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

Pseudo-likelihood training

- N variables $\mathbf{x} \in \mathbb{R}^N$, dataset of P datapoints $\xi^\mu \in \mathbb{R}^N$, load $\alpha = \frac{P}{N}$
- Energy-based parametrization $p_J(x) = \exp\{-\lambda E(x)\}/Z_J$
- Likelihood training: minimize $\mathcal{L} = -\sum_{\mu=1}^{P} \log p_J(\xi^\mu)$.
  Problem: untractable partition function Z_J

Pseudo-likelihood approximation: $\mathcal{L} = -\sum_{\mu=1}^{P}\sum_{i=1}^{N} \log p_i(\xi_i^\mu | \xi^\mu_{\setminus i})$

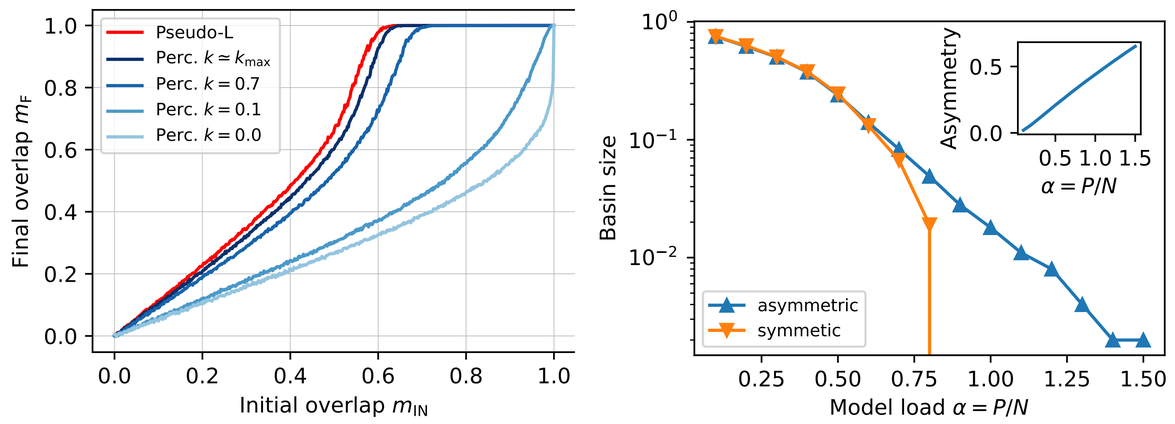
Two-bodies models: $E(x) = -\sum_{i\neq j} J_{ij} x_i x_j$

Pseudo-likelihood loss $\mathcal{L} = -\sum_{i,\mu}\left[\lambda\, \xi_i^\mu \sum_{j\neq i} J_{ij}\xi_j^\mu - \log 2\cosh\left(\lambda \sum_{j\neq i} J_{ij}\xi_j^\mu\right)\right]$

Zero-temperature dynamics: $x_i^{(t+1)} = \mathrm{sign}(\sum_{j(\neq i)} J_{ij} x_j^{(t)})$

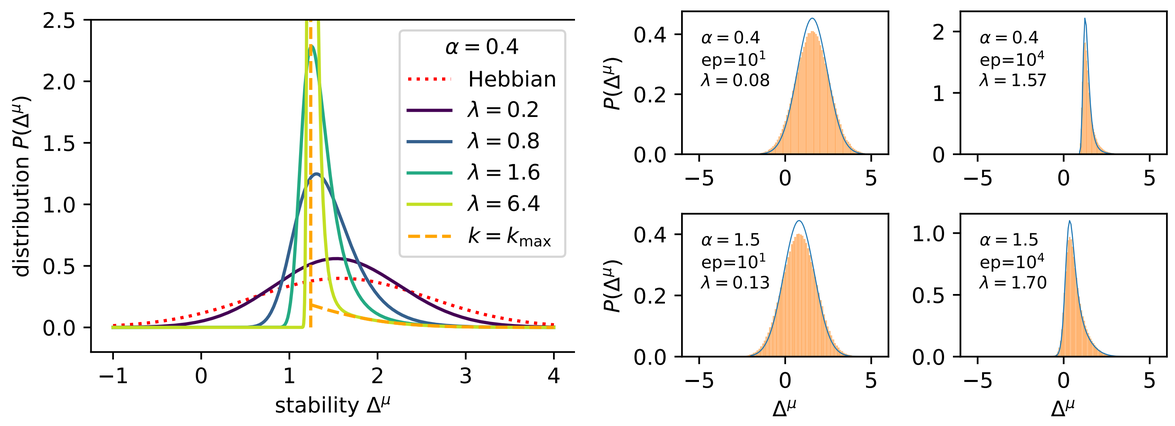
Pseudo-likelihood produces associative memories

- Associative memories store memories as fixed point
- Initial condition $\mathbf{x}^{\mathrm{IN}} \Longrightarrow$ *basin of attraction* of a memory ξ^μ related to minimum overlap $m_{\mathrm{IN}} = \frac{1}{N}\xi^\mu \cdot \mathbf{x}^{\mathrm{IN}}$ s.t. dynamics converges to ξ^μ.
- (*left*) Pseudo-likelihood training produces optimal basins
- (*right*) Works for symmetric and asymmetric couplings.
  Theoretical bounds: $\alpha_C = 1$ for symmetric and $\alpha_C = 2$ for asymmetric.



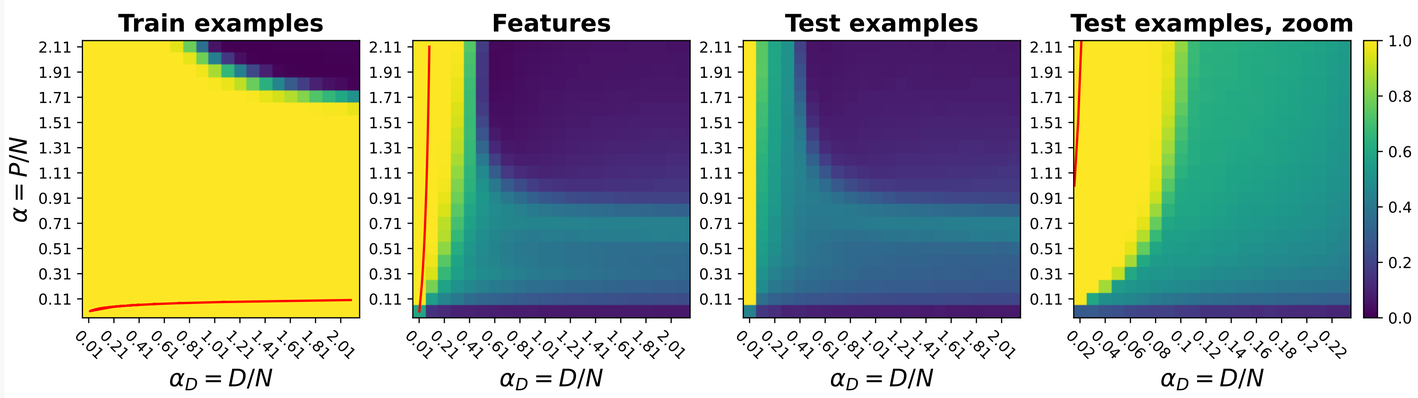
Quantitative theory during training

- Stabilities $\Delta_i^\mu = \frac{1}{\sqrt{N}}\xi_i^\mu \sum_{j\neq i} J_{ij}\xi_j^\mu$ to predict storage of ξ^μ
- Pseudo-likelihood maximized independently by each spin
  $\Longrightarrow$ mapping to training of N independent perceptrons
- (*left*) Fixed norm Perceptron minimizes $V_\lambda(\Delta) = \lambda\Delta - \log 2\cosh(\lambda\Delta)$
  $\Longrightarrow$ Gardner replica computation to obtain $P(\Delta)$ given (α, λ)
- (*right*) $P(\Delta)$ with unbounded norm similar to $\lambda(t) = |\mathbf{J}(t)|$



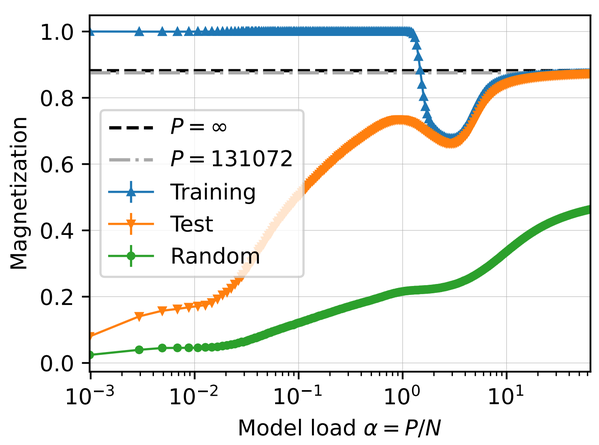
Storage, learning and generalization in a random features dataset (Hidden Manifold Model)

- D random i.i.d. features $\mathbf{f}^k \in \mathbb{R}^N$, with $f_i^k = \pm 1$
- Datapoints $\xi^\mu = \mathrm{sgn}(\sum_k^D c_k^\mu f_i^k)$
- Random coefficients $c_k^\mu = \pm 1$
- Storage of training ξ^μ data, learning of features $\mathbf{f}^k$ and generalization to unseen ξ^{test} examples

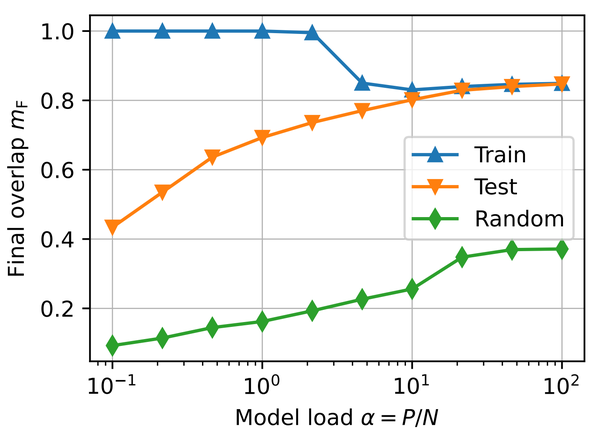


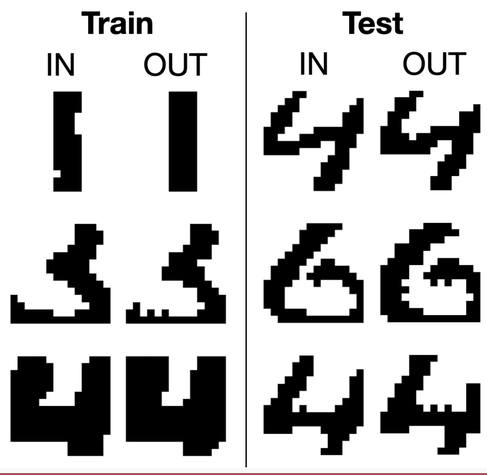
Results on various datasets

Edwards-Anderson 2D

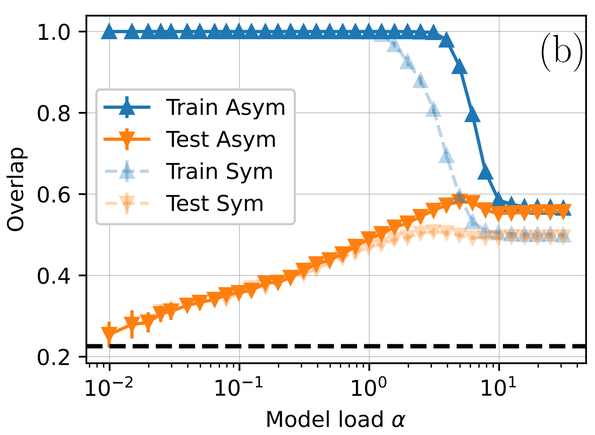


Binarized MNIST 14×14





Protein sequences



Take-home message

- PL produces energy minima over data
- Energy landscape controlled by norm of weights
- With random features fixed-points are developed over features and test data as well
- On real dataset storage-to-generalization

Related works

- *Self-attention as an attractor network: transient memories without backpropagation*, F. D'Amico, M. Negri (2024)
- *Supervised perceptron learning vs unsupervised Hebbian unlearning: Approaching optimal memory retrieval in Hopfield-like networks*, M. Benedetti, E. Ventura, E. Marinari, G. Ruocco, F. Zamponi (2022)
- *Implicit bias produces neural scaling laws in learning curves, from perceptrons to deep networks*, F. D'Amico, D. Bocchi, M. Negri (2025)
- *Random Features Hopfield Networks generalize retrieval to previously unseen examples*, S. Kalaj, C. Lauditi, G. Perugini, C. Lucibello, E. M. Malatesta, M. Negri (2024)

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

francesco.damico@uniroma1.it

[1]Università di Roma Sapienza, [2] CNR-NANOTEC

Funded by the European Union NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research
SAPIENZA UNIVERSITÀ DI ROMA