Statistical mechanics of autoregressive models: towards a theory of Self-Attention

Francesco D'Amico

Supervisors:

Matteo Negri

Chiara Cammarota

SAPIENZA
UNIVERSITÀ DI ROMA

Dipartimento di Fisica

PhD talks, October 16, 2025

Outline of the talk

1. Introduction
   - Associative memories (Hopfield networks)
   - Self-Attention and LLMs
2. Our results
   - Self-Attention as pseudolikelihood optimization
   - Pseudolikelihood produces associative memories
   - Vector-spin associative memories
3. Last step: work in progress

Introduction

- Associative memories (Hopfield networks)
- Self-Attention and LLMs

Associative memories (Hopfield networks)

- Network of N binary neurons, $\vec{\sigma} \in \{-1,+1\}^N$
- $$H = -\sum_{(i,j)}^{N} J_{ij}\sigma_i\sigma_j \quad , \quad J_{ij} := \frac{1}{N}\sum_{\mu}^{P} \xi_i^{\mu}\xi_j^{\mu} \quad \leftarrow \textbf{Hebb's rule}$$

Associative memories (Hopfield networks)

- Network of N binary neurons, $\vec{\sigma} \in \{-1,+1\}^N$
- $$H = -\sum_{(i,j)}^{N} J_{ij}\sigma_i\sigma_j \quad , \quad J_{ij} := \frac{1}{N}\sum_{\mu}^{P} \xi_i^{\mu}\xi_j^{\mu} \quad \leftarrow \textbf{Hebb's rule}$$
- P random *patterns* $\vec{\xi}^{\mu} \in \{-1,+1\}^N$ are the *memories*
- $\alpha = \frac{P}{N}$ is the control parameter

Associative memories (Hopfield networks)

- Network of N binary neurons, $\vec{\sigma} \in \{-1,+1\}^N$
- $$H = -\sum_{(i,j)}^{N} J_{ij}\sigma_i\sigma_j \quad , \quad J_{ij} := \frac{1}{N}\sum_{\mu}^{P} \xi_i^{\mu}\xi_j^{\mu} \quad \leftarrow \textbf{Hebb's rule}$$
- P random *patterns* $\vec{\xi}^{\mu} \in \{-1,+1\}^N$ are the *memories*
- $\alpha = \frac{P}{N}$ is the control parameter
- T=0 dynamical rule:

$$\sigma_i(t+\Delta t) = \text{sign}\left[\sum_{j(\neq i)} J_{ij}\sigma_j(t)\right]$$

Associative memories (Hopfield networks)

$\xi_i^\mu = \text{sign}\left[\xi_i^\mu + \mathcal{O}\left(\sqrt{\frac{P}{N}}\right)\right] \Rightarrow$ memories are fixed points of dynamics

Associative memories (Hopfield networks)

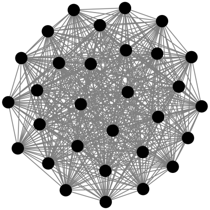
$$\xi_i^\mu = \text{sign}\left[\xi_i^\mu + \mathscr{O}\left(\sqrt{\frac{P}{N}}\right)\right] \Rightarrow \text{memories are fixed points of dynamics}$$

Associative memories (Hopfield networks)

$\xi_i^\mu = \text{sign}\left[\xi_i^\mu + \mathcal{O}\left(\sqrt{\frac{P}{N}}\right)\right] \Rightarrow$ memories are fixed points of dynamics

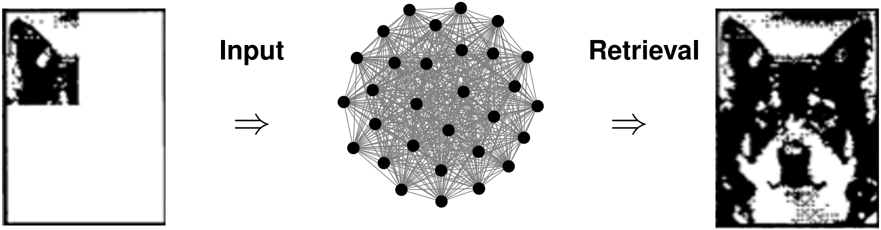
Input

$\Rightarrow$



Associative memories (Hopfield networks)

$$\xi_i^\mu = \text{sign}\left[\xi_i^\mu + \mathcal{O}\left(\sqrt{\frac{P}{N}}\right)\right] \Rightarrow \text{memories are fixed points of dynamics}$$

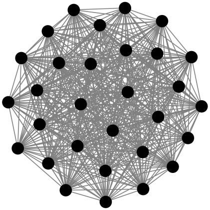


Associative memories (Hopfield networks)

$$\xi_i^\mu = \text{sign}\left[\xi_i^\mu + \mathcal{O}\left(\sqrt{\frac{P}{N}}\right)\right] \Rightarrow \text{memories are fixed points of dynamics}$$

Input

$\Rightarrow$

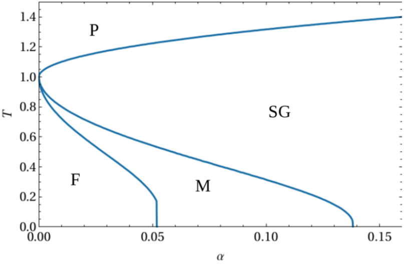


Retrieval

$\Rightarrow$

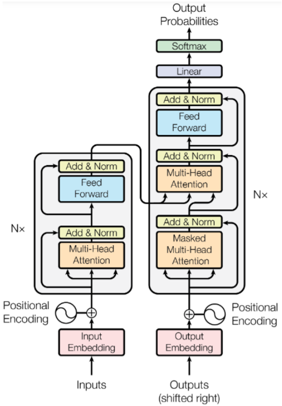
Phase diagram of retrieval

Amit et al. (1987)
Statistical Mechanics of Neural Networks Near Saturation



Self-Attention and LLMs

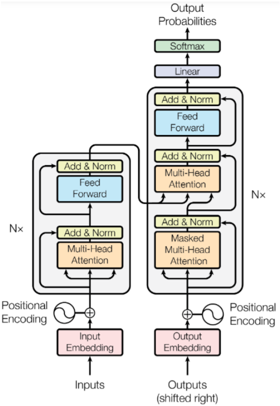
- Transformer architecture:
  Vaswani et al. (2017) Attention Is All You Need



Self-Attention and LLMs

- Transformer architecture:
  Vaswani et al. (2017) Attention Is All You Need
- Input and output: N vectors in d dimension

$$\{\vec{X}_i\}_{i=1,\ldots,N}\ ;\ \vec{X}_i \in \mathbb{R}^d$$



Self-Attention and LLMs

- Transformer architecture:
  Vaswani et al. (2017) Attention Is All You Need

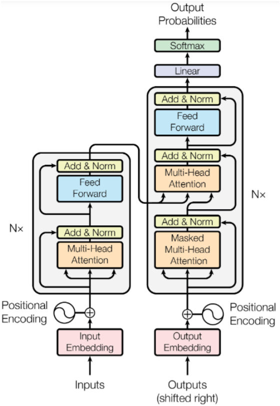
- Input and output: N vectors in d dimension

$$\{\vec{X}_i\}_{i=1,\ldots,N} \; ; \; \vec{X}_i \in \mathbb{R}^d$$

- Building block: Self-Attention layer

$$\vec{X}_i^{\text{out}} = \sum_{j=1}^{N} \text{softmax}_j \left[\left(\mathbf{K}\vec{X}_i \right)^T \left(\mathbf{Q}\vec{X}_j \right) \right] \mathbf{V}\vec{X}_j$$

$\mathbf{K}, \mathbf{Q} \in \mathbb{R}^{r \times d}$, $\mathbf{V} \in \mathbf{R}^{d \times d}$ are learnable matrices



Self-Attention and LLMs

- Transformer architecture:
  Vaswani et al. (2017) Attention Is All You Need

- Input and output: N vectors in d dimension

$$\{\vec{X}_i\}_{i=1,\dots,N}\ ;\ \vec{X}_i \in \mathbb{R}^d$$

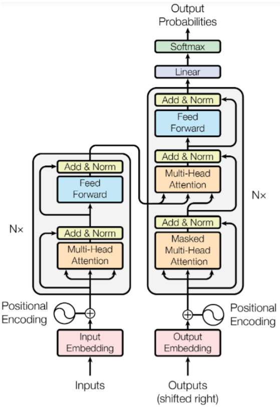
- Building block: Self-Attention layer

$$\vec{X}_i^{\text{out}} = \sum_{j=1}^{N} \text{softmax}_j \left[\left(\mathbf{K}\vec{X}_i\right)^T \left(\mathbf{Q}\vec{X}_j\right) \right] \mathbf{V}\vec{X}_j$$

$\mathbf{K}, \mathbf{Q} \in \mathbb{R}^{r\times d}$, $\mathbf{V} \in \mathbf{R}^{d\times d}$ are learnable matrices

- Simplified version: $\mathbf{K}^T \cdot \mathbf{Q} = \mathbf{V} = J$

$$\vec{X}_i^{\text{out}} = \sum_{j=1}^{N} \text{softmax}_j \left[\vec{X}_i^T \mathbf{J} \vec{X}_j \right] \mathbf{J}\vec{X}_j$$



Our work

- Self-Attention as pseudolikelihood optimization
- Pseudolikelihood produces associative memories
- Vector-spin associative memories

Self-Attention as pseudolikelihood optimization

Self-attention as an attractor network: transient memories without backpropagation

1st Francesco D'Amico 2nd Matteo Negri

- Layer of simplified Self-Attention: weights tensor $\mathbb{J} \in \mathbb{R}^{N \times N \times d \times d}$

$$\vec{X}_i^{t+1} = \sum_{j(\neq i)} \alpha_{i \leftarrow j} \mathbf{J}_{ij} \vec{X}_j^t \tag{1}$$

$$\alpha_{i \leftarrow j} = \text{softmax}_j \left[\lambda \vec{X}_i^t \cdot \mathbf{J}_{ij} \vec{X}_j^t \right] \tag{2}$$

Self-Attention as pseudolikelihood optimization

Self-attention as an attractor network: transient memories without backpropagation

1st Francesco D'Amico   2nd Matteo Negri

- Layer of simplified Self-Attention: weights tensor $\mathbb{J} \in \mathbb{R}^{N \times N \times d \times d}$

$$\vec{X}_i^{t+1} = \sum_{j(\neq i)} \alpha_{i \leftarrow j} \mathbf{J}_{ij} \vec{X}_j^t \tag{1}$$

$$\alpha_{i \leftarrow j} = \text{softmax}_j \left[\lambda \vec{X}_i^t \cdot \mathbf{J}_{ij} \vec{X}_j^t \right] \tag{2}$$

- t layer index $\rightarrow$ promoted to time index

Self-Attention as pseudolikelihood optimization

Self-attention as an attractor network: transient memories without backpropagation

1st Francesco D'Amico 2nd Matteo Negri

- Layer of simplified Self-Attention: weights tensor $\mathbb{J} \in \mathbb{R}^{N \times N \times d \times d}$

$$\vec{X}_i^{t+1} = \sum_{j(\neq i)} \alpha_{i \leftarrow j} \mathbf{J}_{ij} \vec{X}_j^t \tag{1}$$

$$\alpha_{i \leftarrow j} = \text{softmax}_j \left[\lambda \vec{X}_i^t \cdot \mathbf{J}_{ij} \vec{X}_j^t \right] \tag{2}$$

- t layer index $\rightarrow$ promoted to time index
- Eq. 1: minimization dynamics of cost

$$F(\{\vec{X}_i\}; J) = -\frac{1}{\lambda} \sum_i \log \left[\sum_{j(\neq i)} \exp\left(\lambda \vec{X}_i \cdot \mathbf{J}_{ij} \vec{X}_j\right) \right] = \sum_i e_i(\{\vec{X}_i\}; \mathbb{J}) \tag{3}$$

$$\vec{X}_i^{t+1} = -\nabla_{\vec{X}} F(\{\vec{X}_i\}; J) \tag{4}$$

Pseudolikelihood method

- Model with two-bodies interaction: $E(x) = -\sum_{i \neq j} J_{ij} x_i x_j$
- Joint probability $p_J(x) = \exp\{-\lambda E(x)\}/Z_J$

Pseudolikelihood method

- Model with two-bodies interaction: $E(x) = -\sum_{i \neq j} J_{ij} x_i x_j$
- Joint probability $p_J(x) = \exp\{-\lambda E(x)\}/Z_J$
- Dataset of P datapoints $\xi^\mu \in \mathbb{R}^N$
- Likelihood training: minimize $\mathscr{L} = -\sum_{\mu=1}^{P} \log p_J(\xi^\mu)$

Pseudolikelihood method

- Model with two-bodies interaction: $E(x) = -\sum_{i \neq j} J_{ij} x_i x_j$
- Joint probability $p_J(x) = \exp\{-\lambda E(x)\}/Z_J$
- Dataset of P datapoints $\xi^\mu \in \mathbb{R}^N$
- Likelihood training: minimize $\mathscr{L} = -\sum_{\mu=1}^{P} \log p_J(\xi^\mu)$
- Problem: untractable partition function Z_J

Pseudolikelihood method

- Model with two-bodies interaction: $E(x) = -\sum_{i \neq j} J_{ij} x_i x_j$
- Joint probability $p_J(x) = \exp\{-\lambda E(x)\}/Z_J$
- Dataset of P datapoints $\xi^\mu \in \mathbb{R}^N$
- Likelihood training: minimize $\mathscr{L} = -\sum_{\mu=1}^{P} \log p_J(\xi^\mu)$
- Problem: untractable partition function Z_J
- Pseudo-likelihood approximation: $\mathscr{L} = -\sum_{\mu=1}^{P} \sum_{i=1}^{N} \log p_i(\xi_i^\mu | \xi^\mu{}_{\setminus i})$

Pseudolikelihood produces associative memories

Step back:

What happens for the simplest possible model trained with pseudolikelihood?

Pseudolikelihood produces associative memories

Step back:

What happens for the simplest possible model trained with pseudolikelihood?

$\Downarrow$

Pseudo-likelihood produces associative memories able to generalize, even for asymmetric couplings

Francesco D'Amico,[1,2] Dario Bocchi,[1,2] Luca Maria Del Bono,[1,2] Saverio Rossi,[1] and Matteo Negri[1,2]

[1] *Physics Department, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy*
[2] *Institute of Nanotechnology, National Research Council of Italy, CNR-NANOTEC, Rome Unit*

Pseudolikelihood produces associative memories

Pseudolikelihood on random data: same setting as Hopfield

Pseudolikelihood produces associative memories

Pseudolikelihood on random data: same setting as Hopfield

- N neurons, $\vec{\sigma} \in \{\pm 1\}^N$, T=0 dynamics:
  $\sigma_i(t+\Delta t) = \text{sign}\left[\sum_{j(\neq i)} J_{ij}\sigma_j(t)\right]$
- P random memories $\vec{\xi}^{\mu} \in \{\pm 1\}^N$, $\alpha = \frac{P}{N}$ control parameter

Pseudolikelihood produces associative memories

Pseudolikelihood on random data: same setting as Hopfield

- N neurons, $\vec{\sigma} \in \{\pm 1\}^N$, T=0 dynamics:
  $\sigma_i(t+\Delta t) = \text{sign}\left[\sum_{j(\neq i)} J_{ij}\sigma_j(t)\right]$
- P random memories $\vec{\xi}^\mu \in \{\pm 1\}^N$, $\alpha = \frac{P}{N}$ control parameter
- J_{ij} as negative log-pseudolikelihood minimizer at fixed $\|\mathbf{J}\| = \lambda$

$$NLP = \sum_{i=1}^{N} \ell_i(J_i) = \sum_{i=1}^{N}\sum_{\mu=1}^{P} \log\left(1 + e^{-\xi_i^\mu \sum_{j\neq i} J_{ij}\xi_j^\mu}\right)$$

Pseudolikelihood produces associative memories

Pseudolikelihood on random data: same setting as Hopfield

- N neurons, $\vec{\sigma} \in \{\pm 1\}^N$, T=0 dynamics: $\sigma_i(t+\Delta t) = \text{sign}\left[\sum_{j(\neq i)} J_{ij}\sigma_j(t)\right]$
- P random memories $\vec{\xi}^\mu \in \{\pm 1\}^N$, $\alpha = \frac{P}{N}$ control parameter
- J_{ij} as negative log-pseudolikelihood minimizer at fixed $\|\mathbf{J}\| = \lambda$

$$NLP = \sum_{i=1}^{N} \ell_i(J_i) = \sum_{i=1}^{N} \sum_{\mu=1}^{P} \log\left(1 + e^{-\xi_i^\mu \sum_{j\neq i} J_{ij}\xi_j^\mu}\right)$$

- Quantity of interest: stabilities

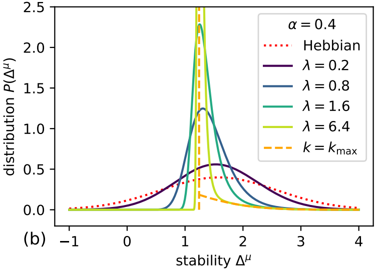
$$\Delta_i^\mu = \xi_i^\mu \left(\sum_{j\neq i} J_{ij}\xi_j^\mu\right)$$

Pseudolikelihood produces associative memories

Pseudolikelihood on random data: Gardner computation

Pseudolikelihood produces associative memories

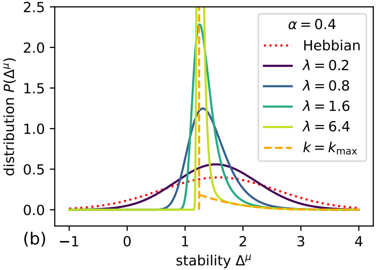
Pseudolikelihood on random data: Gardner computation

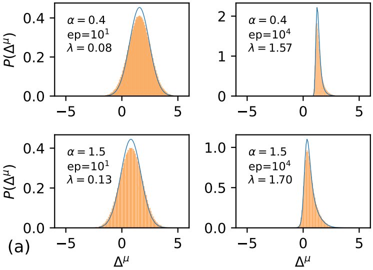


At fixed norm $\|\mathbf{J}\| = \lambda$

Pseudolikelihood produces associative memories

Pseudolikelihood on random data: Gardner computation





At fixed norm $\|\mathbf{J}\| = \lambda$

GD at free norm

Vector spins: towards Self-Attention

Vector spins: towards Self-Attention

Statistical mechanics of vector Hopfield network near and above saturation

Flavio Nicoletti,[1,2,*] Francesco D'Amico,[2,3,†] and Matteo Negri[2,3,‡]

[1] *Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden*
[2] *Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy*
[3] *Institute of Nanotechnology, National Research Council of Italy, CNR-NANOTEC, Rome Unit*

Vector spins: towards Self-Attention

Statistical mechanics of vector Hopfield network near and above saturation

Flavio Nicoletti,[1,2,*] Francesco D'Amico,[2,3,†] and Matteo Negri[2,3,‡]

[1] *Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden*
[2] *Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy*
[3] *Institute of Nanotechnology, National Research Council of Italy, CNR-NANOTEC, Rome Unit*

- N spherical vector spins $\{\vec{S}_i\}_{i=1,..,N} \in \mathbb{R}^d$, $\|\vec{S}_i\| = 1$

Vector-spin associative memories

Vector spins: towards Self-Attention

Statistical mechanics of vector Hopfield network near and above saturation

Flavio Nicoletti,[1,2,*] Francesco D'Amico,[2,3,†] and Matteo Negri[2,3,‡]

[1] *Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden*
[2] *Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy*
[3] *Institute of Nanotechnology, National Research Council of Italy, CNR-NANOTEC, Rome Unit*

- N spherical vector spins $\{\vec{S}_i\}_{i=1,..,N} \in \mathbb{R}^d$, $\|\vec{S}_i\| = 1$
- P memories $\{\vec{\xi}_i\}^\mu$

Vector spins: towards Self-Attention

Statistical mechanics of vector Hopfield network near and above saturation

Flavio Nicoletti,[1,2,*] Francesco D'Amico,[2,3,†] and Matteo Negri[2,3,‡]

[1] *Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden*
[2] *Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy*
[3] *Institute of Nanotechnology, National Research Council of Italy, CNR-NANOTEC, Rome Unit*

- N spherical vector spins $\{\vec{S}_i\}_{i=1,..,N} \in \mathbb{R}^d$, $\|\vec{S}_i\| = 1$
- P memories $\{\vec{\xi}_i\}^\mu$
- $H = -\frac{1}{2}\sum_{i\neq j}^{1,N} \vec{S}_i \cdot \mathbb{J}_{ij}\vec{S}_j$

Vector-spin associative memories

Vector spins: towards Self-Attention

Statistical mechanics of vector Hopfield network near and above saturation

Flavio Nicoletti,[1,2,*] Francesco D'Amico,[2,3,†] and Matteo Negri[2,3,‡]

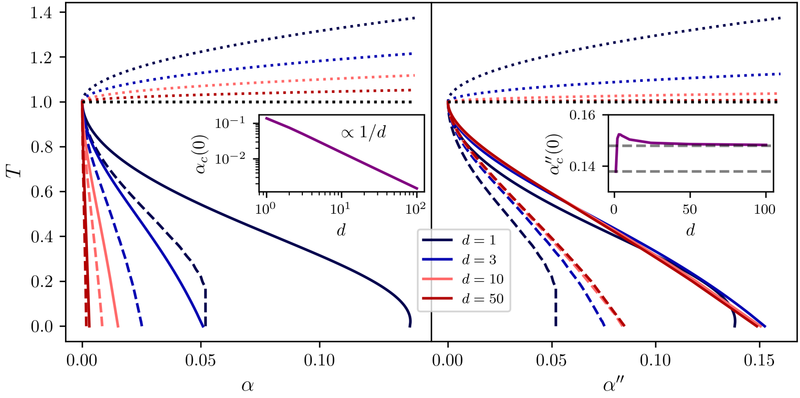
[1] *Department of Computer Science and Engineering Chalmers University of Technology and University of Gothenburg SE-41296 Gothenburg, Sweden*
[2] *Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy*
[3] *Institute of Nanotechnology, National Research Council of Italy, CNR-NANOTEC, Rome Unit*

- N spherical vector spins $\{\vec{S}_i\}_{i=1,..,N} \in \mathbb{R}^d$, $\|\vec{S}_i\| = 1$
- P memories $\{\vec{\xi}_i\}^\mu$
- $H = -\frac{1}{2}\sum_{i\neq j}^{1,N} \vec{S}_i \cdot \mathbb{J}_{ij}\vec{S}_j$
- Hebb's couplings $\mathbb{J}_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \vec{\xi}_i^{\,\mu} \times \vec{\xi}_j^{\,\mu} \Rightarrow \mathbb{J}_{ij} \in \mathbb{R}^{d\times d}$

Phase diagram of retrieval at equilibrium

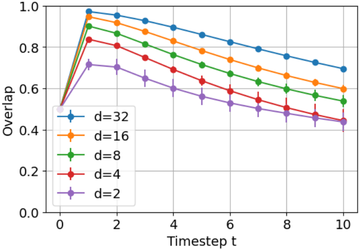
Phase diagram of retrieval at equilibrium

Two order parameters: $\alpha = \frac{P}{N}$, $\alpha'' = \frac{Pd}{N}$



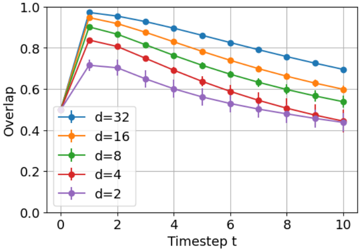
First step denoising

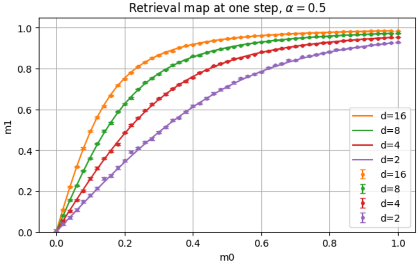
First step denoising



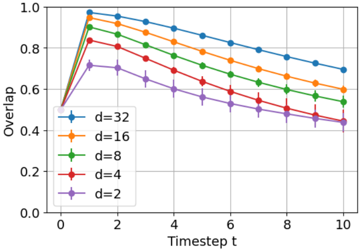
Vector-spin associative memories

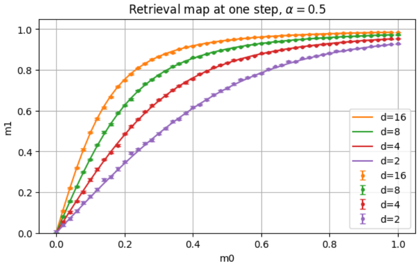
First step denoising





First step denoising





$$\Rightarrow P = \alpha' N d$$

Last step

$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$

Pseudolikelihood

$\boldsymbol{s}_i \in \{\pm 1\}$

✓ *Amit et al. (1987)*

$\boldsymbol{s}_i \in \boldsymbol{s}_{d-1}$

Last step

$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$

Pseudolikelihood

$s_i \in \{\pm 1\}$

✓ *Amit et al. (1987)*

→

✓ *D'Amico, Bocchi, Del Bono, Rossi, Negri (2025)*

$s_i \in s_{d-1}$

Last step

	$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$		Pseudolikelihood
$\boldsymbol{s}_i \in \{\pm 1\}$	✓ *Amit et al. (1987)*	→	✓ *D'Amico, Bocchi, Del Bono, Rossi, Negri (2025)*
	↓		
$\boldsymbol{s}_i \in \boldsymbol{s}_{d-1}$	✓ *Nicoletti, D'Amico, Negri (2025)*		

Last step

	$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$	Pseudolikelihood
$\boldsymbol{s}_i \in \{\pm 1\}$	✓ *Amit et al. (1987)*	✓ *D'Amico, Bocchi, Del Bono, Rossi, Negri (2025)*
$\boldsymbol{s}_i \in \boldsymbol{s}_{d-1}$	✓ *Nicoletti, D'Amico, Negri (2025)*	?