Power-law feature statistics explain test reconstruction gaps in Associative Memories

Sergio E.G. Manfrin - Francesco D'Amico - Marco Gherardi - Aurélien Decelle - Beatriz Seoane - Matteo Negri

Università degli Studi di Milano | Università di Roma Sapienza | Università degli Studi di Milano | Universidad Politécnica de Madrid | Universidad Complutense de Madrid | CY Cergy Paris Université

TL;DR Toy data with power-law feature statistics reproduces real settings in AM

Model

Cross-entropy binary Hopfield

- Deterministic recurrent network of **N** Binary neurons with weights J_{ij}

$$x_i^{(t+1)} = \mathrm{sgn}\Big[\sum_{j(\neq i)}^{N} J_{ij}\, x_j^{(t)}\Big]$$

- Given a training dataset of **P** examples $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$ we minimize cross-entropy / pseudo-likelihood loss to obtain J_{ij}

$$\mathcal{L}(\mathbf{J}) = \sum_{\mu,i} \log\left(1 + e^{-\lambda \xi_i^\mu \sum_{j\neq i} J_{ij}\xi_j^\mu}\right)$$

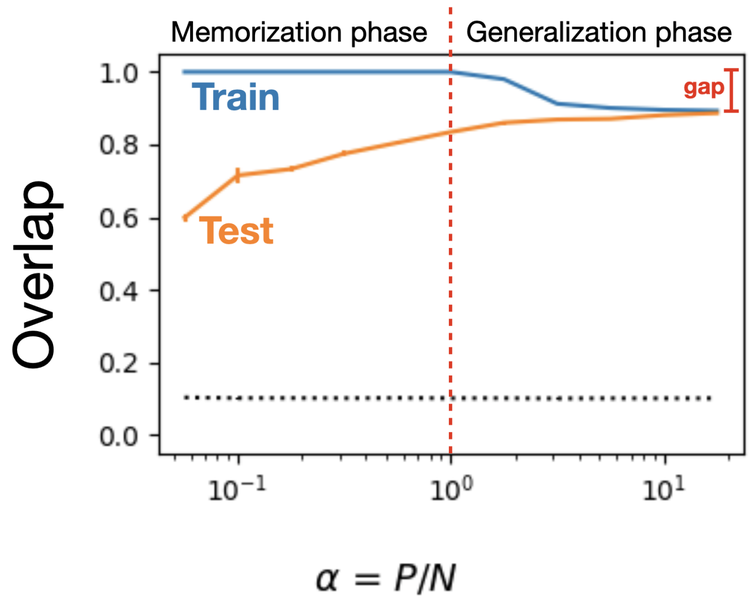
Artificial data:

Hidden manifold model with power-law features

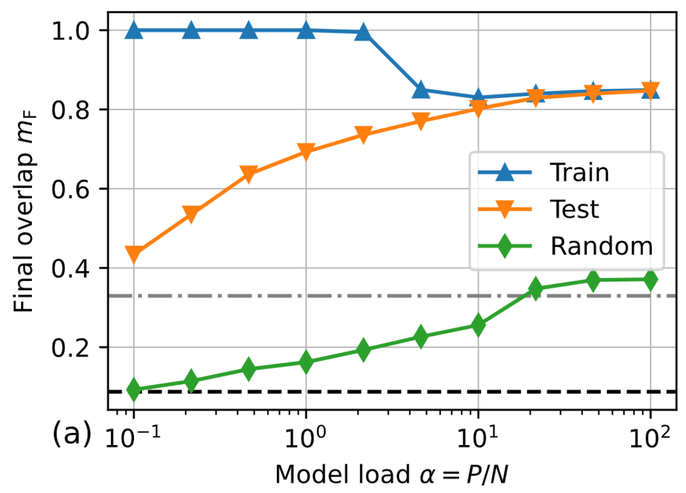
- **P** correlated data $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$ as combinations of **D** random features $\{\boldsymbol{f}^k\}_{k=1}^D$
- Data and features are vectors of **N** binary variables.
- Features are i.i.d. $f_{ki} \sim \mathrm{Unif}(\pm 1)$
- Data are combinations of features $\xi_i^\mu = \mathrm{sgn}\left[\sum_{k=1}^{D} c_k^\mu f_{ki}\right]$, with $c_k^\mu \in \{0, \pm 1\}$
- Combinations are sparse, only $L = \mathcal{O}(1)$ non-zero coefficients for datum μ
- The **L** features are chosen with a power-law probability: $\pi_k = k^{-\eta}/Z(\eta, D)$

1 This artifical dataset reproduces memorization-to-generalization transition observed in real data

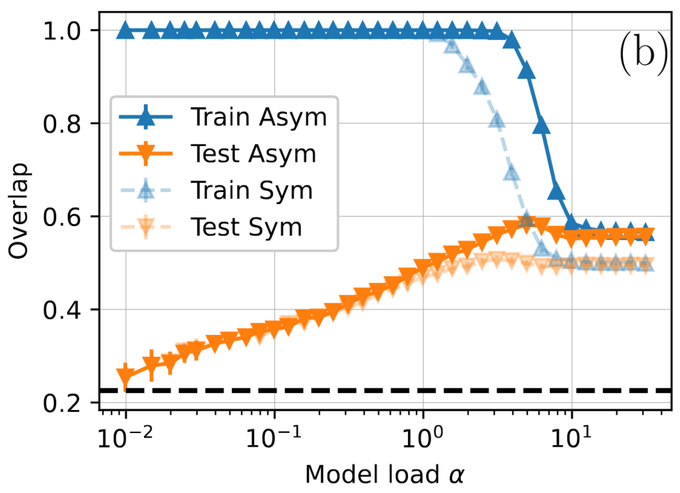
Artificial data



MNIST binarized



Real protein sequences



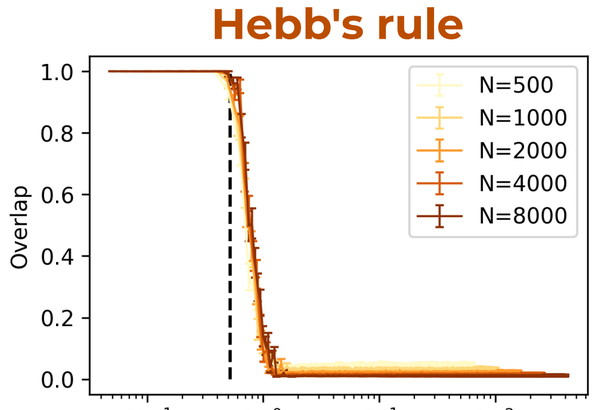
2 The cut-off in number of learned features explains generalization

- Signal-to-noise ratio for Hebb's rule predicts local fields

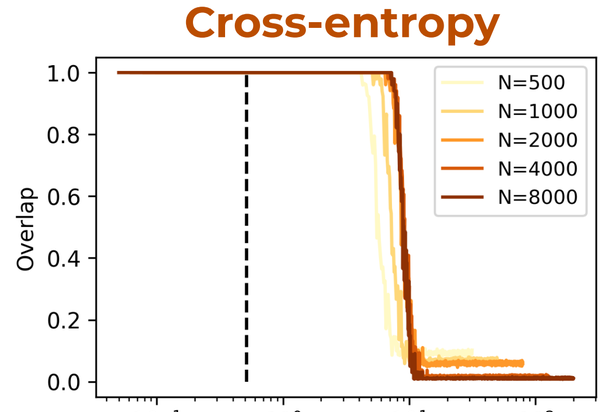
$$h_{ik} \simeq \frac{P_k}{L}\left(f_{ki} + O\left(\sqrt{\frac{Z(2\eta, D)}{N\, k^{-2\eta}}}\right)\right) \Rightarrow \text{last learnable feature } k_{\max}(N) = \mathcal{O}(N^{\frac{1}{2\eta}})$$

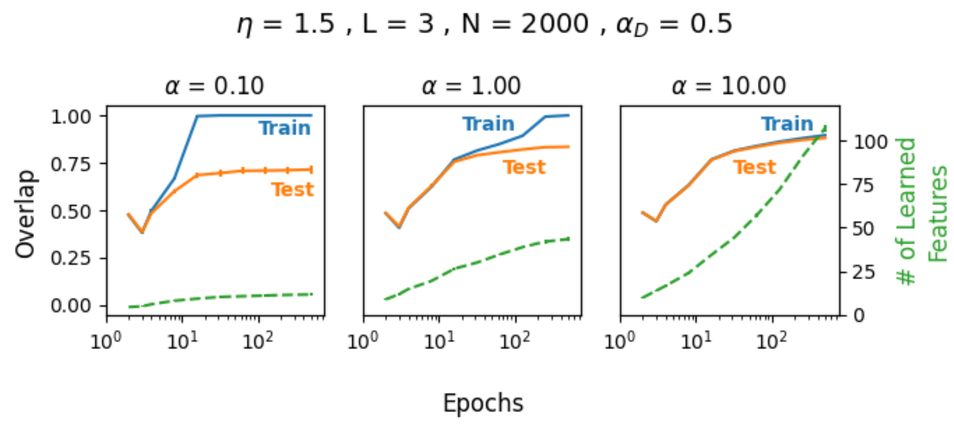
- The number of learned features explains the test reconstruction gap in cross-entropy associative memory dynamics

Hebb's rule



Cross-entropy





- **Hidden manifold model with power-law random features is a good model for real data**
- **The exponent selects how many features are learnable**
- **Test reconstruction gap is directly controlled by the number of features learned**
- **Feature acquisition is sequential: frequent ones are incorporated earlier in training time and dataset size**