EFFECTS OF FEATURE CORRELATIONS ON ASSOCIATIVE MEMORY CAPACITY

Stefan Bielmeier*
Department of Computer Science and Electrical Engineering
University of California, Berkeley
Berkeley, CA 94720
stefan.bielmeier@tum.de

Gerald Friedland†
Department of Computer Science and Electrical Engineering
University of California, Berkeley
Berkeley, CA 94720
fractor@icsi.berkeley.edu

ABSTRACT

We investigate how feature correlations influence the capacity of Dense Associative Memory (DAM), a Transformer attention-like model. Practical machine learning scenarios involve feature-correlated data and learn representations from input space, but current capacity analyses do not account for this. We develop an empirical framework to analyze the effects of data structure on capacity dynamics. Specifically, we systematically construct datasets that vary in feature correlation and pattern separation using Hamming distance from information theory, and compute the model's corresponding storage capacity using a simple binary search algorithm. Our experiments confirm that memory capacity scales exponentially with increasing separation in the input space. Feature correlations do not alter this relationship fundamentally, but reduce memory capacity slightly at constant separation. This effect is amplified at higher polynomial degrees in the energy function, suggesting that Associative Memory is more limited in depicting higher-order interactions between features than patterns. Our findings bridge theoretical work and practical settings for DAM, and might inspire more data-centric methods.

1 INTRODUCTION

Hopfield networks have gained renewed attention because they exhibit parallels with Transformer architectures and demonstrate strong empirical performance in diverse applications (Vaswani, 2017; Ramsauer et al., 2020; Widrich et al., 2020). Originally, their fundamental ability to store and retrieve patterns was limited to a fraction of the number of its neurons (Hopfield, 1982; 1984). Researchers recently expanded its storage capabilities up to exponentially many memories as Dense Associative Memory (DAM) (Krotov & Hopfield, 2016; Demircigil et al., 2017), and simplified their integration into modern machine learning architectures for practical problem solving (Ramsauer et al., 2020).

However, capacity dynamics under real-world data remain under-explored. While we know that the distribution of memories, and pattern separation in particular, affects storage limits (Wu et al., 2024), most capacity analyses assume theoretically convenient distributions of nearly ideally-separated patterns with independent variables (Krotov & Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2020). This assumption does not hold for practical settings, and real-world data capacity already deviates in part from theoretical predictions (Chaudhry et al., 2023). Therefore understanding how

*Research conducted while a Visiting Research Scholar at UC Berkeley.
†Work done prior to joining Amazon.

Hopfield models behave in data scenarios with imperfect separation and codependent feature is an open question.

To address this, we examine Dense Associative Memory (DAM) in raw input space—mirroring scenarios where feature dimensions are fixed—to reflect practical machine learning contexts (Kempe et al., 2024). We develop a rigorous framework to systematically measure DAM capacity under various data scenarios. Our approach generates patterns with controlled separations to isolate the role of feature correlations in memory storage and retrieval, and measures memory capacity for various model configurations. This work offers insights into capacity behavior that complement existing theoretical analyses.

2 Related Work and Background

Dense Associative Memory. Hopfield networks are a model of auto-associative memory, which is able to recover learned patterns from disrupted versions of themselves (Hopfield, 1982). Dense Associative Memory (DAM) is a variant of it that has a rectified polynomial of degree n as its activation function, and large storage capacity that increases with n (Krotov & Hopfield, 2016).

Memory Capacity. This is the number of patterns the network can store and retrieve reliably at a given error probability. Our target is $P_{error} = 0$. We follow common practice and initialize the network with each pattern in question and perform one asynchronous update to check retrieval performance (Lucibello & Mézard, 2024; Ramsauer et al., 2020; Demircigil et al., 2017).

Kernel Capacity Methods. Recent work exploits kernel transformations to increase Hopfield-like model capacity by making patterns more separable in a high-dimensional feature space via either the energy function of the model, or by reshaping the dataset directly (Wu et al., 2024; Hu et al., 2024). These techniques presuppose freedom to alter data and semantic representations—an assumption that does not always hold in practical settings with fixed input dimensions, such as raw images.

Pattern Separation and Hamming distance. Higher average pairwise pattern separation increases memory capacity, and Hamming distance can be used to identify and remove overlapping patterns to do the same (Hu et al., 2024; 2023; Manandhar & Sadananda, 2002). We view stored patterns as Hamming codes from information theory Hamming (1950); Shannon (1948). For a formal definition of Hamming distance (HD) in context: given two bitstrings ζ_1 and ζ_2, the Hamming distance between them counts the number of differing bit positions. In our framework, this metric is especially useful because each bit difference translates into a discrete, uniform step of dissimilarity, making it easier to design construction algorithms that control average separation.

3 Method and Experimental Setup

We want to measure memory capacity for a set of patterns of different separation. For this, we need to (1) create datasets for both synthetic and real data for a given level of average separation to isolate effects of feature correlations, and (2) compute memory capacity efficiently.

Dataset Construction. We create two main datasets to compare capacity dynamics. The first dataset is the baseline, an "artificial" dataset that contains 50 subsets of skewed i.i.d. Rademacher patterns. This distribution is parameterized by a single parameter which determines whether we sample 1 or -1. We vary the parameter between (0.5, 1] in steps of $eps = 0.01$, resulting in a dataset of size $50 \times\ 50 \times\ 784$. The average Hamming distance consequently ranges from $(0, 392)$. The second dataset consists of binarized MNIST images at threshold 128. We use a greedy selection procedure—an iterative hill-climbing algorithm—to ensure a specified Hamming distance in each subset. Specifically, we randomly sample a pattern and accept it into our subset only if it increases or maintains the target HD across all previously accepted patterns. Since data structure inherent to MNIST limits average separation in subsets to $30 \leq HD \leq 190$, we create 53 subsets to evenly space out data points. This results in a dataset of size $53 \times 50 \times 784$.

Model Setup. We fix the number of neurons at $N = 784$ for a 28×28 input size, and use a rectified polynomial as in Krotov & Hopfield (2016), with n controlling the polynomial degree. We compute capacities for $n \in [6, 38]$ with step size 2 for values larger than 11. Our program also stores input patterns in-memory instead of training weights for simplicity and speed.

Capacity Measurement. We would need to train and test each model configuration on all possible datasets, which is very resource-intensive. In practice, we can only to test a representative subset that we need to identify: given a constructed subset of size S, we apply binary search to find memory capacity K_{max}, which is the largest $K < S$ such that our DAM can store and retrieve the first K patterns perfectly (normalized inner product = 1). This method reduces computational costs from $O(s)$ to $O(log\ s)$, s being the number of evaluated Ks. Importantly, we re-calculate the mean separation after we determine K_{max} to accurately represent separation in plots. We also exclude values of $K_{max} > 49$ from results to avoid distortions: since all subsets contain 50 patterns $S = 50$, memory capacity K_{max} saturates at 50 for high polynomial degrees of n even though the actual K_{max} might be larger than 50. Lastly, we also use early stopping based on K_{max} thresholds to optimize computational efficiency.

4 RESULTS

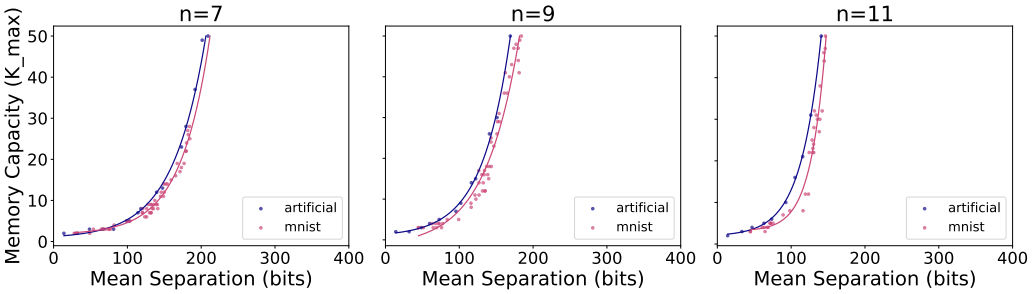

Figure 1: Comparing memory capacity scaling for synthetic and real data. We observe a slightly increasing constant difference between them at various levels of n.

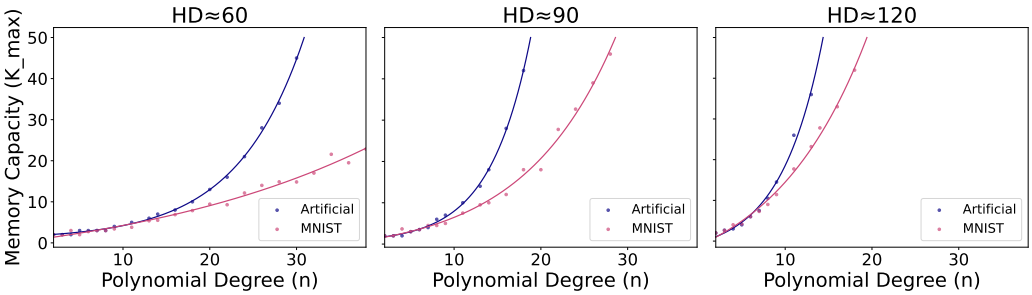

Figure 2: Capacity scaling for increasing n at different Hamming distance buckets.

As a baseline analysis, we confirm empirically that memory capacity $K_{\max}$ scales exponentially with the average separation in the dataset, shown in Appendix A.1. In Figure 1, we fix n and sweep the mean separation. As expected, capacity $K_{\max}$ grows sharply with separation, but synthetic data achieves consistently higher $K_{\max}$ for a given separation than MNIST. This effect becomes slightly more apparent at higher n. To better understand this dynamic, we plot K_{max} against polynomial degree n for both synthetic (blue) and MNIST (red) data across different mean Hamming distance buckets in Figure 2. Buckets have a HD leeway of 10 on both sides.

As expected, capacity grows exponentially with n. While pattern separation primarily drives capacity, correlated data is disadvantaged at higher n. At higher polynomial degrees, correlated data (MNIST) notably suffers an increasing capacity drop relative to synthetic data. Higher separation somewhat mitigates the strength of this divergence, but its effect is substantial for the the practical dataset in question: the MNIST train set with 3000 of each digit has an avg. $HD \approx 112$, where divergence for large n is still significant.

5 Discussion and future directions

With this paper, we bridge the gap between theoretical analyses and nuanced realities of real data. Our results highlight that feature correlations affect memory capacity of Hopfield networks significantly and are relevant for practical settings. Ignoring feature correlations can overestimate real-world capacity. These observations might also imply that Transformer architectures are constrained in their learning capacity in the same way. By extension, Large Language Models (LLMs) might also be affected by some type of bottleneck caused by patterns with high feature correlations.

We leave a rigorous theoretical explanation and intuition of how these effects come about to future work. Another, more immediate next research step could be how continuous-space input data affects capacity dynamics. It would also be interesting to study bounds below optimal memorization and more towards generalization, or recall under the assumption of non-zero error rates, both of which usually happen in practical machine learning scenarios.

References

Hamza Tahir Chaudhry, Jacob A. Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan. Long Sequence Hopfield Memory, November 2023. 1

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a Model of Associative Memory with Huge Storage Capacity. *Journal of Statistical Physics*, 168(2):288–299, July 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1806-y. 1, 2

Richard W. Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950. 2

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982. doi: 10.1073/pnas.79.8.2554. 1, 2

J J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, May 1984. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.81.10.3088. 1

Jerry Yao-Chieh Hu, Donglin Yang, Dennis Wu, Chenwei Xu, Bo-Yu Chen, and Han Liu. On sparse modern hopfield model. *Advances in Neural Information Processing Systems*, 36:27594–27608, 2023. 2

Jerry Yao-Chieh Hu, Dennis Wu, and Han Liu. Provably Optimal Memory Capacity for Modern Hopfield Models: Transformer-Compatible Dense Associative Memories as Spherical Codes, October 2024. 2, 5

Julia Kempe, Dmitry Krotov, Hilde Kuehne, Daniel Lee, and Sara A. Solla. New Frontiers in Associative Memories. In *ICLR 2025 Workshop Proposals*, December 2024. 2

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29:1172–1180, 2016. 1, 2

Carlo Lucibello and Marc Mézard. Exponential Capacity of Dense Associative Memories. *Physical Review Letters*, 132(7):077301, February 2024. doi: 10.1103/PhysRevLett.132.077301. 2

S.K. Manandhar and R. Sadananda. Effect of Hamming distance of patterns on storage capacity of Hopfield network. In *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.*, volume 1, pp. 253–256 vol.1, November 2002. doi: 10.1109/ICONIP.2002.1202172. 2

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need, July 2020. 1, 2

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2

A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, and Günter Klambauer. Modern Hopfield Networks and Attention for Immune Repertoire Classification, April 2020. 1

Dennis Wu, Jerry Yao-Chieh Hu, Teng-Yun Hsiao, and Han Liu. Uniform Memory Retrieval with Larger Capacity for Modern Hopfield Models, November 2024. 1, 2

A Appendix

A.1 Baseline Analysis.

We empirically confirm the known fact that memory capacity $K_{\max}$ scales exponentially with average separation in the dataset (Hu et al., 2024).

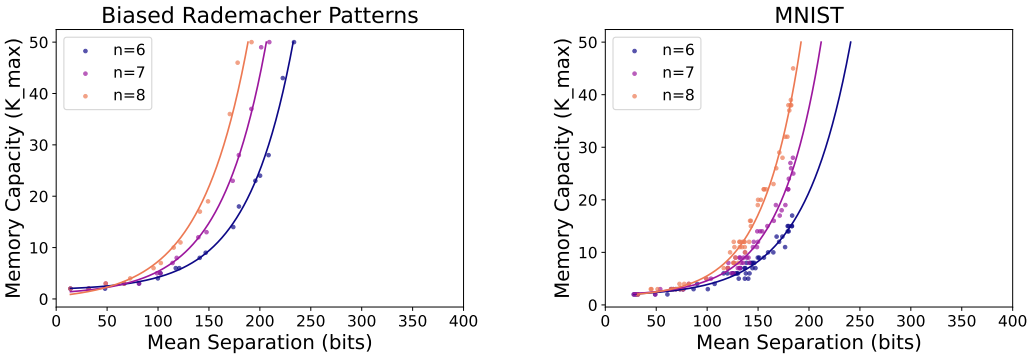

Figure 3: We calculate memory capacity K_{max} of Dense Associative Memory for 50 sets of biased Rademacher patterns and 53 sets of MNIST of increasing separation. Capacity grows exponentially with data separation for both types. This is in line with Hu et al. (2024)'s theoretical proposition that average separation scales in $O(\ln M)$, where M is K_{max} as memory capacity. When patterns are sufficiently separated, the chance of retrieval error drops dramatically due to lack of crosstalk. The general scaling relationship for feature-correlated data is not different from synthetic data, but there exist important differences which are explained in the results section.

A.2 Experiment and Code.

All numerical computations were performed using Apple's MLX framework, with patterns and network states represented as 32-bit floating-point arrays.

Our implementation of methods and experiments is publicly available on GitHub.