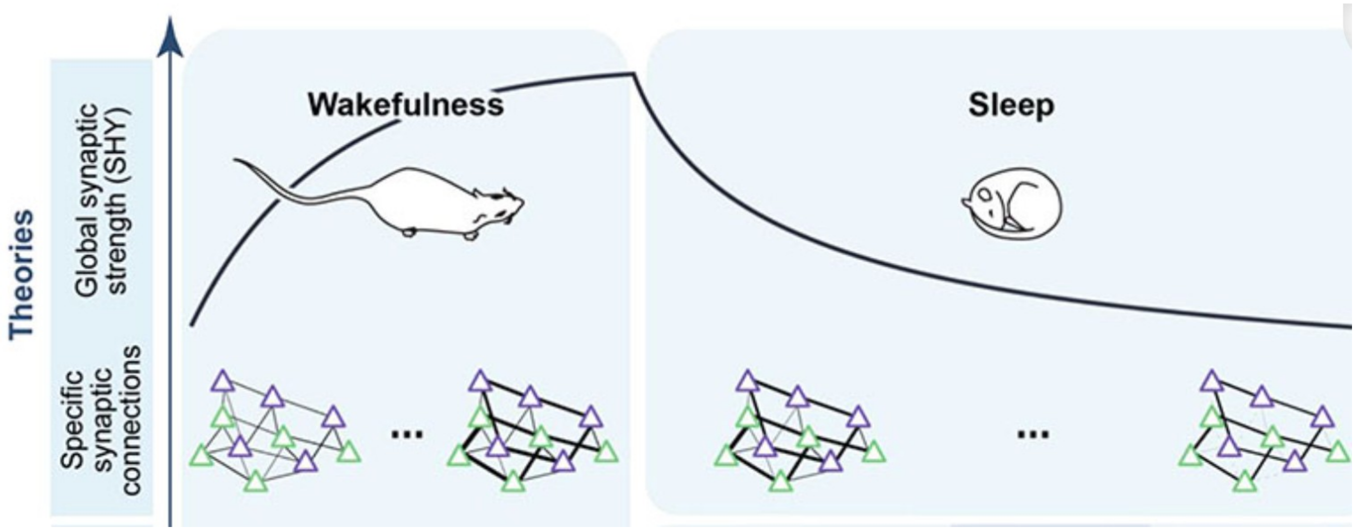
$$S_i^{(t+1)} = \text{sign}\left(h_i^{(t)}\right)$$

Deterministic dynamics

[Peretto, 1984; Amit, 1989]

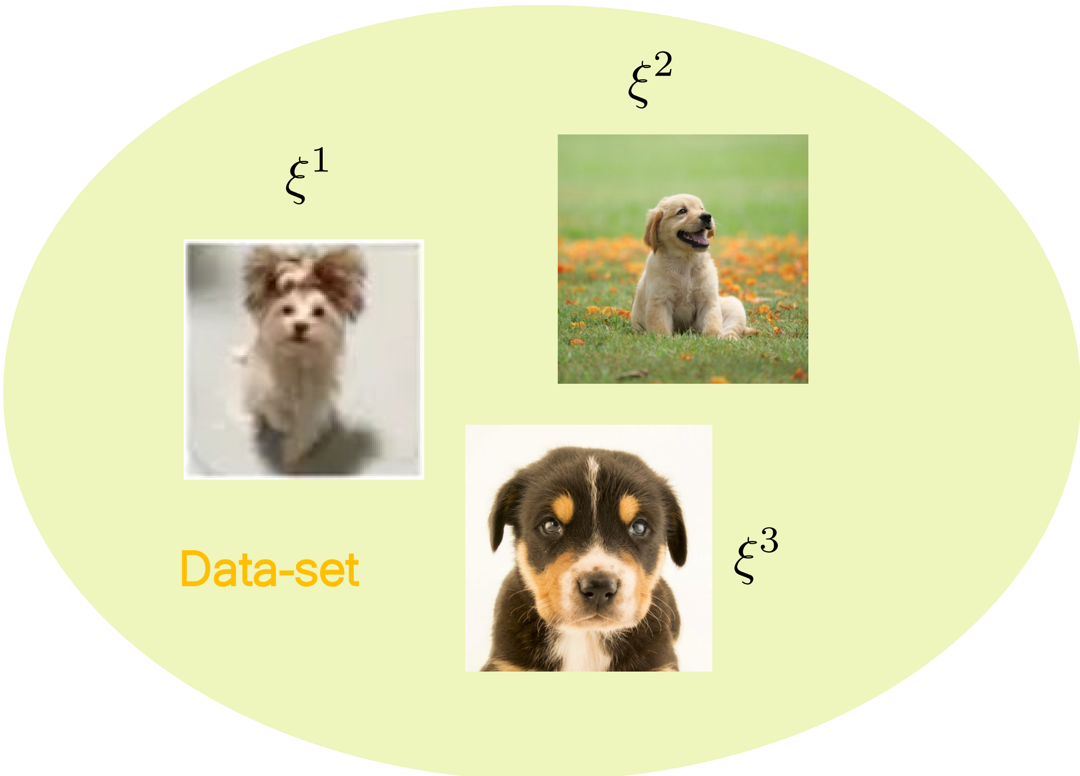
Recurrent Neural Networks

$$\langle S_i S_j \rangle_{data} \qquad\qquad -\langle S_i S_j \rangle_t$$



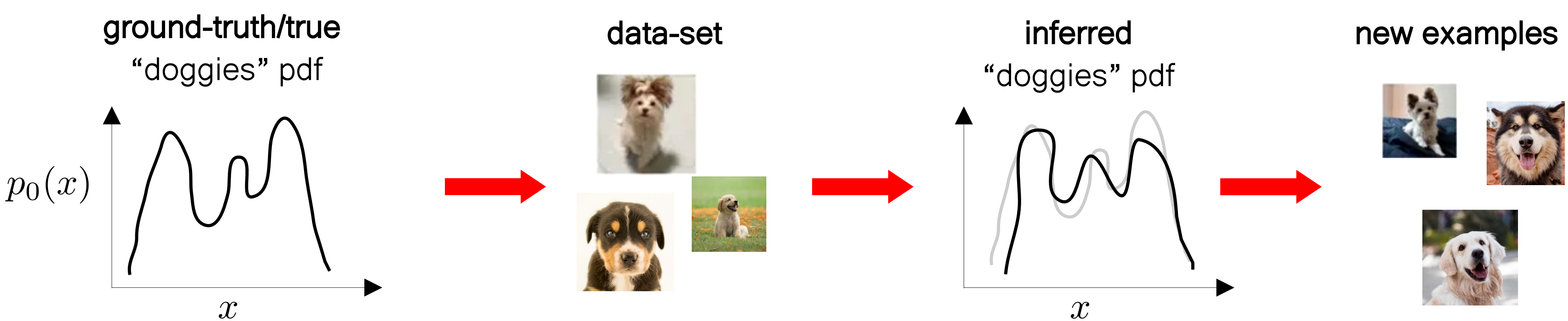
[Girardeau et al. (2020), Hoel (2021)]

Diffusion Models



$$\xi^\mu \in \mathbb{R}^N \qquad \mu = 1, .., P$$

$$P = e^{\alpha N}$$

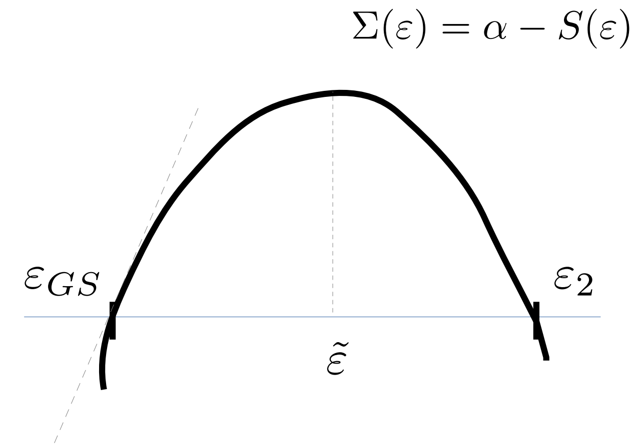


REM formalism

Diffusion Models can be mapped into **Random Energy Models (REM)**.

System with N degrees of freedom can assume $P = e^{\alpha N}$ **energy levels** $\{\varepsilon^\mu\}_{\mu=1}^P$ and $\varepsilon^\mu \sim p_\varepsilon$ i.i.d.

$$p_\varepsilon = \int_{\varepsilon_{GS}}^{\varepsilon_2} e^{-Ns(\varepsilon)} d\varepsilon \approx e^{-NS(\tilde{\varepsilon})}$$ concentrates when $N \to \infty$



$$\Sigma_t = \log \int_{\varepsilon_{GS}}^{\varepsilon_2} e^{N(\alpha - S(\varepsilon))} d\varepsilon$$ **entropy**

$$Z_t = \sum_{\mu=1}^{P} e^{\frac{N}{t}\varepsilon^\mu} \approx e^{N(\alpha - S(\tilde{\varepsilon}) + \frac{\tilde{\varepsilon}}{t})}$$ **partition function**

$$\phi_\alpha(t) = \lim_{N\to\infty} \frac{t}{N} \mathbb{E}_\varepsilon \log \sum_\mu e^{\frac{N}{t}\varepsilon^\mu} = \begin{cases} \tilde{\varepsilon} + t(\alpha - S(\tilde{\varepsilon})) & (t > t_c) \\ \varepsilon_{GS} & (t < t_c) \end{cases}$$ **free-energy function**

[Derrida (1981)]

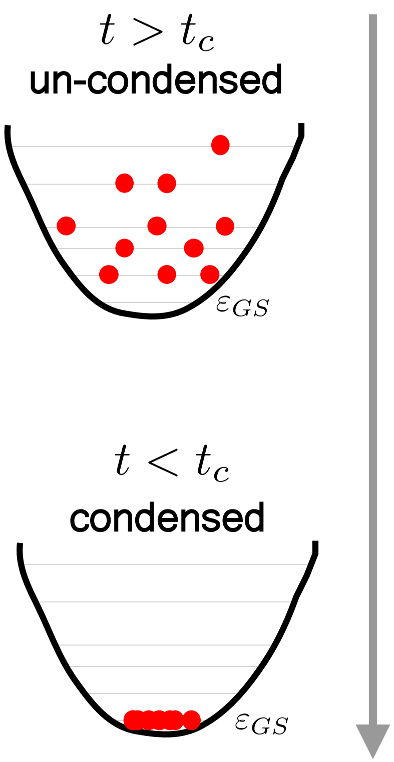
REM formalism

Diffusion Models can be mapped into **Random Energy Models (REM)**.

System with N degrees of freedom can assume $P = e^{\alpha N}$ **energy levels** $\{\varepsilon^\mu\}_{\mu=1}^P$ and $\varepsilon^\mu \sim p_\varepsilon$ i.i.d.

$$Z_t = \sum_{\mu=1}^{P} e^{\frac{N}{t}\varepsilon^\mu}$$ **partition function**

$$\phi_\alpha(t) = \lim_{N\to\infty} \frac{t}{N} \mathbb{E}_\varepsilon \log \sum_\mu e^{\frac{N}{t}\varepsilon^\mu}$$ **free-energy function**



$$p_t^{emp}(x = \xi^1 + \omega\sqrt{t}) = \frac{1}{P\sqrt{2\pi t}^N}\left(e^{-\frac{\|\omega\|^2}{2}} + \sum_{\mu\geq 2} e^{-\frac{1}{2t}\|(\xi^1-\xi^\mu)+\omega\sqrt{t}\|^2}\right)$$

$$= \frac{1}{P\sqrt{2\pi t}}\left(Z_1 + Z_{2,\ldots,P}\right)$$

Condensation when Z_1 is sub-leading with respect to $Z_{2,..,P}$

[Derrida (1981), Biroli et al. (2024), Biroli & Mézard (2025), Lucibello & Mézard (2024)]

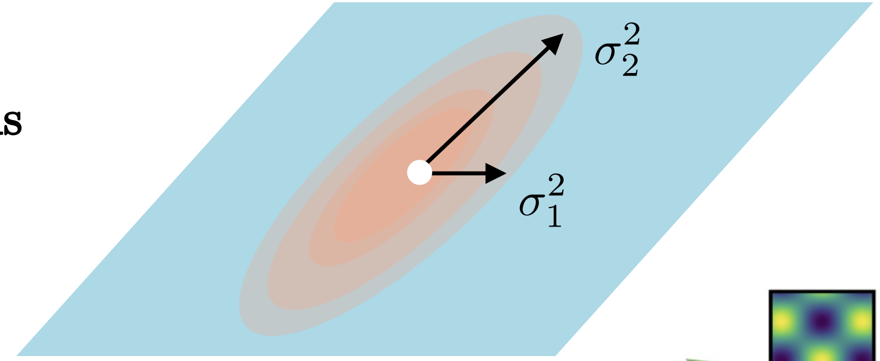
Diffusion Models under the Manifold Hypothesis

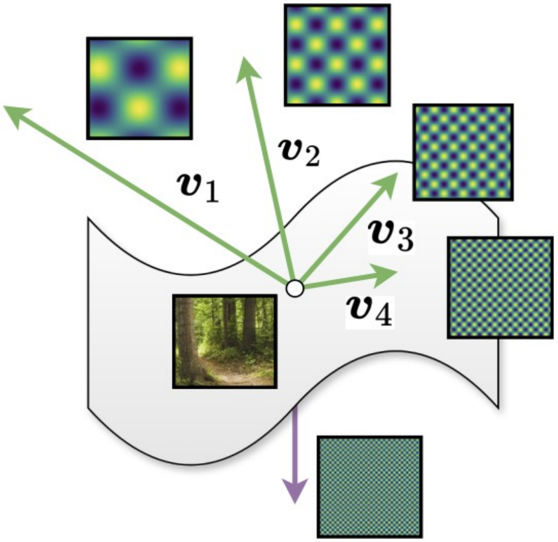
Random Matrix Approach $(t > t_c)$

$$F_{ij} \sim \mathcal{N}(0, \sigma_j^2)$$

sub-manifolds

$$\xi^\mu = \boldsymbol{F} z^\mu$$





We are interested in computing the eigenspectrum of the Jacobian of the Score function, because

$$\vec{S}(x + dx, t) \approx \vec{S}(x, t) + \boldsymbol{J}(t) \cdot x$$

Gaps in the eigenspectrum reveal **forbidden diffusive directions**.

$$\boldsymbol{J}(t) = \frac{1}{t} \boldsymbol{F} \left[I_D + \frac{1}{t} \boldsymbol{F}^\top \boldsymbol{F} \right]^{-1} \boldsymbol{F}^\top - I_N$$

Diffusion Models under the Manifold Hypothesis

Random Matrix Approach $(t > t_c)$

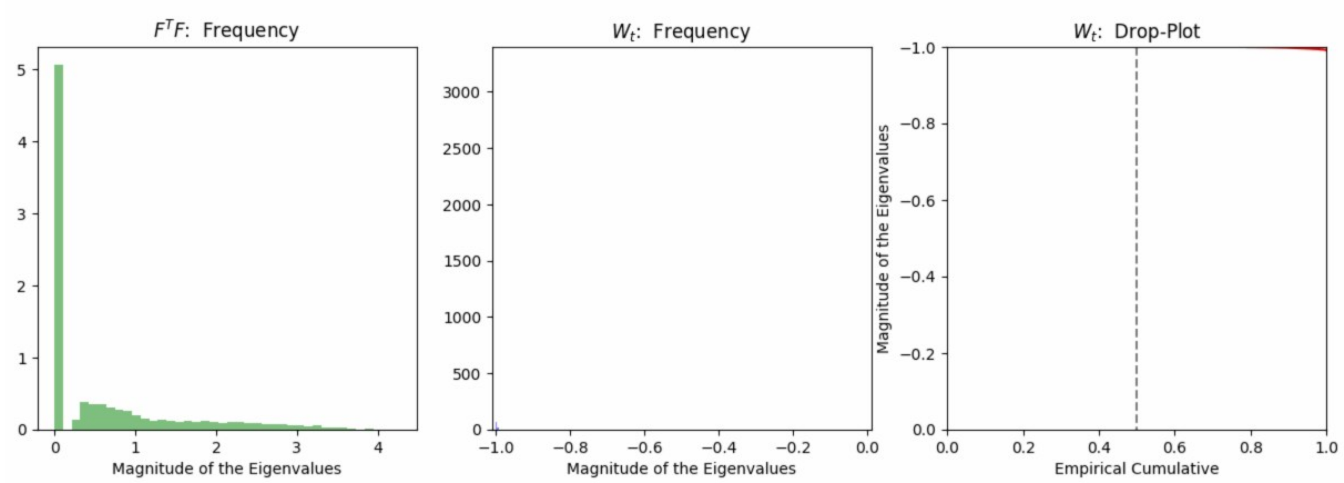
Matrix $$A = \frac{1}{N} F F^T$$

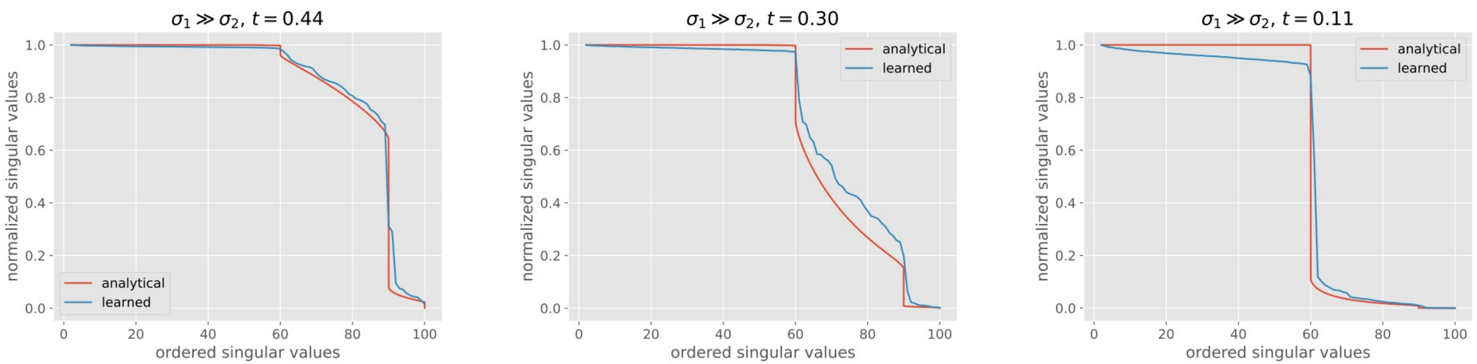
Steltjes transform

$$\begin{aligned} \mathbb{E}\left[g_A(z)\right] &= -\frac{2}{N}\frac{\partial}{\partial z}\mathbb{E}\left[\log \frac{1}{\sqrt{\det\left(zI_N - A\right)}}\right] \\ &= -\frac{2}{N}\frac{\partial}{\partial z}\lim_{n\to 0}\mathbb{E}\left[\frac{Z^n - 1}{n}\right] \end{aligned}$$

Diffusion Models under the Manifold Hypothesis

Random Matrix Approach $(t > t_c)$



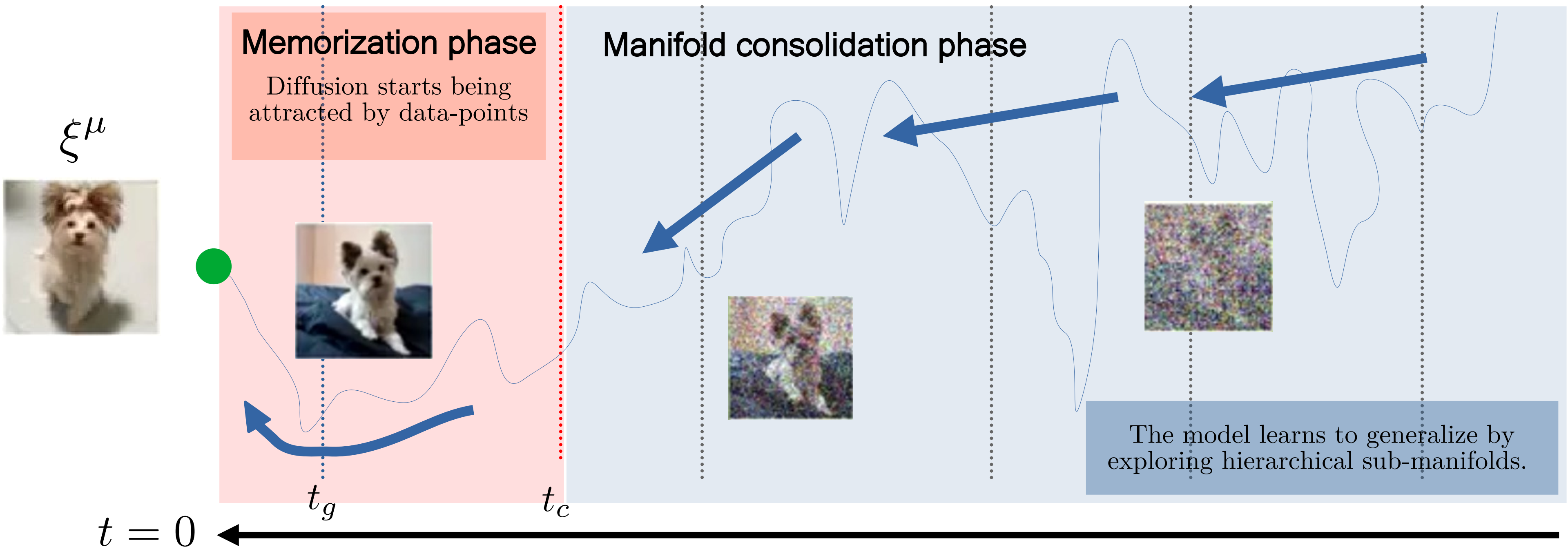


Diffusion Models under the Manifold Hypothesis

We study generalization through two analytical approaches:

1. Random Matrix study of the **geometry of the score** $\vec{S}(x,t) = -\vec{\nabla}_x \log p_t(x)$ with respect to the manifold.

2. $t_g = \text{argmin}_t D_{KL}\left(p_0 \| p_t^{emp}\right) = \text{argmin}_t \left[\Phi_t\right]$ With Φ_t a **REM free-energy** function to minimize.



Generalization occurs while memorizing and the system benefits from structure.

Diffusion Models under the Manifold Hypothesis

Random Matrix + REM approach:

As sub-manifolds with different variances are progressively reconstructed during the backward process, they are also **memorized with the same ordering** (dynamics of memorization).

- $Z_t^{REM}(x)$ with x arbitrary $\longrightarrow$ $t_c(x)$ memorization depends on topology.

- The eigenspectrum of the Jacobian of the score can be computed inside the memorization phase

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \approx \frac{1}{\tilde{N}_t(\boldsymbol{x})} \sum_{\mu=1}^{\tilde{N}_t(\boldsymbol{x})} \left(\boldsymbol{y}^\mu - \boldsymbol{x}\right)/t$$

Diffusion Models under the Manifold Hypothesis

Random Matrix + REM approach:

As sub-manifolds with different variances are progressively reconstructed during the backward process, they are also **memorized with the same ordering** (dynamics of memorization).

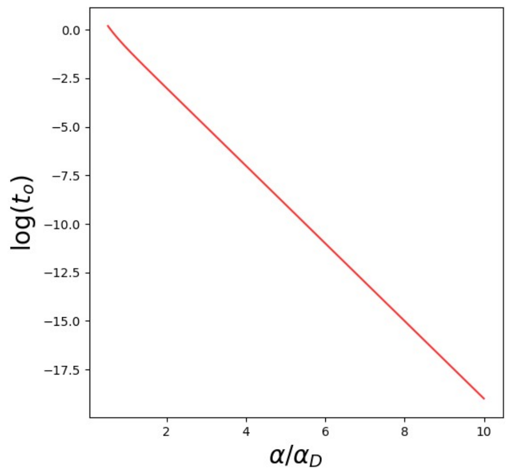
- $Z_t^{REM}(x)$ with x arbitrary $\longrightarrow$ $t_c(x)$ memorization depends on topology.
- The eigenspectrum of the Jacobian of the score can be computed inside the memorization phase

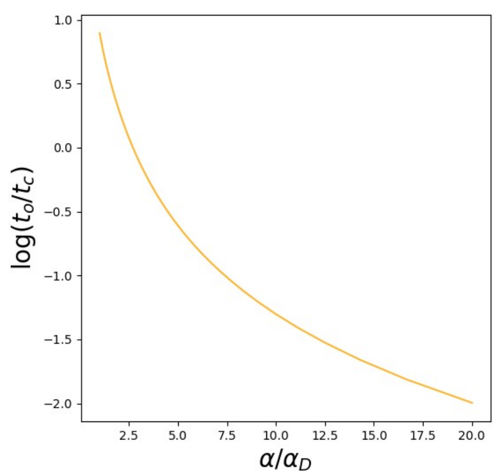
$$J_{ij}(t) \sim \mathcal{N}\left(-\delta_{ij}\left(t+\sigma_i^2\right)^{-1}, \frac{\sigma_i^2}{t\left(t+\sigma_i^2\right)}\left[\phi(t,\mathbf{0})+\phi(t,\boldsymbol{e}_j\cdot\sqrt{t})\right]\right)$$

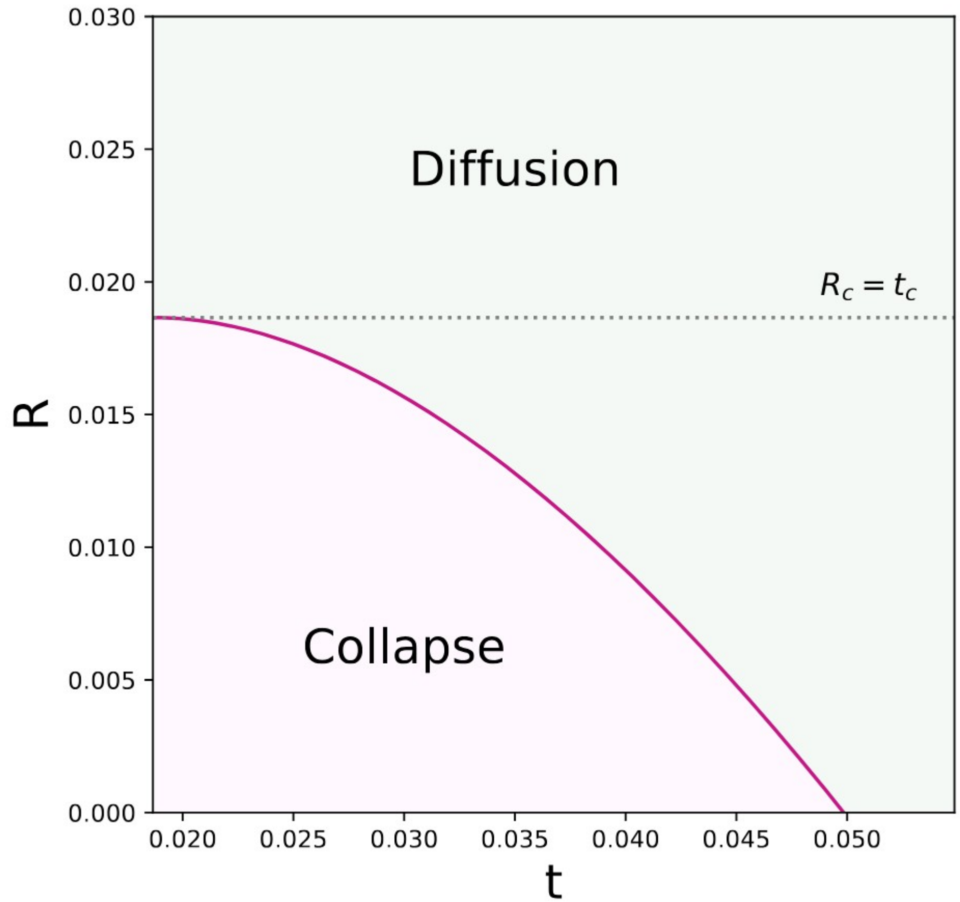
$$\phi(t,\boldsymbol{x}) = \max\left(1/N, t^{-1} - t_c^{-1}(\boldsymbol{x})\right)$$

Emergence of Attractors

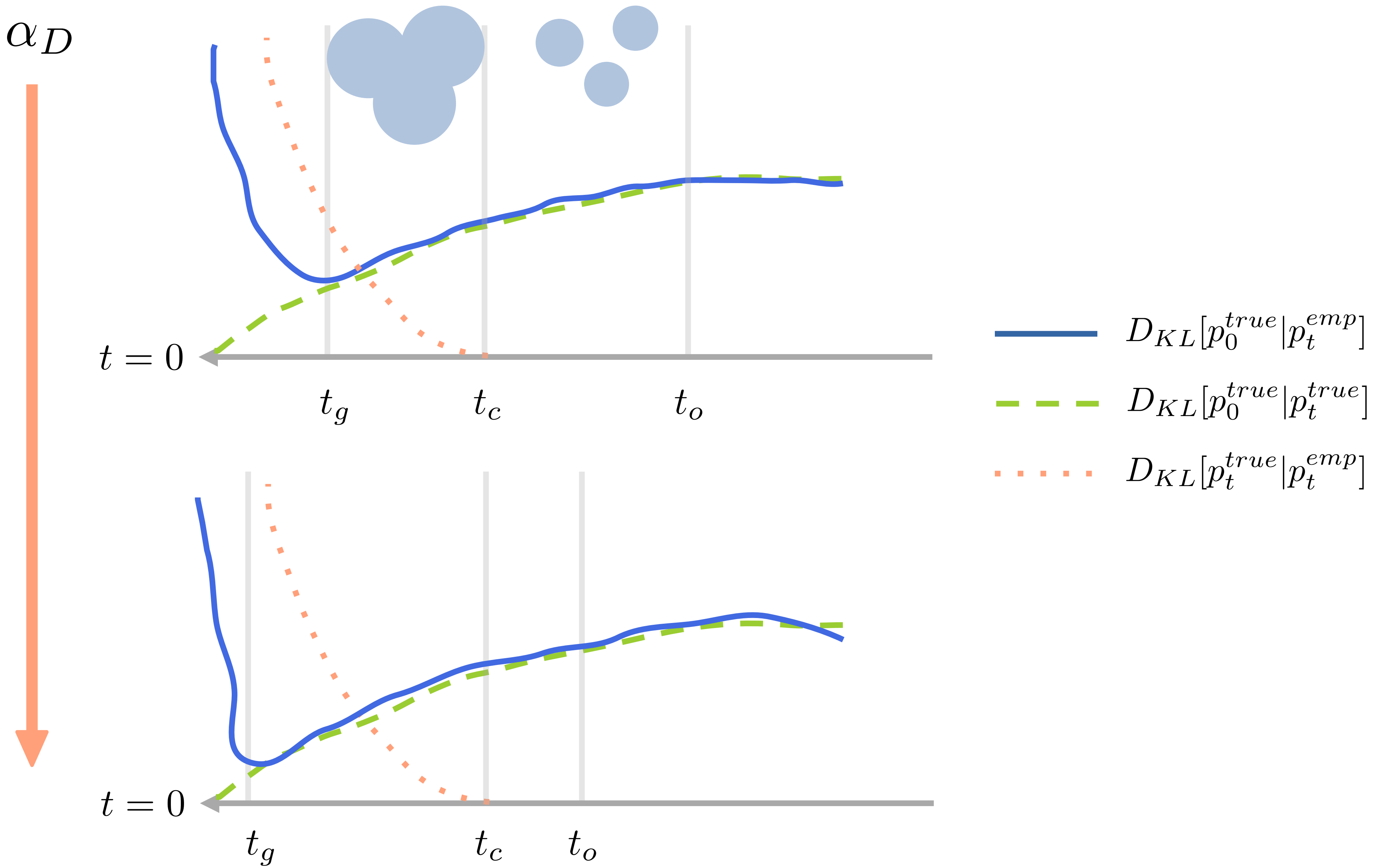
$$\zeta_{t,R}(\lambda) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\xi^1,\omega} \log \mathbb{E}_{\xi} e^{-\frac{\lambda}{2t}\|(\xi^1-\xi)+\omega\sqrt{R}\|^2}$$







Full Picture



Memorization in Moment-Matching Algorithms

$$\dot{J}_{ij} = \langle S_i S_j \rangle_{data} - \langle S_i S_j \rangle_{t=0}$$

$$\langle O(\vec{S}) \rangle_{t=0} = \frac{1}{\Omega} \sum_{\vec{S}^*} O(\vec{S}) \delta(\vec{S} - \vec{S}^*)$$

$\vec{S}^*$ local minima of the energy

Theorem [Ventura (in preparation, 2024)]:

Given $\vec{S} \in \{-1, +1\}^N$ and $\vec{\xi}^\mu \in \{-1, +1\}^N \; \forall \mu$ then $\langle S_i S_j \rangle_{data} = \langle S_i S_j \rangle_{t=0}$ holds if and only if the only local minima of the energy function correspond to the data-points.

Proof:

Follows from the absence of rotational invariance on the N-dimensional hypercube.

Heuristics

- In both RNNs and Diffusion Models we can pass from memorization to generalization by increasing the "temperature of learning" of a certain amount.

Heuristics:

I can achieve **memorization** with any pdf $p_t(x) = \frac{1}{Z_t} e^{-\frac{1}{t}E(x)}$ such that:

$$\lim_{t \to 0} p_t(x) = \frac{1}{P} \sum_{\mu=1}^{P} \delta(x - \xi^\mu)$$

- Diffusion Models can memorize **whatever number** of data-points.
- Recurrent Neural Networks can memorize up to a **sub-exponential number** of data-points ($P_{max} \simeq N/2$). No proof of the storage capacity yet.

Heuristics

- In both RNNs and Diffusion Models we can pass from memorization to generalization by increasing the “temperature of learning” of a certain amount.

Heuristics:

Generalization depends on the way such pdf converges to the mixture of Dirac deltas:

$$\lim_{t \to 0} p_t(x) = \frac{1}{P} \sum_{\mu=1}^{P} \delta(x - \xi^\mu)$$

- Diffusion Models are **“rigid”** learning systems.
- Recurrent Neural Networks are **“liquid”** learning systems.

Question: What about deep neural network-trained Diffusion Models?

AMP approach to Diffusion

Understanding "microscopically" how a trained diffusion model fits the data-manifold.

$$p_t(x) = \int D\mathbf{z} \frac{1}{\sqrt{2\pi\Delta_t}^N} \; e^{-\frac{1}{2t}\|\mathbf{x}-\sigma(F\mathbf{z})\|^2}$$

$$s_t(x) = -\nabla \log p_t(x) \qquad \textbf{exact score}$$

This problem can be mapped in a Generalized Linear Model solvable through AMP.

$$-\nabla \log p_t^{AMP}(\mathbf{x}) = -\frac{1}{t}\left(\mathbf{x} - \langle \sigma(F\mathbf{z}) \rangle_t\right)$$

Then one can compute, for different parameters of a neural network:

$$\mathbb{E}_{\mathbf{x}\sim p_t} \; \|\nabla \log p_t^{AMP}(\mathbf{x}) - \hat{s}_t(\mathbf{x})\|^2$$

where $\hat{s}_t(\mathbf{x})$ is the trained score function