BEYOND MEMORY STORAGE CAPACITY OF ASSOCIATIVE MEMORY NETWORKS

MUHAN GAO, MILTON LIN

Contents

1. Memory networks beyond memory capacity

For a survey of computational memory models, see [5]. Associative memory networks (the Hopfield network) is one of the very first computational models of memory search. Recent interest arises from i) improved memory storage capacities from polynomial [7], exponential [2] and to kernalized. ii) novel architectures [8], [4], [1], which incorporated hopfield models to current deep learning frameworks.

Goal. A trained memory network should have the weights of its weight matrix as stored memory patterns, we refer to these as **weight memory patterns**, these are denoted as $(\xi_\mu)_{\mu=1}^K$ in paper, see Equation 1. The fraction of data which converges to some weight memory pattern with $(1-\varepsilon)\%$ tolerance is denoted as D_c^ε, Definition 3.2. In literature, there is a theoretical upper bound, $K^{\max}$, which allows an input to converge to weight memory with high probability (under an iid set up) Equation 3, referred to as **storage capcaity**. Numerous number of follow-up works have been attempting to extend this bound, however, little has been said whether it is *ok* to store more than necessary memory in the context of practical tasks. Our goal: **to study the regime when one stores memory beyond the storage capacity,** $K \gg K^{\max}$.

We expect this can give us insight on i) focus of future study in Hopfield networks ii) the nature of generalization in the context of modern networks. We suggest some open problems to continue on from the current experiments in Section 1.2.

Date: November 15, 2024.

1.1. **Results from experiments.** We trained dense associative memory networks on MNIST dataset.

- In the context in incorporating Hopfield models to modern deep neural net, **storage capcity is not necessarily for task performance**, we ran experiments for the number of stored memory patterns, K, from 500 to $3 * 10^4$, which is way beyond theoretical capacity, Equation 3. The general trend is similar to the case $K = 5500$, Figure 2.
- It would be interesting to study limiting behavior **increasing interaction**. In the context of multilayer perception, this **does not seem to affect the tasks performance**. Figure 2 and Figure 1.

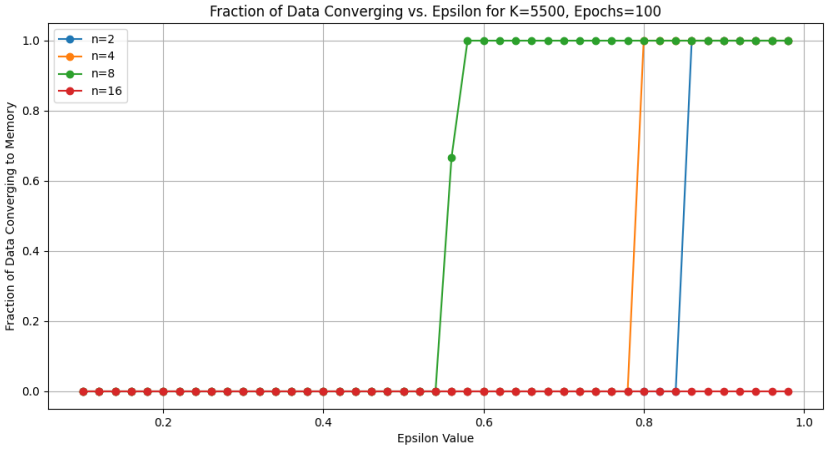

FIGURE 1. Fraction of Data Converging vs. Epsilon for $K = 5500$, and various interaction orders $n = 2, 4, 8, 16$. As expected from theory, there are more data converging to a pattern.

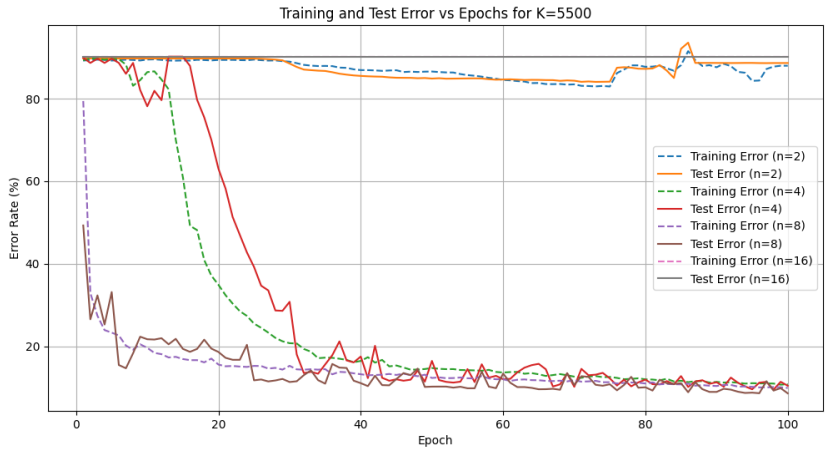

FIGURE 2. Training and Test Error for $K = 5500$. The test error rate does not highly differ for $n = 4, 8$.

1.2. **Open problems.** Below we list a number of interesting phenomena and questions arise from the experiments.

- How does the behavior of stored memory change under change of task, and could this explain **catostrophic forgetting**? [6].
- Correletated dataset strongly affect convergence to weight memory pattern as shown in experiments. Can one say more about the theory?
- One iteration of Hopfield network barely converges to a memory. A DAM can be phrased as iterative multilayer perception with shared weights [7, Ch.5]. It would be interesting to theoretically study the number of iterations required before convergence.

1.3. **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 the Hopfield network.

2. 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

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

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

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

To make computation on the right, it is useful to use the erf function. 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]

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

where !! means double factorial and meant to to compute product of all odd integers (even) to n if n is odd (even).

[1] update one component once at a time

[2] The paper in [7] seems to imply right hand side is independent of ε.

Example 2.1. In [7], 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 $N = 784$ So the values should still be in a manageable range, when we study K from 0 to 10^4.

This says that the maximal storage we can have for perfect recovery i.e. that there is a $(1-\delta)\%$ chance that we get all bits right for a given state is given by the formula on right hand side as a function of N. Our goal study the regimes whe

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

3. 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 $P_{\text{error}} < \varepsilon/N$.

3.1. 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 3.1. [7, App. B] The **recovery of state σ after n iterations:**

$$\text{Cvg}_K^{(n)}(\sigma) := \left| \max_{\mu=1}^{K} \left\langle \xi^\mu, \text{HN}^{(n)}(\sigma) \right\rangle \right| \in [0, N], \tag{4}$$

Here $\text{HN}(\sigma) = (\text{HN}_1(\sigma), \ldots, \text{HN}_N(\sigma))$, as defined in 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 3.2. The fraction of data that converges to some weight memory pattern with ε error is thus

$$D_\epsilon^c := \frac{|\{\sigma : \text{Cvg}_K(\sigma) \geq (1-\varepsilon)N\}|}{|D|} \tag{5}$$

As a first approximation : it seems that this should be quite small in a completely random setting.

Proposition 3.3. *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 3.4. This metric is quite different from what is required in [8], which is a *continuous Hopfield network*.

Example 3.5. 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	$\|D\|$	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. 3).

[4] 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).

[5] 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).

[6] 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).

[7] 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, 3, 4).

[8] 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, 5).