BEYOND MEMORY STORAGE CAPACITY OF ASSOCIATIVE MEMORY NETWORKS

MUHAN GAO, MILTON LIN

Contents

1. Introduction

Associative memory networks, particularly Hopfield networks, were among the earliest computational models for memory search and retrieval [6]. Recent developments have significantly advanced these models along two fronts: i) *Improved storage capacity*, progressing from polynomial [8], to exponential [2], and in other point of views, [4] ii) *Integration into modern deep learning architectures*, such as attention mechanisms [9], energy-based transformers [5], and higher-order models like simplicial Hopfield networks [1].

2. Goal

In associative memory networks, the weight matrix stores memory patterns, termed **weight memory patterns** $(\xi_\mu)_{\mu=1}^K$. The fraction of inputs that converge to these patterns within a $(1-\varepsilon)\%$ tolerance is denoted as D_c^ε (Definition 6.2). Theoretical studies have established an upper bound on D_c^0, termed the **storage capacity** $K^{\max}$, representing the maximum number of patterns that can be stored under an independent and identically distributed (i.i.d.) setup (Equation 4).

While much effort has focused on extending $K^{\max}$, there is limited empirical work on the regime where the number of stored patterns K significantly exceeds $K^{\max}$. In practical tasks, where exact recall of stored patterns may not be critical, it is unclear whether this overloading hinders or benefits performance. Thus, our study investigates *partial recovery*, measured by D_c^ε, and explores the consequences of storing $K \gg K^{\max}$.

Date: December 24, 2024.

Our Goal: To empirically study the behavior of associative memory networks in the overloading regime and its implications for:

(1) The focus of future research in associative networks,

(2) Understanding generalization in modern memory architectures,

(3) Informing the design of next-generation memory networks.

3. Results

We conducted experiments using Dense Associative Memory (DAM), a modern Hopfield network variant with exponential storage capacity [8], trained on the MNIST dataset. Our key findings are as follows:

3.1. **Storage Capacity Is Not a Hard Constraint on Task Performance.**

We varied the number of stored patterns K from 500 to $30,000$, far exceeding the theoretical storage capacity (Equation 4). Surprisingly, the network's performance trends, in terms of training and test error, remained similar to those observed near $K = 5500$ (Figure 2). This suggests that for certain tasks, surpassing $K^{\max}$ does not necessarily degrade performance, possibly due to the robustness of partial recall mechanisms or task-specific generalization properties.

3.2. **Impact of Higher-Order Interactions.**

We examined the effect of increasing interaction order n (Equation 2), corresponding to polynomial activations in multilayer perceptrons. Different values of n exhibited broadly consistent performance trends, as shown in Figure 2 and Figure 1. These findings suggest diminishing returns for higher-order interactions and invite further exploration into the trade-offs between model complexity and retrieval fidelity.

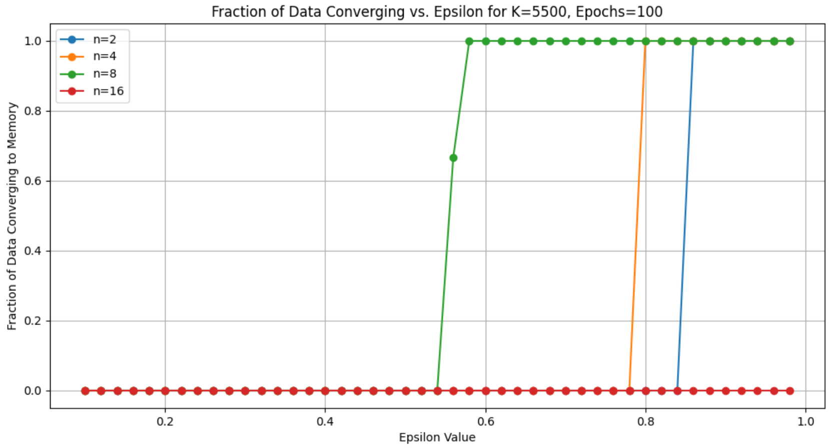

Figure 1. Fraction of Data Converging vs. Epsilon for $K = 5500$, and various interaction orders $n = 2, 4, 8$. As expected from theory, there are more data converging to a pattern.

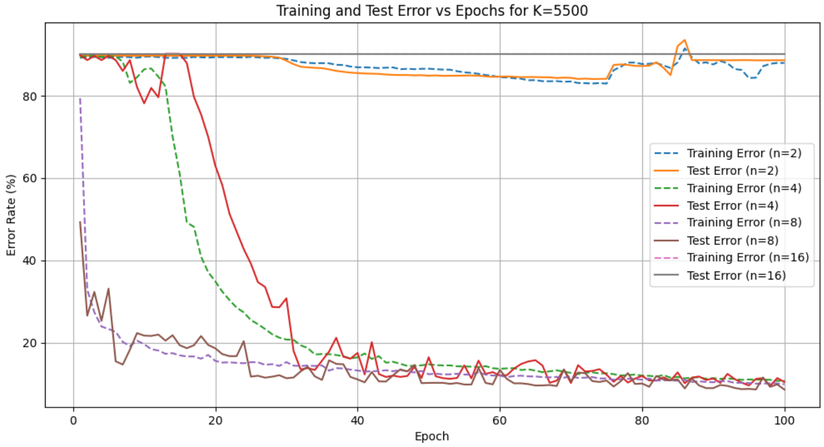

FIGURE 2. Training and Test Error for $K = 5500$. The test error rate does not highly differ for $n = 4, 8$.

4. SUMMARY AND FURTHER PROBLEMS

Scaling Laws and Generalization: The insensitivity of task performance to K in the overloading regime echoes trends seen in scaling laws for deep learning models. Just as overparameterized neural networks often generalize well beyond their nominal capacity, memory networks might leverage a similar phenomenon. Investigating this connection could provide a unified understanding of generalization across tasks and architectures.

Invariant of interaction order The consistent behavior under varying n, Equation 2, hints at an architectural invariance worth deeper theoretical investigation. Linking these observations to broader frameworks, such as energy-based models would be insightful.

The experimental results highlight several phenomena and unresolved questions that connect memory networks to broader challenges in machine learning, including catastrophic forgetting, interpretability, and the scaling of modern architectures. Below, we outline key directions for future work:

- **Task Dependency and Catastrophic Forgetting:** The behavior of stored memory patterns appears highly sensitive to the nature of the task. How does task variability influence memory retrieval, and could this sensitivity offer insights into *catastrophic forgetting*? Understanding this phenomenon, especially in the context of continual learning, could bridge memory networks with advances in lifelong machine learning [7].
- **Correlated Data and Memory Convergence:** Experimental evidence shows that correlated datasets significantly alter convergence behavior to stored memory patterns. Can these observations be formalized theoretically? A deeper understanding of how data structure impacts memory retrieval could inform both theoretical bounds and practical applications.

- **Iterations Required for Convergence:** One iteration of a Hopfield network rarely ensures convergence to a memory pattern. Dense Associative Memories (DAMs) can be viewed as iterative multilayer perceptrons with shared weights [8, Ch.5]. Studying the number of iterations required for reliable convergence, particularly under varying capacities and input perturbations, could yield new insights into optimizing both memory networks and broader architectures.

4.1. **Other approaches to study memory in LLMs.** It is studied in [3], a different approach. For a fixed algorithm $\mathcal{A}$ an training dataset, the amount of label memorization by $\mathcal{A}$ example (x_i, y_i) in dataset S, is defined as

$$\mathrm{mem}(\mathcal{A}, S, i) := \mathbb{P}(h_{\mathcal{A}(S)}(x_i) = (y_i)) - \mathbb{P}(h_{\mathcal{A}(S\setminus i)}(x_i) = y_i)$$

This quantifies the extent to which (x_i, y_i) was important in the memorization. This suggests another form of study with associative memory networks.

5. Background on dense associative memory netowrks

Consider a state $\xi \in \mathbb{R}^N$. For we stored memory patterns $(\xi^\mu)_{\mu=1}^K$, the asynchronous [1] update rule is given by

$$\mathrm{HN}_i(\sigma^{(t)}) := \sigma_i^{(t+1)} := \mathrm{sgn}\left[\sum_{\mu=1}^{K} F(\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)}) - F(-\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)})\right] \tag{1}$$

where

$$F(x) = \begin{cases} x^n & x \geq 0 \\ 0 & x \leq 0 \end{cases} \tag{2}$$

The probability that the ith bit of ξ_i for some $i \in \{1, \ldots, N\}$ of being unstable is

$$P_{\text{error}} := P(\xi_i \text{ is unstable}) = \int_{\langle \Delta E\rangle}^{\infty} \frac{dx}{\sqrt{2\pi\Sigma^2}} e^{\frac{-x^2}{2\Sigma^2}} \approx \sqrt{\frac{(2n-3)!!}{2\pi}\frac{K}{N^{n-1}}} e^{-\frac{N^{n-1}}{2K(2n-3)!!}} \tag{3}$$

For the whole state ξ to be stable we ask that

$$(1 - P_{\text{error}})^N > (1 - \delta)$$

i.e. that there is $(1-\delta)\%$ probability that we get all bits right. For N sufficiently large we make the approximation that

$$P_{\text{error}} < \delta/N$$

by binomial expansion as $(1 - P_{\text{error}})^N \approx 1 - NP_{\text{error}}$. It is shown that [2]

$$K_\delta^{\max}(N) \approx \frac{1}{2(2n-3)!!} \frac{N^{n-1}}{\log N} \tag{4}$$

where !! means double factorial and meant to to compute product of all odd integers (even) to n if n is odd (even). i.e. that there is a $(1-\delta)\%$ chance that we get all bits right for a

[1] update one component once at a time

[2] The paper in [8] seems to imply right hand side is independent of δ.

given state is given by the formula on right hand side as a function of N. Our goal study the regimes whe

$$K < K_\delta^{\max}(N) \text{ and } K \gg K_\delta^{\max}(N)$$

Example 5.1. In [8], the authors fixed $n = 3$, and consider $N \in [50, 200]$. In the case of $n = 3$, we have perfect theoretical convergence of $\frac{1}{2*5!!}\frac{200^2}{\log 200} \approx 1258.26$ when $N = 200$ whilst $\frac{1}{2*3!!}\frac{784^2}{\log 784} \approx 15371.6$ when $N = 784$.

6. Experimental set up

Let

- D denote the set of data, which is the MNIST dataset, this is the set of 1000 digits (100 digits for each class).
- N be the number of nodes.
- $M_\delta^{\max}(N)$ be the maximal memory storage capacity with error ε (for iids) for N bits. This is the largest value of stored memory M so that $P_{\text{error}} < \varepsilon/N$.

6.1. **Convergence metric.** On the empirical side, to measure how much a fixed state input $\sigma \in D$ is recovered, we define the following metric:

Definition 6.1. [8, App. B] The **recovery of state σ after n iterations:**

$$\text{Cvg}_K^{(n)}(\sigma) := \left| \max_{\mu=1}^{K} \left\langle \xi^\mu, \text{HN}^{(n)}(\sigma) \right\rangle \right| \in [0, N], \tag{5}$$

Here $\text{HN}(\sigma) = (\text{HN}_1(\sigma), \ldots, \text{HN}_N(\sigma))$, where each component is defined as Equation 1 is a synchronous update of all vectors. $\text{Cvg}_K^{(n)}$ measures the distance of $\text{HN}^{(n)}(\sigma) \in \{-1, 1\}^N$ and the closest memory, $\xi^\mu \in \{-1, 1\}^N$. This ranges from 0 to N.

Definition 6.2. The fraction of data that converges to some weight memory pattern with ε error is

$$D_\epsilon^c := \frac{|\{\sigma \,:\, \text{Cvg}_K(\sigma) \geq (1-\varepsilon)N\}|}{|D|} \tag{6}$$

As a first approximation : it seems that this should be quite small in a completely random setting, so that a convergence of ϵ% error is still a nontrivial "memory retrieval". Previous studies has focused on perfect recovery when $\varepsilon = 0$.

Proposition 6.3. *If a data y and a memory weight patterns ξ^μ are iid uniform on $\{-1, 1\}$ for each component, then as $N \to \infty$,*

$$P(|\langle \xi^\mu, y \rangle| \geq (1-\varepsilon)N) \approx P(|Z| \geq (1-\varepsilon)\sqrt{N})$$

where $Z \sim N(0, 1)$ is the standard normal distributionn.

Proof. Sketch Here $S_N := \sum_i^N X_i$, where $X_i = \xi_i^\mu y_i$, of which ξ_i^μ and y_i are uniform iid on $\{-1, 1\}$. This also implies X_i itself is too. Now apply Central Limit theorem, where we may suppose $S_N \sim \mathcal{N}(0, N)$. □

Remark 6.4. This metric is quite different from what is required in [9], which is a *continuous Hopfield network.*

Example 6.5. When $N = 4, \varepsilon = 1/2$, then the chance of a random binary state, y matches to weight pattern is approximately 15%. This decreases significantly as $N \to \infty$.

Parameter	Symbol	Value/Range
Data set size	$\lvert D\rvert$	10000 (100 per class, 10 class)
Number of nodes	N	100 to 1000
Maximum memory capacity	$M_\delta^{\max}(N)$	Varies with error δ
Error tolerance	ε	0 to 1
Fraction of data retrieved / Converges to a memory pattern	α	0 to 1

TABLE 1. Experimental Parameters and Settings

References

[1] Burns, Thomas F and Fukai, Tomoki. *Simplicial Hopfield networks*. 2023. arXiv: 2305.05179 [cs.NE]. URL: https://arxiv.org/abs/2305.05179 (cit. on p. 1).

[2] Demircigil, Mete, Heusel, Judith, Löwe, Matthias, Upgang, Sven, and Vermet, Franck. "On a Model of Associative Memory with Huge Storage Capacity". In: 168 (May 2017), pp. 288–299 (cit. on p. 1).

[3] Feldman, Vitaly and Zhang, Chiyuan. *What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation*. 2020. arXiv: 2008.03703 [cs.LG]. URL: https://arxiv.org/abs/2008.03703 (cit. on p. 4).

[4] Hillar, Christopher J. and Tran, Ngoc Mai. "Robust Exponential Memory in Hopfield Networks". In: *Journal of Mathematical Neuroscience* 8 (2014). URL: https://api.semanticscholar.org/CorpusID:11295055 (cit. on p. 1).

[5] Hoover, Benjamin, Liang, Yuchen, Pham, Bao, Panda, Rameswar, Strobelt, Hendrik, Chau, Duen Horng, Zaki, Mohammed J., and Krotov, Dmitry. *Energy Transformer*. 2023. arXiv: 2302.07253 [cs.LG]. URL: https://arxiv.org/abs/2302.07253 (cit. on p. 1).

[6] Kahana, Michael J. "Computational Models of Memory Search." In: *Annual review of psychology* (2020). URL: https://api.semanticscholar.org/CorpusID:203624267 (cit. on p. 1).

[7] Kemker, Ronald, Abitino, Angelina, McClure, Marc, and Kanan, Christopher. "Measuring Catastrophic Forgetting in Neural Networks". In: *ArXiv* abs/1708.02072 (2017). URL: https://api.semanticscholar.org/CorpusID:22910766 (cit. on p. 3).

[8] Krotov, Dmitry and Hopfield, John J. *Dense Associative Memory for Pattern Recognition*. 2016. arXiv: 1606.01164 [cs.NE]. URL: https://arxiv.org/abs/1606.01164 (cit. on pp. 1, 2, 4, 5).

[9] Ramsauer, Hubert, Schäfl, Bernhard, Lehner, Johannes, Seidl, Philipp, Widrich, Michael, Gruber, Lukas, Holzleitner, Markus, Adler, Thomas, Kreil, David, Kopp, Michael K, Klambauer, Günter, Brandstetter, Johannes, and Hochreiter, Sepp. "Hopfield Networks is All You Need". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=tL89RnzIiCd (cit. on pp. 1, 6).