A Different Route to Exponential Storage Capacity

Elvis Dohmatob
FAIR, Meta

Abstract

Recent developments have sought to overcome the inherent limitations of traditional associative memory models, like Hopfield networks, where storage capacity scales linearly with input dimension. In this paper, we present a new extension of Hopfield networks that grants precise control over inter-neuron interactions while allowing control of the level of connectivity within the network. This versatile framework encompasses a variety of designs, including classical Hopfield networks, models with polynomial activation functions, and simplicial Hopfield networks as particular cases. Remarkably, a specific instance of our construction, resulting in a new self-attention mechanism, is characterized by quasi-exponential storage capacity and a sparse network structure, aligning with biological plausibility.

1 Introduction

Hopfield networks [1, 21, 15] are widely recognized as one of the most prominent mathematical models for *associative memory*. In this model, the retrieval of an item is possible by merely recognizing a fragment of its content, such as reconstructing a complete image from a partial view. These models have a rich history which goes back as far as [35, 40, 22, 16]. Apart from being analytically tractable, Hopfield networks are attractive to biologists because in principle, they can be implemented by the neurons and synapses in the brain. Indeed, a model could fail to be biologically plausible if the connections are not synaptic (neuron-to-neuron) connections or if the connectivity structure is not sparse (i.e is dense). Sparse connectivity means that each input neuron is connected to only a small number of other neurons (i.e, vanishing edge-density in the connectivity graph). This is the case of so-called *dilute* Hopfield networks [10, 36, 5, 23]. Finally, it is well-known that matrix models like Hopfield networks have limited *storage capacity* [16, 15, 34, 4]: they can only store and reliable retrieve cN memory patterns where c is an absolute constant.

The present study introduces a simple yet powerful approach for constructing general Hopfield networks with desirable properties. Our main contributions are summarized as follows.

- *Abstract Hopfield Networks.* Our proposed models utilize setwise connections based on collections, called *skeletons*, of subsets of input dimensions. We provide analytic expressions for the energy functional and update rule for such models, which extend the definition of traditional Hopfield networks [15]. These *Abstract Hopfield Networks* (or AHNs for short), encompass the classical Hopfield network [15], and its various extensions [17, 9, 6].
- *A New Type Self-Attention Layer.* As our second contribution, we show in Section 3 that a specific choice of skeleton leads to an AHN leads to a new type of self-attention layer which we call *Product-of-Sums Hopfield network (PSHN)* due to its structure. Note that a duality between traditional self-attention layers [38] and a certain type of Hopfield network has also been established [28]. Our proposed PSHN model enjoys the following desirable properties:
  - *High robust storage capacity*. Indeed, we show that our PSHN modern can store $e^{cN\log\log(N)/\log N}$ memories, which is (quasi-)exponential in the input dimension N
  - *Biological-Plausibility. It can be neurobiologically implemented by introducing k hidden neurons which have sparse synaptic connections to the N input neurons.*

See Appendix A for an overview of the relevant literature.

2 Abstract Hopfield Networks

In this section, we develop a simple and general extension of Hopfield networks.

The Skeleton. Given an arbitrary pattern $y = (y_1, \ldots, y_N) \in \{\pm 1\}^N$ and a subset $\sigma \subseteq [N] := \{1, 2, \ldots, N\}$ of neurons, define a variable $y_\sigma \in \mathbb{R}$ by $y_\sigma := \Pi_{n \in \sigma} y_n$, with the convention that $y_\emptyset = 1$. For example, if $\sigma = \{1, 5, 7\}$, then y_σ is the product $y_1 y_5 y_7$. Let $\mathfrak{S}$ be any (nonempty) collection of subsets of $[N]$. We shall call $\mathfrak{S}$ a *skeleton*, borrowing terminology from [6] which considered the special case $\mathfrak{S} = \binom{[N]}{D}$, the collection of all subsets of $[N]$ which contain D or less elements. A skeleton induces a correlation function on $\{\pm 1\}^N$ given by $\langle x, y\rangle_{\mathfrak{S}} := \sum_{\sigma \in \mathfrak{S}} x_\sigma y_\sigma$ for every pair of patterns $x, y \in \{\pm 1\}^N$. This can also be seen as an inner-product in the feature space given by the mapping $y \mapsto (y_\sigma)_{\sigma \in \mathfrak{S}}$. The parameters of the model are the memories $x^{(1)}, \ldots, x^{(M)}$. For any neuron $n \in [N]$, define

$$\partial_n \mathfrak{S} := \{\sigma \setminus \{n\} \mid \sigma \in \mathfrak{S} \text{ and } n \in \sigma\}. \tag{1}$$

In words, $\partial_n \mathfrak{S}$ is the collection of subsets of $[N]$ which don't contain the neuron n and can be turned into an element of $\mathfrak{S}$ by including n. For example, in the case of classical Hopfield networks [15],

$$\mathfrak{S} = \{\sigma \subseteq [N] \text{ s.t } |\sigma| = 2\} \text{ and } \partial_n \mathfrak{S} = \{\{n'\} \mid n' \in [N] \setminus \{n\}\} \cong [N] \setminus \{n\}, \forall n \in [N]. \tag{2}$$

Thus, in this case $|\partial_n \mathfrak{S}| = N - 1 \leq N$ for all $n \in [N]$, and we shall see later (Theorem 4.1) that this accounts for the linear storage capacity of the classical Hopfield network. One can therefore hope to obtain higher storage capacity by appropriate choices for the skeleton $\mathfrak{S}$.

Energy Functional and Update Rule. The energy of an input pattern $y \in \{\pm 1\}^N$ is given by

$$E(y) := -\sum_{\sigma \in \mathfrak{S}} \omega(\sigma) y_\sigma = -\sum_{\mu=1}^{M} \langle x^{(\mu)}, y\rangle_{\mathfrak{S}}, \text{ with } \omega(\sigma) := \sum_{\mu=1}^{M} x_\sigma^{(\mu)}. \tag{3}$$

The (one-step) update rule $T : \{\pm 1\}^N \to \{\pm 1\}^N$ is defined component-wise by

$$T_n(y) := \operatorname{sign}\left(\sum_{\mu=1}^{M} c_n^{(\mu)}(y) x_n^{(\mu)}\right), \text{ with } c_n^{(\mu)}(y) := \langle x^{(\mu)}, y\rangle_{\partial_n \mathfrak{S}} = \sum_{s \in \partial_n \mathfrak{S}} x_s^{(\mu)} y_s. \tag{4}$$

for any neuron $n \in [N]$. This construction is a generalization of the energy of the classical Hopfield network [15] by considering arbitrary multi-neuron interactions. For the particular case of classical Hopfield networks [15], it is easy to see from (2) that the energy (3) reduces to $E(y) = -\sum_{\mu=1}^{M} \sum_{n, n' \in [N],\, n' \neq n} x_n^{(\mu)} x_{n'}^{(\mu')} y_n y_{n'}$, while the update rule (4) reduces to update rule $T_n(y) = \operatorname{sign}\left(\sum_{\mu=1}^{M} x_n^{(\mu)} \sum_{n' \in [N],\, n' \neq n} x_{n'}^{(\mu)} y_{n'}\right)$, both of which are well-known formulae.

Definition 2.1. *Given a nonempty collection $\mathfrak{S}$ of subsets of neurons $[N]$, the energy* (3) *and update rule* (4) *define an Abstract Hopfield Network (AHN) with skeleton $\mathfrak{S}$.*

Thus, once the skeleton $\mathfrak{S}$ is prescribed, everything else about an AHN is completely determined. In particular, when $\mathfrak{S}$ is the collection of subsets of $[N]$ with at most D elements, i.e a simplicial complex of dimension D, we obtain the model proposed in [6].

Generality of Our Construction. We now show that for specific choices of the skeleton $\mathfrak{S}$, various well-known extensions of Hopfield networks are instances of our AHNs. This also allows us to recover the storage capacity of these models in a unified manner (Appendix D).

Theorem 2.1. *The classical Hopfield network [15], the polynomial Hopfield network [17, 9], and the simplicial Hopfield network [6],as well as all diluted versions of these networks are all instances of AHNs corresponding to specific choices for the skeleton $\mathfrak{S}$.*

Comparison to [24]. This recent work makes the observation that the update functional T for classical [15] and modern (dense) Hopfield networks [17, 9, 28] can be written as a composition of a linear projection (proj), a separation / activation function (sep), and similarity function (sim). For a general abstract Hopfield network (proposed by our work), it is not clear whether such a proj-sep-sim decomposition is always possible. Moreover, [24] doesn't show how any of these choices control for the properties (e.g storage capacity) of the resulting memory model. In contrast, our work proposes a different route to generalizing Hopfield network: by replacing pairwise connections by many-body connections based on an arbitrary collection of subsets (of input neurons), the skeleton. Moreover, the properties (storage capacity, etc.) of the resulting network derive from the topological properties of the skeleton (see Sections 4.1 and 5).

3 Product-of-Sums Hopfield Network (PSHN)

We now construct an instance of AHNs (Section 2) with remarkable properties like high storage capacity, biological plausibility, and connections to transformers [38].

The Skeleton. Let $G_1, \ldots, G_k$ form a partition of $[N]$, and consider an AHN whose skeleton $\mathfrak{S}$ is the collection of all subsets of $[N]$ which contain exactly one item from each G_i, i.e

$$\mathfrak{S} = \mathcal{T}(G_1, \ldots, G_k) := \{\sigma \subseteq [N] \text{ s.t } |\sigma \cap G_i| = 1 \text{ for all } i\}. \tag{5}$$

Observe that this is isomorphic to the Cartesian product of the G_i's in an obvious way. We call the resulting network a Product-of-Sums Hopfield network (PSHN), a terminology which will become clear later once we make its energy functional explicit. For fixed G_i's of equal size $N_i = N/k$ for all $i \in [k]$, we simply write $\mathcal{T}(N, k)$ for $\mathcal{T}(G_1, \ldots, G_k)$. The inherent product structure of such a skeleton enables it to integrate information across long-range interactions among input neurons. For example, we show in Appendix E.3 that these models can solve the XOR problem [25], a 3-dimensional problem known to be unsolvable with HNs comprising fewer than 4 neurons.

Energy and Update Rule. The energy functional (3) now takes on a special form.

Lemma 3.1. *For $\mathfrak{S} = \mathcal{T}(G_1, \ldots, G_k)$, the energy* (3) *is given by $E(y) = \sum_{\mu=1}^{M} E_\mu(y)$, where*

$$E_\mu(y) = -\prod_{i=1}^{k} \sum_{n \in G_i} x_n^{(\mu)} y_n, \text{ for any input pattern } y \in \{\pm 1\}^N. \tag{6}$$

The RHS of (6) justifies the name of the resulting network, namely: Product-of-Sums Hopfied network (PSHN). Figure 4 (Appendix C.3) gives a schematic illustration of (6).

Lemma 3.2. *For any pattern $y \in \{\pm 1\}^N$, group index $i \in [k]$, and memory index $\mu \in [M]$, define $a_i^{(\mu)}(y) := \sum_{n' \in G_i} x_{n'}^{(\mu)} y_{n'}$. Then, for any $n \in G_i$, the update* (4) *is $T_n(y) = \operatorname{sign}(\Delta_n(y))$, where*

$$\Delta_n(y) = \sum_{\mu=1}^{M} c_i^{(\mu)}(y) x_n^{(\mu)}, \text{ with } c_i^{(\mu)}(y) := \prod_{j \neq i} a_j^{(\mu)}(y). \tag{7}$$

A New Type of Self-Attention Layer. As already mentioned in the introduction, transformers (aka self-attention layers) are the core component of LLMs. We now show that a specific instance of our proposed PSHN model corresponds to a new type of self-attention layer. So, consider the special case where the skeleton is $\mathfrak{S} = \mathcal{T}(N, k)$, i.e where $N_i = N_1 = N/k$ for all i. Thus, $N = k \times N_1$. Stack the memories $(x^{(\mu)})_{\mu=1}^{M}$ into a matrix $X \in \mathbb{R}^{M \times N}$ and consider a batch of m queries $y^{(1)}, \ldots, y^{(m)}$ stacked in to a matrix $Q \in \mathbb{R}^{m \times N}$, and consider the following code snippet.

Code Listing 1: PyTorch GPU-friendly implementation of our PSHN model / self-attention layer.

```
X = X.reshape((M, k, N1))   # database of memories (e.g clean images)
Q = Q.reshape((m, k, N1))   # incoming queries (e.g noisy/occluded images)
Z = torch.einsum("mkg,Mkg->mMk", Q, X)   # correlate
C = Z.prod(axis=2, keepdims=True) / Z   # this replaces softmax operator
C = torch.nan_to_num(C, nan=0.)   # This is a trick to compute the ci's
TQ = torch.sign(torch.einsum("mMk,Mkg->mkg", C, X))   # output
TQ = TQ.reshape((m, N))   # original shape of input query matrix Q
```

Thanks to Lemma 3.2, the above code snippet computes the update rule T for our PSHN model. It can take full advantage of optimized matrix multiplication (GPUs). See Appendix C.2 for implementation tips. We come to the realization that in the case of equally sized groups, our proposed PSHN model is a new type of self-attention layer schematized like so

$$T(Q) = \operatorname{sign}(\Delta(Q)),\ \text{with}\ \Delta(Q) := \sigma(Q \cdot X) \cdot X, \tag{8}$$

where σ is the nonlinear mapping which produces C from Z in the above code snippet, "$\cdot$" denotes inner-product of tensors along an appropriate axis, and we have omitted the reshaping operators for clarity. Thus, our proposed PSHN model provides new perspectives for building transformers.

4 Analysis of Storage Capacity

As before, let $N \geq 1$ be the input dimension, that is the number of feature dimensions, or simply input *neurons* (e.g number of pixels in an image). Thus, for simplicity $[N] := \{1, 2, \ldots, N\}$ is the set of (indices of) neurons. As an example, this could be the number of pixels in an image. Let $x^{(1)}, \ldots, x^{(M)} \in \{\pm 1\}^N$ be a collection of M iid Rademacher memory patterns we wish to store. Fix a noise threshold $\theta \in [0, 1)$, and let $y^{(\mu)}$ be obtained from $x^{(\mu)}$ by setting to -1, the value of $\lfloor N\theta \rfloor$ coordinates of the latter selected uniformly at random and independently of all the memories.

Definition 4.1 (Robust Storage Capacity)**.** *We say that an associative memory network with update function $T : \{\pm 1\}^N \to \{\pm 1\}^N$ has θ-robust storage capacity $M_{N,\theta}$ if it holds that*

$$\lim_{\varepsilon \to 0^+} \lim_{N \to \infty} \frac{1}{M_{N,\theta}} \max\{1 \leq M \leq 2^N \mid \inf_{\mu \in [M]} \mathbb{P}(T(y^{(\mu)}) = x^{(\mu)}) \geq 1 - \varepsilon\} = 1. \tag{9}$$

In particular, $M_N := M_{\theta,0}$ is nonrobust storage capacity (i.e for retrieving uncorrupted memories).

4.1 A Generic Lower-Bound for Nonrobust Storage Capacity

The following is a generic lower-bound for the capacity of an AHN.

Theorem 4.1. *For any AHN with skeleton $\mathfrak{S}$, it holds that $M_N(\mathfrak{S}) \geq \underline{d}(\mathfrak{S})/(2 \log N)$, where $\underline{d}(\mathfrak{S}) := \min_n d(n)$, and $d(n) := |\partial_n \mathfrak{S}|$.*

In the presence of corruptions ($\theta \neq 0$), one cannot in general hope to get nontrivial lower-bounds for the robust storage capacity without further information about $\mathfrak{S}$. This is done in Appendix D.4.

4.2 A Lower-Bound for the PSHN Model

We now establish lower-bounds for the storage capacity of our proposed PSHN model (Section 3). Let us first consider the case of retrieving clean / uncorrupted patterns.

Theorem 4.2. *Consider a PSHN model with k groups each of size $N_1 = N/k$. Then, $M_N \geq N_1^{k-1}/(2 \log N)$. In particular, if $2 \leq N_1 = O(1)$, then $M_N \geq e^{cN}$ for some positive constant c.*

We now turn to the case of robust storage capacity Fix a corruption level $\theta \in [0, 1)$ and let $p := 1 - \theta/2$, $a := 1 - \theta$, and $b := e^{-1/(2a)}$. The next theorem is one of our main results.

Theorem 4.3. *Consider the PSHN model with k equal groups each of size $N_1 \geq C \log N$ with $C \geq 73/p$ where $p := 1 - \theta/2$. Then, $M_{N,\theta}(\mathfrak{S}) \geq (abN_1)^{k-1}$. In particular, for $N_1 = C \log N$, then $M_{N,\theta}(\mathfrak{S}) \geq e^{cN \cdot \frac{\log \log N}{\log N}}$, where c is a positive constant that only depends on θ and C.*

The logarithmic scaling $N_1 \asymp (1/p) \log N$ is crucial for achieving the quasi-exponential lower-bound. Results of some experiments are presented in Appendix B and C.1.

Computation of upper-bounds for the robust storage capacity of AHNs in general and the PSHN model in particular, is left for future work.

5 Biological Plausibility of the PSHN Model

We now provide strong arguments which show that our proposed PSHN model (described in Section 3) can be implemented in neurobiology (the brain), at least in principle.

Synaptic Connections. Note that our PSHN model can be realized with $O(k)$ hidden neurons for computing sums and products in (7), with direct synaptic connections to input neurons. Unlike the biological plausibility of the "sum" neurons, the "product neurons" which implement k-fold multiplication in (7), need some explanation because such an operation might not be implementable biologically by a single neuron. However, this operation can be carried out via a series of k 2-fold / binary multiplication neurons $(a, b) \to a * b$, which are known to be biologically plausible [13, 37]. In fact, k-fold multiplication is the basis of so-called sigma-pi networks [11] and pi-sigma networks [31, 12].

Remark 5.1. *As observed by a reviewer, we have hand-waved the unfortunate fact that multiplication with biological components is likely to be inexact, and errors might compound as the number of groups (i.e hidden neurons) k grows. However, we believe that if the errors are kept under control, the computations can still be done reliably. A rigorous analysis of this point is left for future work.*

Sparsity of Connexions. Concerning the connectivity structure, we see from (7) with $M = 1$ (i.e a single memory pattern) that the graph representing an PSHN model only contains $O(Nk)$ edges in total, corresponding to an edge density of $O(Nk/N^2) = O(k/N)$. If the number of groups k is of order $o(N)$ (e.g $k = O(N/\log N)$), then the computation graph for the corresponding PSHN model is extremely sparse (vanishing edge density), and thus is biologically plausible. Importantly our construction can simultaneously achieve sparsity and (quasi-)exponential robust storage capacity. In contrast, diluted HNs [10, 5, 6] also enjoy sparse connectivity, but can only boast of polynomial storage capacity.

Number of Connections vs Storage Capacity. Motivated by [6] which showed that it is important to compare storage capacity of different Hopfield Networks with the same total number of connections, let us now provide a back-of-envelope (but rigorous) calculation which shows that compared to classical, polynomial, or simplicial Hopfield networks, our proposed PSHN network has a much higher storage capacity w.r.t number of connections. Indeed, the number of connections in a polynomial (resp. simplicial) HN is of order N^k where $k = d - 1$ and d is the degree of the polynomial (dimension of the simplicial complex). The (robust) storage capacity of these models is at most of order N^{k-1}. In contrast, the number of connections in our PSHN model with $k = d - 1$ groups of equal size $N_1 = N/k$ is of order Nk (i.e, much smaller!), and the storage capacity is of order N_1^k. Thus if we fix the number of connections to say N^2 (as in [6]), then the storage capacity of our network is quasi-exponential $e^{N \log\log(N)/\log N}$ (since the requirement k groups of equal size $N_1 = C \log N$ in Theorem 4.3 amounts to $O(Nk) = O(N^2/(C \log N)) = O(N^2/\log N)$ connections, which is definitely less than N^2 for large N), while the robust storage capacity of classical, polynomial, or simplicial Hopfield networks is only of order N (i.e linear).

Thus, simultaneously, our proposed PSHN model is biologically plausible and with (quasi-) exponential robust storage capacity. Our arguments for biological plausibility are stronger than the indirect arguments presented in [18] for polynomial and exponential Hopfield networks.

6 Concluding Remarks

In this work, we have introduced a versatile framework for extending classical Hopfield networks by incorporating long-range interactions defined via collections of subsets of input features, called "skeletons". We have also demonstrated that many classical Hopfield network extensions are specific examples of our broader construction corresponding to specific choices of the skeleton. Importantly, one specific instantiation of our model introduces a novel self-attention layer (PSHN) with exponential storage capacity. Moreover, we have shown that the later model is biologically-plausible: it can be implemented by sparse two-body synaptic connections between neurons, providing a high storage capacity w.r.t for a fixed number of connections, compared to previous extensions of Hopfield networks.

Our findings open new possibilities for enhancing machine learning models with powerful associative memory modules. This direction will be explored further in future work.

References

[1] S. Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, 1972.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255.

[4] Anton Bovier. Sharp upper bounds on perfect retrieval in the hopfield model. *Journal of Applied Probability*, 36(3):941–950, 1999.

[5] Anton Bovier and Véronique Gayrard. Rigorous bounds on the storage capacity of the dilute Hopfield model. *Journal of Statistical Physics*, 69(3-4), November 1992.

[6] Thomas F. Burns and Tomoki Fukai. Simplicial hopfield networks. In *ICLR*, 2023.

[7] Rishidev Chaudhuri and Ila Fiete. Bipartite expander hopfield networks as self-decoding high-capacity error correcting codes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[8] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016.

[9] 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*, 168:288–299, 2017.

[10] B. Derrida and J. P. Nadal. Learning and forgetting on asymmetric, diluted neural networks. *Journal of Statistical Physics*, 1987.

[11] J.A. Feldman and D.H. Ballard. Connectionist models and their properties. *Cognitive Science*, 1982.

[12] Joydeep Ghosh and Yoan Shin. Efficient higher-order neural networks for classification and function approximation. *Int. J. Neural Syst.*, 1992.

[13] Lukas N. Groschner, Jonatan G. Malis, Birte Zuidinga, and Alexander Borst. A biophysical account of multiplication by a single neuron. *Nature*, 2022.

[14] Christopher J. Hillar, Tenzin Chan, Rachel Taubman, and David Rolnick. Hidden hypergraphs, error-correcting codes, and critical learning in hopfield networks. *Entropy*, 23, 2021.

[15] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79 (8):2554–2558, April 1982. ISSN 0027-8424.

[16] Teuvo Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, C-21(4): 353–359, 1972.

[17] Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[18] Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *ICLR*, 2021.

[19] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2010.

[20] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017.

[21] W.A. Little. The existence of persistent states in the brain. *Mathematical Biosciences*, 1974.

[22] H. C. Longuet-Higgins, D. J. Willshaw, and O. P. Buneman. Theories of associative recall. *Quarterly Reviews of Biophysics*, page 223–244, 1970.

[23] Matthias Löwe and Franck Vermet. The hopfield model on a sparse erdös-renyi graph. *Journal of Statistical Physics*, 2011.

[24] Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. PMLR, 2022.

[25] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[26] Charles M. Newman. Memory capacity in neural network models: Rigorous lower bounds. *Neural Networks*, 1:223–238, 1988.

[27] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016.

[28] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlovic, Geir Kjetil Sandve, Victor Greiff, David P. Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. *CoRR*, 2020.

[29] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. 2007.

[30] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, page 437–446, USA, 2012. Society for Industrial and Applied Mathematics.

[31] Y. Shin and J. Ghosh. The pi-sigma network: an efficient higher-order neural network for pattern classification and function approximation. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, 1991.

[32] Matthew Smart and Anton Zilman. On the mapping between hopfield networks and restricted boltzmann machines. In *ICLR*, 2021.

[33] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel Distributed Process*, 1, 1986.

[34] Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.*, 1990.

[35] K. Steinbuch. Die lernmatrix. *Kybernetik*, 1961.

[36] A Treves and D J Amit. Metastable states in asymmetrically diluted hopfield networks. *Journal of Physics A: Mathematical and General*, jul 1988.

[37] Juan C. Valle-Lisboa, Andrés Pomi, and Eduardo Mizraji. Multiplicative processing in the modeling of cognitive activities in large neural networks. *Biophysical Reviews*, 2023.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[39] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[40] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960–962, 1969.

Appendix

A Related Work

We now overview a cross-section of the relevant literature. Extensions of Hopfield networks that break the linear of classical Hopfield networks [15] have been proposed. [26], then [17] have shown that a modification of the energy of the classical Hopfield network leads to a polynomial increase in memory capacity. This has been followed up by [9], and more recently [28] who proposed another modification leading to exponential memory capacity. These so-called modern / dense Hopfield networks use a nonlinear activation function to make the energy and an update rule that is more sharply peaked around the stored memories in the space of neuron's configurations compared to the traditional Hopfield network. See [18] for a detailed review. One of the main insights from this recent resurgence of Hopfield networks is their connection to transformers [38], which have become the core components in the design of of large language models (LLMs) for example. Indeed, it was shown in [28] that the update / retrieve function in their proposed Hopfield network amounts to the self-attention layer in a transformer model [2, 8, 27, 20, 38]. This connection provides hope for a theoretical understanding and explanation of the *emergent* capabilities of modern LLMs [39]. Refer to Table 1 for a comparison between different types of Hopfield networks.

[14] established a duality between certain error-correcting codes on hyper-graphs, and Hopfield networks. This link allowed them to derive an extension of Hopfield networks with quasi-exponential storage capacity. On a similar route, [6] considered an extension of the traditional Hopfield network wherein the complete graph characterizing the connectivity structure of the neurons is replaced by a simplicial complex. Finally, let us also mention [7, 32] who have established a direct mapping between Hopfield networks (and extensions thereof) and Restricted Boltzmann Machines (RBMs) [33, 29] whereby the memory patterns of stored by the Hopfield network correspond to parameters that control the activity of the hidden layer in the RBM.

Type of HN	**Reference Paper**	**Bio Plausible**	**Robust Storage Capacity**
Classical	[15]	Yes	$cN/\log N$ (linear)
Polynomial	[17]	Yes†	$cN^{d-1}/\log N$ (poly)
Simplicial	[6]	Yes*	$cN^{D-1}/\log N$ (poly)
Exponential	[9, 28]	Yes†	$\exp(cN)$ (expo)
Little-Hopfield 2	[14]	–	$\exp(\frac{cN}{\log N})$ (quasi-expo)
PSHN	Our work	Yes	$\exp(\frac{cN \log\log N}{\log N})$ (quasi-expo)

Table 1: Comparing different types of Hopfield networks (HNs) according to their biological (im)plausibility and robust storage capacity (formally defined in Section 4.1). Our proposed product-of-sums Hopfield network (PSHN) is described in Section 3. For the polynomial (resp. simplicial) Hopfield network, d (resp. D) is the degree (resp. dimension). The c's in the exponents of the storage capacity bounds are positive constants which typically depend on the level of robustness required (and also on d and D in the case of polynomial and simplicial Hopfield networks respectively). Yes† means the corresponding Hopfield network is only biologically plausible in an indirect sense: it provides an effective description for a more microscopic theory that has additional (hidden) neurons and only requires two-body interactions between them [18]. Finally, the simplicial Hopfield network [6] only becomes biologically plausible when diluted, i.e a large number of connections are suppressed. This reduces its storage capacity.

B An Experiment: Storing and Retrieving Correlated Patterns

We empirically demonstrate our theoretical results by running a small experiment on the popular MNIST dataset [19]. For this computer vision dataset, each of the 70K examples is a gray-scale image of resolution 28×28 pixels. We sample $M = 10$K out of 70K images from this dataset and examine how these can be stored and retrieved by our Product-of-Sums Hopfield network (PSHN) model described in Section 3.

Experimental Setup. We normalized the intensity values of the image so that they are ± 1. Thus each of the $M = 10$K images is now a vector in $\{\pm 1\}^N$ with $N = 784$. For each memory pattern

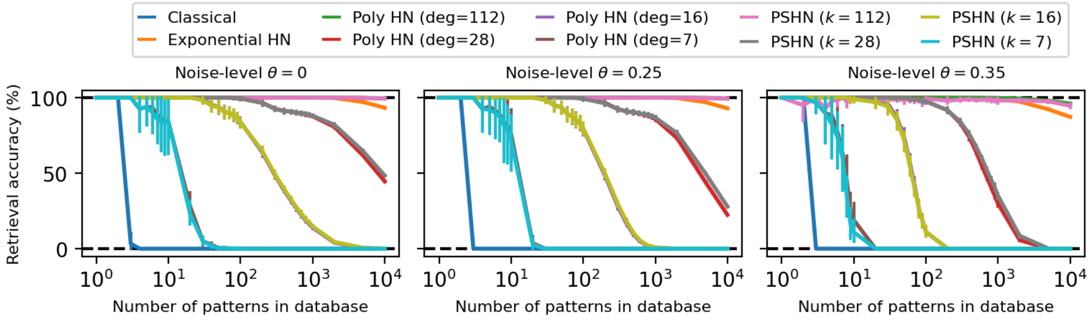


Figure 1: Comparing storage capacity on MNIST. Our proposed PSHN model (Section 3) is instantiated with k groups each of size $N_1 = N/k$, where $N = 784$. The y-axis represents how many memory patterns are perfectly recovered. Error bars are variations across 10 runs (different sub-samplings of 10K out of 70K images). For this experiment, we see that the optimal number of groups is $k = 112$, each of size $N_1 = N/k = 7$. We also show results for the classical Hopfield network, exponential, and polynomial Hopfield networks discussed in the introduction (Section 1).

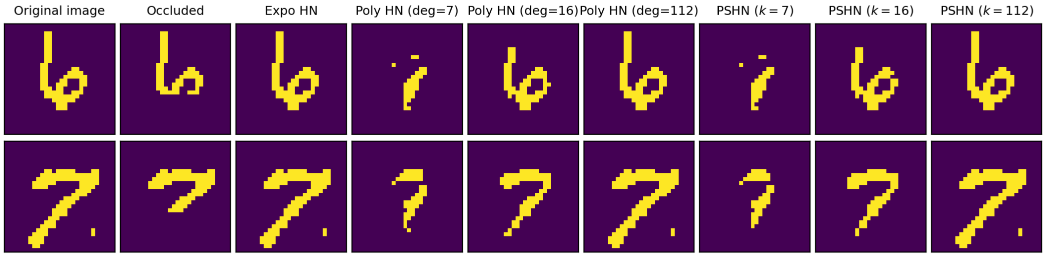


Figure 2: Visual Inspection of reconstructed image for each method. As the order of the long-range interactions in the model (k for our PSHN model and degree "deg" for Poly HN) increases, the model moves from feature-extractors to prototype-builders. This is in accordance with the "Feature vs Prototype" theory advocated in [17]. See Figure 3 for additional results.

$x^{(\mu)} \in \{\pm 1\}^N$, the intensity values of a fraction θ of the pixels (bottom-most) are set to -1. We do this for $\theta = 0$ (corresponding to nonrobust storage), $\theta = 0.25$, and $\theta = 0.35$. We create instances of our PSHN model with k equally sized groups, for different values of k ranging in $\{7, 16, 28, 112\}$. The experiment is run 10 times (on a machine with a single T4 GPU), each time with a different random sub-sampling of $M = 10$K out of the 70K images in the MNIST dataset.

Empirical Results. Figure 1 reports robust storage capacity for our model alongside alongside other types of high-capacity Hopfield network discussed in Section 1. Notice how the performance for our PSHN model matches a polynomial Hopfield network of degree k, in accordance to Corollary D.1. For $k = 7$, we observe the best performance for our model, which is consistent with the (quasi-)exponential storage capacity in established in Theorem 4.3. The good performance for the exponential Hopfield network [9, 28] observed in the figure is also consistent with its exponential storage capacity. These models and ours rely on long-range interactions between features to cope with the strong correlations present in the data. This is unlike the classical Hopfield network [15] which only relies on short-range (pairwise) interactions.

All the models had comparable running times. The entire experiment (10 runs of all models) executes in under 30 minutes on a single T4 GPU. See supplemental for code to reproduce all figures.

C Miscellaneous

C.1 Additional Results: Visual Inspection for Retrieval of MNIST Images

Figure 2 is a longer version of Figure 1. See supplemental for code to reproduce all figures.

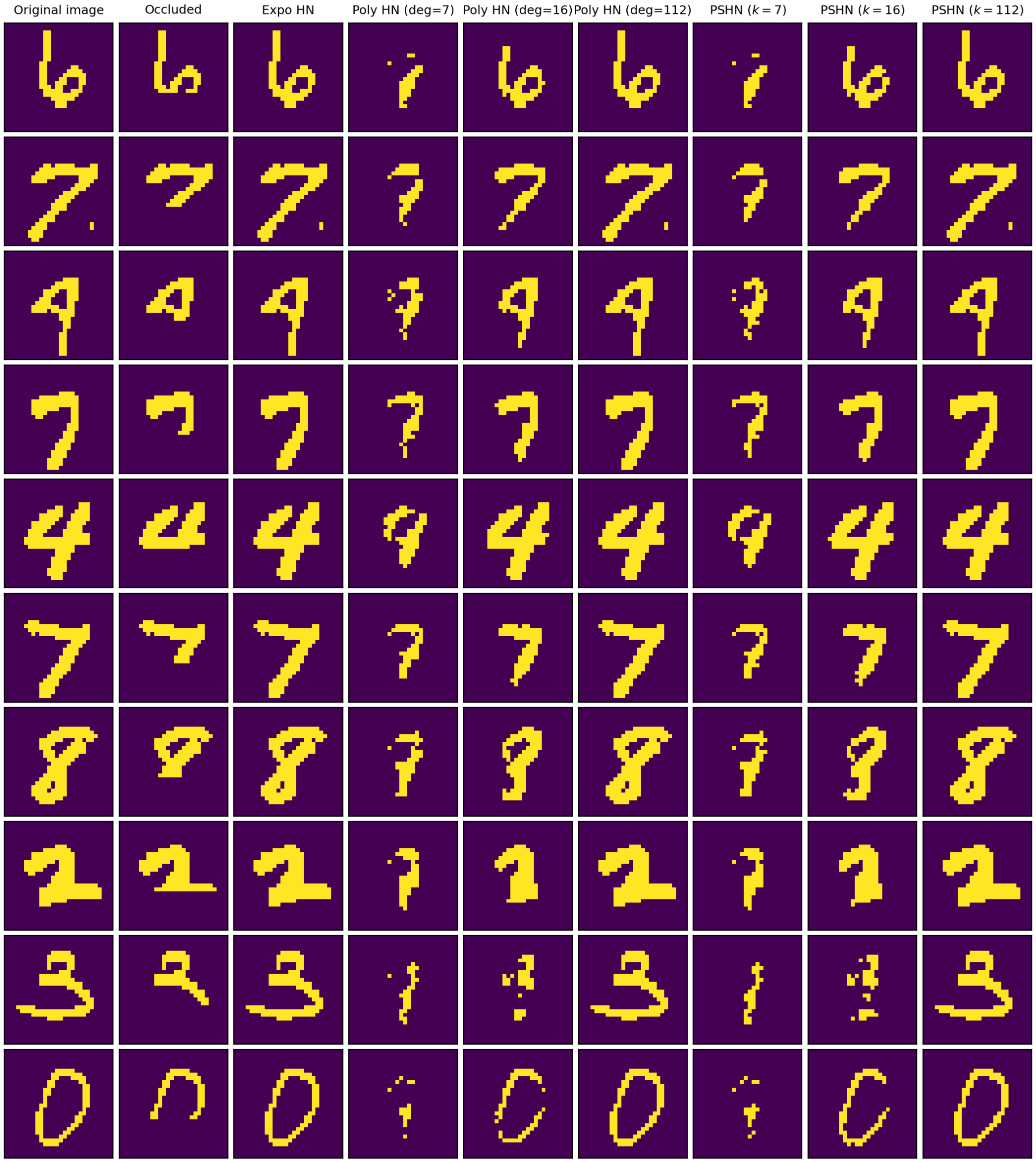


Figure 3: Observe that, as the order of the long-range interactions in the model (k for our PSHN model and degree "deg" for Poly HN) increases, the model moves from feature-extractors to prototype builders. This is in accordance with the "Feature-Extractor vs Prototype" theory advocated in [17, 18].

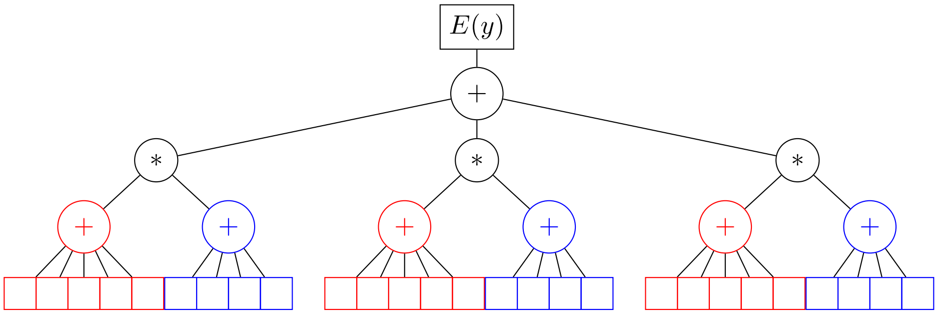
C.2 Technical Details for Implementing our PSHN Self-Attention Layer

Observe that the code snippet in Code Listing 1 (Section 3 of the main text) is vectorizable and can take full advantage of optmized linear algebra on GPUs. Also, line 3 of Code Listing 1 is effectively doing matrix multiplication of k pairs of $m \times N_1$ and $M \times N_1$ matrices, and can be carried out in parallel on GPUs, for example. In particular, the case $k = 1$ reduces to the usual matrix product $Z = QX^\top$. A similar comment applies to line 6. All in all, the complexity of our proposed PSHN model is comparable to a traditional self-attention layer [38], and to classical dense associative memory models [17, 9, 28].

C.3 Schematic Illustration of Energy Functional of PSHN Model

Figure 4 shows a schematic representation of the energy $E(y)$ according to (6) for an input ± 1-pattern y in $N = 9$ dimensions, according to formula (6). Here, there are $M = 3$ memory patterns $x^{(1)}$, $x^{(2)}$, and $x^{(3)}$. Each of the M horizontal blocks of N cells each corresponds to an element-wise product $z^{(\mu)} = x^{(\mu)} \odot y \in \{\pm 1\}^N$, for each $\mu \in [M]$. For this example, the skeleton of the PSHN model is as in (5), with $k = 2$ groups of neurons $G_1 = \{1, 2, 3, 4, 5\}$ and $G_2 = \{6, 7, 8, 9\}$. Each colored subgraph can be seen as a *tokenizer* which correlates the input y and memory patterns $x^{(\mu)}$ along a the input dimensions corresponding to subset of neurons G_i.

Figure 4: Energy Computation for PSHN Model.



$$
\begin{aligned}
E(y) &= -\sum_{\mu=1}^{M}\prod_{i=1}^{k}\sum_{n\in G_i} x_n^{(\mu)} y_n \\
&= -\sum_{\mu=1}^{M} (x_1^{(\mu)}y_1 + x_2^{(\mu)}y_2 + x_3^{(\mu)}y_3 + x_4^{(\mu)}y_4 + x_5^{(\mu)}y_5)(x_6^{(\mu)}y_6 + x_7^{(\mu)}y_7 + x_8^{(\mu)}y_8 + x_9^{(\mu)}y_9),
\end{aligned}
\tag{10}
$$

D Analysis of Storage Capacity for Abstract Hopfield Networks

D.1 Technical Notations

Let us recall some notations used in the manuscript and define others which we will continue to use in this appendix. We will continue to use $[N]$ to denote the set of integers $\{1, 2, \ldots, N\}$. The collection of subsets of $[N]$ with d exactly d elements is denoted $\binom{[N]}{d}$, while $\binom{[N]}{\leq D} := \cup_{d=0}^{D}\binom{[N]}{d}$ is the collection of subsets of $[N]$ with D elements or fewer. As simplicial complex K on $[N]$ is a collection of subsets of subsets of K such that if $s_0 \subseteq s \in K$, then $s_0 \in K$. For example, $\binom{[N]}{\leq D}$ is a simplicial complex of dimension D.

Given nonnegative real functions f and g, we write $f(N) \lesssim g(N)$, or equivalently $f(N) = O(g(N))$ to mean that there exists an absolute constant C such that $f(N) \leq Cg(N)$ for sufficiently large N, while $f(N) \asymp g(N)$ means $f(N) \lesssim g(N) \lesssim f(N)$. Finally, $f(N) = o(g(N))$, or equivalently $f(N) \ll g(N)$, means $f(N)/g(N) \to 0$ as $N \to \infty$. In particular, $f(N) = O(1)$ means f is bounded, while $f(N) = o(1)$ means $f(N) \to 0$ in the limit $N \to \infty$. For example, $\log N = o(N)$ since $\log(N)/N \to 0$ in the limit $N \to \infty$.

$X \overset{D}{=} Y$ denotes equality in distribution of two random variables X and Y.

D.2 Generic Signal to Noise Ratio Computation

Let us consider the problem of robustly storing the first pattern $x^{(1)} \in \{\pm 1\}^N$. Let the pattern y be a corrupted version of $x^{(1)}$ as in Definition 4.1. That is we will to study the probability that $\mathbb{P}(T(y) = x^{(1)})$. The following argument is adapted from [4] which established sharp bounds on the storage capacity of classical Hopfield network (corresponding to $\mathfrak{S} = \binom{[N]}{2}$ in our case). First observe that the update $T_n(y)$ for the the nth neuron satisfies $T_n(y)x_n^{(1)} = \text{sign}(\Delta_n)$, where

$$\begin{aligned}\Delta_n &= \sum_{\mu=1}^{M}\sum_{s\in\partial_n\mathfrak{S}} x_n^{(\mu)}x_s^{(\mu)}y_s x_n^{(1)} = \sum_{s\in\partial_n\mathfrak{S}}(x_n^{(1)})^2 x_s^{(1)}y_s + \sum_{s\in\partial_n\mathfrak{S}}\sum_{\mu=2}^{M} x_n^{(1)}y_s x_n^{(\mu)}x_s^{(\mu)}\\ &= \sum_{s\in\partial_n\mathfrak{S}} x_s^{(1)}y_s + \sum_{s\in\partial_n\mathfrak{S}}\sum_{\mu=2}^{M} x_n^{(\mu)}x_s^{(\mu)},\ \text{since } x_n^{(1)}y_s x_n^{(\mu)}x_s^{(\mu)} \overset{D}{=} x_n^{(\mu)}x_s^{(\mu)}\ \text{(Lemma D.2)} \qquad (11)\\ &= \underbrace{A_n}_{signal} + \underbrace{Z_n}_{noise}\end{aligned}$$

where the signal term A_n and the noise term Z_n (also known as the *crosstalk* term) are given by

$$A_n := \sum_{s\in\partial_n\mathfrak{S}} x_s^{(1)}y_s, \tag{12}$$

$$Z_n := \sum_{s\in\partial_n\mathfrak{S}}\sum_{\mu=2}^{M} x_n^{(\mu)}x_s^{(\mu)}. \tag{13}$$

Note that the noise term Z_n as given in (13) is a sum of $(M-1)\cdot|\partial_n\mathfrak{S}| =$ iid Rademacher random variables $x_n^{(\mu)}x_s^{(\mu)}$, and so in terms of (anti)concentration, we expect it to behave like a properly scaled Gaussian random variable. In fact,

Lemma D.1. *A_n and Z_n are statistically independent, and we have the following identities*

$$\mathbb{E}\,Z_n = 0, \tag{14}$$

$$\text{var}(Z_n) = (M-1)d(n), \tag{15}$$

$$\mathbb{E}\,A_n = \sum_{s\in\partial_n\mathfrak{S}}(1-\theta)^{|s|}. \tag{16}$$

In particular, if $\theta = 0$, then $A_n = d(n) := |\partial_n\mathfrak{S}|$, i.e deterministic.

The hard part of the business is that, in the noisey regime where $\theta = 0$, the variance of A_n will in general depend intricately on the topology of the skeleton $\mathfrak{S}$.

Proof. Note that $x_s^{(1)}y_s = \prod_{n'\in s} x_{n'}^{(1)}y_{n'}$, which is a product of $|s|$ idd random variables, each with mean $(1-\theta)$. The claimed expression for $\mathbb{E}\,A_n$ follows.

We now analyze the noise term Z_n. Observe that Z_n and $-Z_n$ have the same distribution. In particular, Z_n has zero mean. Also, the sum in the equation defining Z_n is a sum of $(M-1)\cdot|\partial_n\mathfrak{S}| = (M-1)\cdot d(n)$ iid Rademacher random variables $x_s^{(\mu)}x_n^{(\mu)}$, where and $2 \le m \le M$ and $s \in \partial_n\mathfrak{S}$. We deduce that Z_n has mean 0 and variance given b $\text{var}(Z_n) = (M-1)\cdot d(n)$. □

Lemma D.2. *Suppose $n \in [N]$ and $s \subseteq [N]$ such that $n \notin s$. Then, for any memory pattern index $\mu \neq 1$, it holds that $x_n^{(1)}y_s x_n^{(\mu)}x_s^{(\mu)} \overset{D}{=} x_n^{(\mu)}x_s^{(\mu)}$, where $y \in \{\pm 1\}^N$ is obtained from the pattern $x^{(\mu)}$ as in Definition 4.1.*

D.3 Proof of Theorem 4.1

Thanks to (11), in order to ensure the "good" event $T_n(x^{(1)})x_n^{(1)} \ge 0$, we need the std of $A_n + Z_n$, to be dominated by its mean. Thanks to Lemma D.1 we know that in the nonrobust regime ($\theta = 0$), the former is $\sqrt{(M-1)d(n)}$ and the former is $d(n)$. Thus, it would from Chebychev's inequality that

$$\begin{aligned}\mathbb{P}(T(x^{(1)}) \neq x^{(1)}) = \mathbb{P}(\exists n \in [N]\ \text{s.t}\ T_n(x^{(1)})x_n^{(1)} \le 0) &\le \sum_{n=1}^{N}\mathbb{P}(Z_n \ge d(n))\\ &\le \sum_{n=1}^{N}\frac{M}{d(n)} \le N\cdot\frac{M}{\underline{d}(\mathfrak{S})}.\end{aligned}$$

This would give the storage capacity bound $M_N(\mathfrak{S}) \gtrsim \underline{d}(\mathfrak{S})/N^{1+o(1)}$, which is sub-optimal with regards to our target. Instead, we use a slightly more involved argument, following a line thought similar to [4]. First, a standard union-bound gives

$$\begin{aligned}\mathbb{P}(T(x^{(1)}) \neq x^{(1)}) &= \mathbb{P}(\exists n \in [N] \text{ s.t } T_n(x^{(1)})x_n^{(1)} \leq 0)\\ &\leq \mathbb{P}(\exists n \in [N] \text{ s.t } Z_n \geq d(n)) \leq \sum_{n=1}^{N} \mathbb{P}(Z_n \geq d(n)).\end{aligned}$$

Now, for any $n \in [N]$, we know that $Z_n = \sum_{s \in \partial_n \mathfrak{S}} x_n^{(\mu)} x_s^{(\mu)}$ has the same distribution as $\sum_{s \in \partial_n \mathfrak{S}} x_s^{(\mu)}$, because $x_n^{(\mu)}$ and $x_s^{(\mu)}$ are independent. Moreover, the later is a sum of $(M-1) \cdot d(n)$ idd Rademacher random variables $x_s^{(\mu)}$, and so exhibits Gaussian concentration around zero [3]. We deduce that

$$\begin{aligned}\mathbb{P}(T(x^{(1)}) \neq x^{(1)}) &\leq \sum_{n=1}^{N} \mathbb{P}\left(\sum_{s \in \partial_n \mathfrak{S}} x_s^{(\mu)} \geq d(n)\right)\\ &\leq \sum_{n=1}^{N} \exp\left(-\frac{d(n)^2}{2Md(n)}\right) = \sum_{n=1}^{N} \exp\left(-\frac{d(n)}{2M}\right) \qquad (17)\\ &\leq N \cdot \exp\left(-\frac{\underline{d}(\mathfrak{S})}{2M}\right).\end{aligned}$$

To make the RHS go to zero in the limit $N \to \infty$, it suffices that $\underline{d}(\mathfrak{S})/M \geq (2+\gamma)\log N$, i.e $M \leq \dfrac{\underline{d}(\mathfrak{S})}{(2+\gamma)\log N}$ where γ is a positive constant. Since γ is arbitrary, we conclude that the storage capacity is lower-bounded as claimed. □

D.4 Robust Storage Capacity

In the case of nonzero corruption level $\theta \in (0,1)$, quantitative analysis of storage capacity must exploit further information about the topology of the skeleton $\mathfrak{S}$. Indeed, we cannot generally hope to get nontrivial lower-bounds for robust storage capacity of an AHN without assumptions on the skeleton $\mathfrak{S}$. For example, the AHN induced by the largest possible collection of subsets of neurons, namely $\mathfrak{S} = 2^{[N]}$, has exponential nonrobust capacity $M_N(\{0,1\}^N) \geq e^{cN}$. This follows from Theorem 4.1 and the fact that $d_{\{0,1\}^N}(n) = 2^{N-1}$ for any $n \in [N]$. However, the basin of attraction around each stored pattern has width zero! To see this, note that $2^{-N}\sum_{\sigma \in 2^{[N]}} x_\sigma y_\sigma = \delta_{x=y}$ by Lemma E.1. Thus, the corresponding energy E in (3) is either 0 or 2^N. Consequently, it is unable to recover any stored pattern with at a nonzero corruption level, i.e $M_{N,\theta}(\{0,1\}^N) = 0$ for all $\theta \in (0,1)$.

Definition D.1 (Moments). *For any $n \in [N]$ and integer $i \geq 0$, define*

$$\mu_{n,i}(\mathfrak{S}) := \max_{s_0 \in \binom{[N]}{i}} |\{s \in \partial_n \mathfrak{S} \mid s_0 \subseteq s\}|. \qquad (18)$$

Thus, there is no subset of $[N]$ with i elements, contained in more than $\mu_{n,i}(\mathfrak{S})$ elements of $\partial_n \mathfrak{S}$.

Note that in particular, $\mu_{n,0}(\mathfrak{S}) = d(n) := |\partial_n \mathfrak{S}|$. Under the following condition, we can derive a generic lower-bound for the robust storage capacity of an abstract Hopfield network.

Condition D.1 (Smooth Skeleton). *(A) $\max_{\sigma \in \mathfrak{S}} |\sigma| - 1 \leq q$ with $q/\log N \to 0$ as $N \to \infty$. (B) $\max_{1 \leq i \leq q} N_1^i \mu_{n,i} = O(d(n))$, for all $n \in [N]$ and some $N_1 \geq N^{c_0}$ and positive constant c_0.*

The above condition means for any $i \leq q$, there is no i-element subject of $[N]$ which is contained in more than a fraction N^{-Ci} of $s \in \partial_n \mathfrak{S}$, where C is an absolute positive constant.

Theorem D.1. *Fix a corruption level $\theta \in [0,1)$, and consider an AHN with skeleton $\mathfrak{S}$ verifying Condition D.1. The θ-robust storage capacity is given by $M_{N,\theta}(\mathfrak{S}) \geq c(1-\theta)^{2q}\underline{d}(\mathfrak{S})/\log N$, for some positive constant c which only depends on θ.*

As shown in Section 2, the Hopfield network [15, 4], polynomial Hopfield networks [17, 9], and simplifical Hopfield networks [6] are all instances of our AHNs with appropriate choices of the skeleton $\mathfrak{S}$. Moreover, the skeletons of these models verify Condition D.1 under certain conditions.

Corollary D.1. *If $k, d, D = o(\log N)$ as $N \to \infty$, the storage capacity bounds in Table 2 hold.*

Type of HN $\mathfrak{S}$	$\underline{d}(\mathfrak{S})$	q	N_1	Robust Storage Capacity
Classical	N	1	N	$M_{N,\theta} \geq c(1-\theta)^2 N/\log N$
Polynomial	N^{d-1}	$d-1$	N	$M_{N,\theta} \geq c(1-\theta)^{2(d-1)} N^{d-1}/\log N$
Simplicial	N^{D-1}	$D-1$	N	$M_{N,\theta} \geq c(1-\theta)^{2(D-1)} N^{D-1}/\log N$
PSHN	N_1^{k-1}	$k-1$	N_1	$M_{N,\theta} \geq c(1-\theta)^{2(k-1)} N_1^{k-1}/\log N$

Table 2: Combined with Theorem 2.1, our Corollary D.1 recovers lower-bounds previously established in [15, 26, 17, 9, 6]. The PSHN model listed in the table is with k groups of equal size $N_1 = N/k$.

From the corollary, we see that in the small k regime, the storage capacity of our PSHN model with k equally sized groups behaves like that of a polynomial Hopfield network of degree k.

E Some Calculations Related to Our PSHN Model

E.1 Proof of Lemma 3.1

Starting from the general formula (3), we have

$$E(y) = \sum_{\sigma \in \mathfrak{S}} \omega(\sigma) y_\sigma = \sum_{\mu} \sum_{\sigma \in \mathfrak{S}} z_\sigma, \text{ where } z_\sigma := \prod_{n \in \sigma} z_n \text{ and } z_n := x_n^{(\mu)} y_n. \tag{19}$$

Now, by basic algebra, one has

$$\prod_{i=1}^{k} \sum_{n \in G_i} z_n = \sum_{n_1 \in G_1, \ldots, n_k \in G_k} z_{n_1} z_{n_2} \ldots z_{n_2} = \sum_{\sigma \in \mathfrak{S}} z_\sigma, \tag{20}$$

and the result follows upon combing with (19). □

E.2 Proof of Lemma 3.2

Indeed, from (4), we know that $T_n(y) = \text{sign}(\Delta_n(y))$, where $\Delta_n(y) := \sum_{\mu=1}^{M} x_n^{(\mu)} \sum_{s \in \partial_n \mathfrak{S}} x_s^{(\mu)} y_s$, where $\partial_n \mathfrak{S}$ is as defined in (1). Now, because $\mathfrak{S} = \mathcal{T}(G_1, \ldots, G_k) := \{\sigma \subseteq [N] \text{ s.t } |\sigma \cap G_j| = 1\, \forall j\} \cong \prod_j G_j$, it is clear that if $n \in G_i$, then

$$\partial_n \mathfrak{S} = \{s \subseteq [N] \text{ s.t } |s \cap G_i| = 1\, \forall j \neq i\} = \mathcal{T}(G_1, \ldots, G_{i-1}, G_{i+1}, \ldots, G_k) \cong \prod_{j \neq i} G_j. \tag{21}$$

We deduce that $\Delta_n(y) = \sum_\mu x_n^{(\mu)} \prod_{j \neq i} \sum_{n' \in G_j} y_{n'}^{(\mu)} x_{n'}^{(\mu)} = \sum_\mu c_i^{(\mu)} x_n^{(\mu)}$, as claimed. □

E.3 Solving the XOR Problem

Let us present the simplest example of a non-trivial problem which can be solved by our proposed PSHN model, but cannot be solved by a classical associative memory model (e.g traditional Hopfield networks [15]) on the same input space: the XOR problem [25]. Note that the problem was also considered in [17] and shown to be solvable by their polynomial networks as soon as the degree of the polynomial is at least 3. This is because higher-order polynomials induce a capacity limit which surpasses the number of neurons N. Indeed, the XOR problem corresponds to $M = 4$ memory patterns in $N = 3$ dimensions, given by

$$x^{(1)} = (-1, -1, -1),\ x^{(2)} = (-1, 1, 1),\ x^{(3)} = (1, -1, 1),\ x^{(4)} = (1, 1, -1), \tag{22}$$

with the identification $0 \mapsto -1$ and $1 \mapsto 1$. The 3rd / output neuron is the XOR of the first two.

Now, consider consider a PSHN with skeleton $\mathfrak{S} = T(\{1\},\{2\},\{3\}) = \{(1,2,3)\}$. It is easy to see that for any $n \in \{1,2,3\}$, then (1) becomes $\partial_n \mathfrak{S} = \{[3] \setminus \{n\}\}$ for any $n \in \{1,2,3\}$. Thus, for any pattern $y \in \{\pm 1\}^3$, the update (4) for the 3rd neuron is given by $T_3(y) = \operatorname{sign}(\Delta_3(y))$, where

$$\begin{aligned}\Delta_3(y) &= \sum_{\mu=1}^{4} x_3^{(\mu)} \sum_{s \in \partial_3 \mathfrak{S}} z_s^{(\mu)} = \sum_{\mu=1}^{4} x_3^{(\mu)} z_{\{1,2\}}^{(\mu)} = \sum_{\mu=1}^{4} x_3^{(\mu)} z_1^{(\mu)} z_2^{(\mu)} \\ &= \sum_{\mu=1}^{M} x_2^{(\mu)} x_1^{(\mu)} y_1 x_2^{(\mu)} y_2 = \sum_{\mu=1}^{4} x_1^{(\mu)} x_2^{(\mu)} x_3^{(\mu)} y_1 y_2 = -4 y_1 y_2.\end{aligned}$$

We have used the fact that $x_1^{(\mu)} x_2^{(\mu)} x_3^{(\mu)} = -1$ for all μ. Thus, $T_3(y) = -\operatorname{sign}(y_1 y_2) = \operatorname{XOR}(y_1, y_2)$. We deduce that this simple PSHN model solves the XOR problem.

E.4 A Boolean Binomial Identity

Lemma E.1. *For every pair of patterns $x, y \in \{\pm 1\}^N$, it holds that*

$$\sum_{\sigma \subseteq [N]} x_\sigma y_\sigma = \prod_{n \in [N]} (1 + x_n y_n) = \begin{cases} 0, & \text{if } x \neq y, \\ 2^N, & \text{otherwise,} \end{cases}, \tag{23}$$

where $x_\sigma := \prod_{n \in \sigma} x_n$ as usual.

Proof. The proof is by induction on N. The case $N = 1$ is trivial since $\sum_{\sigma \subseteq [1]} x_\sigma y_\sigma = 1 + x_1 y_1 = \prod_{n \in [1]} (1 + x_n y_n)$. Suppose the result is true for $N = N'$. We will prove if for $N = N' + 1$. Indeed, observe that

$$\begin{aligned}\sum_{\sigma \subseteq [N'+1]} x_\sigma y_\sigma &= \sum_{\sigma \subseteq [N']} x_\sigma y_\sigma + \sum_{\sigma \subseteq [N']} x_{\sigma \cup \{N'+1\}} y_{\sigma \cup \{N'+1\}} \\ &= \sum_{\sigma \subseteq [N']} x_\sigma y_\sigma + \sum_{\sigma \subseteq [N']} x_\sigma x_{N'+1} y_\sigma y_{N'+1} \\ &= (1 + x_{N'+1} y_{N'+1}) \sum_{\sigma \subseteq [N']} x_\sigma y_\sigma \\ &= (1 + x_{N'+1} y_{N'+1}) \sum_{\sigma \subseteq [N']} \prod_{n \in [N']} (1 + x_n y_n) \text{ by the induction hypothesis} \\ &= \prod_{n \in [N'+1]} (1 + x_n y_n),\end{aligned}$$

which completes the proof. □

F Proof of Theorem D.1: Storage Capacity of A Class of AHNs

F.1 Controlling the Signal Term A_n in (11)

We will prove something more general than Theorem D.1. Let K be a nonempty collection of subsets of $[N]$. Ultimately, we are interested in the case where $K = \partial_s \mathfrak{S}$. Note that K can be seen as an unweighted hyper-graph with vertex-set $[N]$ and edge-set K. Define a random variable $A(K)$ by

$$A(K) := \sum_{s \in K} z_s, \tag{24}$$

where $z_s := \prod_{n \in s} z_n$ as usual. It is clear that the mean of $A(K)$ is given by

$$\mathbb{E}\, A(K) = \sum_{s \in K} (1 - \theta)^{|s|} \tag{25}$$

Let $q = q(K) \geq 1$ be the maximal cardinality of an element of K, i.e

$$q(K) := \max_{s \in K} |s|. \tag{26}$$

Thus, the random variable $A(K)$ is a random multi-linear polynomial of degree q. Moreover, it is clear that

$$\mathbb{E}\, A(K) \geq (1-\theta)^q |K|, \tag{27}$$

with equality if K is regular in the sense that $|s| = q$ for all $s \in K$. Now, for any integer $0 \leq i \leq q$, define $\mu_i = \mu_i(K) \geq 0$ by

$$\mu_i := \max_{s_0 \in \binom{[N]}{k}} \sum_{s \in K | s_0 \subseteq s} \prod_{n \in s \setminus s_0} \mathbb{E}\,|z_n| = \max_{s_0 \in \binom{[N]}{k}} |\{s \in K \mid s_0 \subseteq s\}|, \tag{28}$$

where we have used the fact that $|z_n| = 1$, since z_n only takes the values ± 1, for any $n \in [N]$. It is clear that $\mu_0 = |K|$. The other μ_i's control the size (on average) of the "partial derivatives" of $A(K)$ w.r.t to the elements of K. We have the following proposition which is a direct consequence of the main result in [30].

Proposition F.1. *With all variables defined as above, it holds for any $\lambda > 0$ that*

$$\mathbb{P}\left(|A(K) - \mathbb{E}\, A(K)| \geq \max_{1 \leq i \leq q} \max(\sqrt{\lambda |K| \mu_i C^q}, \lambda^i \mu_i C^q)\right) \leq e^2 e^{-\lambda}, \tag{29}$$

where $C \geq 1$ is an absolute constant.

The appearance of C^q in the result is troublesome and somewhat unavoidable. A very high degree polynomial cannot be concentrated in any meaningful way. Thus, we will focus on the case where the degree q is low in the following sense.

Condition F.1 (Smoothness). *For some $N_1 \geq 1$ (which may depend on N, q) and absolute positive constant C_1, it holds that*

$$\max_{1 \leq i \leq q} N_1^i \mu_i \leq C_1 |K|. \tag{30}$$

Note that the above condition only depends on the topology of the underlying collection K of subsets of $[N]$. For example, it is satisfied in the case where K is a simplicial complex on $K_{N,\leq D}$ with $D = O(1)$ (here $(N_1, q) = (N, D)$), or a transversal of an equi-partition partitioning of $[N]$, with $k = O(1)$ groups (here, $(N_1, q) = (N/k, k)$).

Proposition F.2. *Under Condition F.1 with $N_1 = N^{\Omega(1)}$ and $q = o(\log N)$ as $N \to \infty$, it holds that*

$$\mathbb{P}\left(\left|\frac{A(K)}{\mathbb{E}\, A(K)} - 1\right| \geq t\right) \leq e^2 e^{-t^2 N^{\Omega(1)}}, \textit{ for any } t \in (0,1). \tag{31}$$

Proof. WLOG, take $C_1 = 1$. Observe that

$$\max_{1 \leq i \leq q} \sqrt{\lambda |K| \mu_i C^q} \leq |K| \max_{1 \leq i \leq q} \sqrt{\lambda N_1^{-i} C^q} = |K| \sqrt{C^q \lambda / N_1}. \tag{32}$$

On the other hand, one has

$$\begin{aligned}\max_{1 \leq i \leq q} \lambda^i \mu_i C^q &\leq |K| \max_{1 \leq i \leq q} \lambda^i N_1^{-i} C^q = C^q |K| \cdot \max_{1 \leq i \leq q} (\lambda / N_1)^i \\ &= |K| C^q \begin{cases} \lambda / N_1, & \text{if } \lambda \leq N_1, \\ (\lambda / N_1)^q, & \text{else.} \end{cases}\end{aligned}$$

Thus, for any $t \in (0, C^{q/2})$, taking $\lambda = t^2 N_1 / C^q \leq N_1$ gives

$$\max_{1 \leq i \leq q} \max(\sqrt{\lambda |K| \mu_i C^q}, \lambda^i \mu_i C^q) \leq |K| \max(\sqrt{C^q \lambda / N_1}, C^q \lambda / N_1) = \max(t, t^2) |K|. \tag{33}$$

Combining this with (29) then gives the following concentration inequality

$$\mathbb{P}\left(|A(K) - \mathbb{E}\, A(K)| \geq \max(t, t^2)|K|\right) \leq e^2 e^{-\lambda} = e^2 e^{-t^2 N_1 / C^q} = e^2 e^{-t^2 N^{\Omega(1)}}, \tag{34}$$

because $N_1 = N^{\Omega(1)}$ and $q = o(\log N)$ by hypothesis. In particular, taking $t \in (0,1)$ gives

$$\mathbb{P}\left(|A(K) - \mathbb{E}\, A(K)| \geq t|K|\right) \leq e^2 e^{-t^2 N^{\Omega(1)}}, \tag{35}$$

Noting that $\mathbb{E}\, A(K) = \sum_{s \in K} (1-\theta)^{|s|} \geq |K|(1-\theta)^q = |K| N^{o(1)}$ because $q = o(\log N)$, we get

$$\mathbb{P}\left(|A(K) - \mathbb{E}\, A(K)| \geq t \mathbb{E}\, A(K)|\right) \leq e^2 e^{-t^2 N^{\Omega(1) - o(1)}} = e^2 e^{-t^2 N^{\Omega(1)}},$$

which completes the proof. □

F.2 Proof of Theorem D.1

For any neuron $n \in [N]$, applying Proposition F.2 with $K = \partial_n \mathfrak{S}$, $A(K) = A_n$ (the signal term in (11)), and (N_1, q) as in the statement of Theorem D.1 we obtain that: w.p $1 - O(e^{-t^2 N^{\Omega(1)}})$, it holds that

$$|A_n/\mathbb{E}\, A_n - 1| \le t \text{ with } \mathbb{E}\, A_n = (1-\theta)^q |K| = (1-\theta)^q d(n).$$

Note that the conditions for Proposition F.2 are verified thanks to Lemma F.1. We thus obtain

$$\mathbb{P}(T_n(y) \neq x_n^{(1)}) = \mathbb{P}(T_n(y) x_n^{(1)} \le 0) \le \mathbb{P}(Z_n \ge A_n) \le \mathbb{P}(Z_n \ge (1-\theta)^q d(n)/2) + e^{-N^{\Omega(1)}}. \quad (36)$$

A union-bound in the spirit of the proof of Theorem 4.1 then gives

$$\begin{aligned}\mathbb{P}(T(y) \neq x^{(1)}) &\le \sum_{n=1}^{N} \mathbb{P}(Z_n \ge (1-\theta)^q d(n)/2) + N e^{-N^{\Omega(1)}} \\ &= N \cdot \exp(-\frac{(1-\theta)^{2q} d(n)^2}{2(M-1) d(n)}) + o(1) \\ &= \exp(-\frac{(1-\theta)^{2q} d(n)}{2(M-1)} + \log N) + o(1),\end{aligned} \quad (37)$$

and the claimed lower-bound follows. □

Lemma F.1. *For large N and any positive integer $q \le N$, it holds for any $1 \le i \le q$ that*

$$\mu_i(\binom{[N]}{q}) = \binom{N-i}{q-i} = O(N)^{q-i}, \quad (38)$$

$$\mu_i(\binom{[N]}{\le q}) = \sum_{d \le q} \binom{N-i}{d-i} = O(N)^{q-i}, \quad (39)$$

where the functionals μ_i are as defined in (28)*. Consequently, if $q = o(\log N)$, then $\binom{[N]}{q}$ and $\binom{[N]}{\le q}$ satisfy Condition D.1.*

F.3 Proof of Corollary D.1

The proof follows from combining Theorem D.1 with Lemma F.1. We only need to compute $\underline{d}(\mathfrak{S}) := \max_{n \in [N]} |\partial_n \mathfrak{S}|$ for all the networks considered in the corollary.

Classical Hopfield Networks. If $\mathfrak{S}$ is the collection all singletons of $[N]$, then $q = 1$ and $\underline{d}(\mathfrak{S}) = N - 1$.

Polynomial Hopfield Networks. If $\mathfrak{S}$ is the collection of d-element subsets of $[N]$, then $q = d - 1$ and $\underline{d}(\mathfrak{S}) = \binom{N-1}{d-1}$. Furthermore, if $N \gg d$, then $\binom{N-1}{d-1} \approx N^{d-1}/d!$.

Simplicial Hopfield Networks. The model proposed in [6] corresponds to taking $\mathfrak{S}$ to be a D-skeleton on the set of neurons, i.e the collection of subsets of neurons with cardinality D or less, then $q = D - 1$ and $\underline{d}(\mathfrak{S}) = \sum_{d=0}^{D-1} \binom{N-1}{d} \asymp N^{D-1}$. □

G Proof of Theorem 4.2: Nonrobust Capacity of PSHN Model

The theorem is a direct consequence of the following lemma.

Lemma G.1. *If the subsets $G_1, \ldots, G_k$ with $|G_i| = N_i \ge 1$ for all i, form a partitioning of the set of neurons $[N]$, then for the abstract Hopfield network with skeleton $\mathfrak{S} = \mathcal{T}(G_1, \ldots, G_k)$, it holds that $\underline{d}(\mathfrak{S}) = (\prod_i N_i)/\max_i N_i$.*

Proof. It is clear that $|\mathfrak{S}| = |G_1 \times \ldots \times G_k| = \prod_{i=1}^{d} N_i$. Now, for any $n \in [N]$ let $G_{i(n)}$ be the unique cluster of neurons which contains n. It is clear that $\partial_n \mathfrak{S}$ is isomorphic to $\prod_{i \neq i(n)} G_i$, and so $d(n) := |\partial_n \mathfrak{S}| = \prod_{i \neq i(n)} N_i = |\mathfrak{S}|/N_{i(n)}$, from which it follows that $\underline{d}(\mathfrak{S}) = (\prod_i N_i)/\max_i N_i$ as claimed. □

Proof of Theorem 4.2. For such a partition of N, we must have $k = N/N_1 = \Theta(N)$ and so $\prod_i N_i/\max_i N_i \geq 2^k/O(1) \geq e^{\Theta(N)}$ thanks to Lemma G.1. The result then follows directly from Theorem 4.1. □

H Proof of Theorem 4.3: Robust Storage Capacity of PSHN Model

H.1 Warmup: A Weak Lower-Bound via Chebychev's inequality

Fix $\theta \in [0, 1/2)$. Let $x \in \{\pm 1\}^N$ be uniformly random pattern and let $y \in \{\pm 1\}^N$ be a pattern obtained from x as in Definition 4.1. Let d and N_1 be positive integers and set $N = dN_1$. Partition $[N] := \{1, 2, \ldots, N\}$ d disjoint from $G_1, \ldots, G_d$ of each of size N_1, and let $\mathcal{T} = \mathcal{T}(d, N_1)$ be a *transversal* of the G_i's, i.e the collection of subsets of $[N]$ which contain exactly one element from each G_i. Note that $\mathcal{T}$ is isormophic to $G_1 \times \ldots \times G_d$ in an obvious way, and thus $|\mathcal{T}| = N_1^d$. Finally, let $z = x \odot y \in \{\pm 1\}^N$ be the component-wise product of x and y, and define a random variable

$$A(\mathcal{T}) := \sum_{T \in \mathcal{T}} z_T, \tag{40}$$

where $z_T := \prod_{t \in T} z_t$ as usual. Note that $A(\mathcal{T})$ is a *random multilinear polynomial* of total degree d. The objective is to design N_1 and (thus d too) as a function of N such that $A(\mathcal{T})$ is as large as possible (and positive !) w.p $1 - o(1)$ in the limit $N \to \infty$.

First observe that we can alternately write for every $i \in [d]$,

$$A(\mathcal{T}) = \prod_{1 \leq i \leq d} S_i, \text{ with } S_i := \sum_{t \in G_i} z_t. \tag{41}$$

Now, it is clear that

- The S_i's are iid random variables taking integral values in the range $[-N_1, N_1]$.
- Each S_i is itself a sum of iid random variables which take values ± 1, with $\mathbb{P}(z_t = 1) = 1 - \theta/2$ and $\mathbb{E}\, z_t = 1 - \theta/2 - \theta/2 = a := 1 - \theta \in [0, 1]$. Thus, $\mathbb{E}\, S_i = aN_1$, and

$$\mathbb{E}\, A(\mathcal{T}) = (aN_1)^d. \tag{42}$$

Proposition H.1. *In the limit $N_1 \to \infty$ such that $d = o(N_1)$, it holds w.p $1 - o(1)$ that $A(\mathcal{T}) \asymp \mathbb{E}\, A(\mathcal{T}) = (aN_1)^d$*

Proof. Indeed, setting $a := 1 - \theta$, one computes

$$\begin{aligned} \mathbb{E}\, S_i^2 &= \sum_{t \in G_i} \sum_{t' \in G_i} \mathbb{E}\,[z_t z_{t'}] = N_i + \sum_{t, t' \in G_i,\, t' \neq t} \mathbb{E}\, z_t \mathbb{E}\, z_{t'} \\ &= N_i + N_i(N_i - 1)a^2 = N_i(1 - a^2) + (N_i a)^2 \\ &= N_i(1 - a^2) + (\mathbb{E}\, S_i)^2, \end{aligned} \tag{43}$$

and so $\operatorname{var}(S_i) = N_i(1 - a^2)$. It follows from the independence of the S_i's that

$$\begin{aligned} \operatorname{var}(A(\mathcal{T})) &= \prod_{i=1}^d \mathbb{E}\, S_i^2 - \prod_{i=1}^d (\mathbb{E}\, S_i)^2 = ((aN_1)^2 + N_1(1 - a^2))^d - ((aN_1)^2)^d \\ &= ((aN_1)^2)^d \left(\left(1 + \frac{1/a^2 - 1}{N_1}\right)^d - 1\right) = (\mathbb{E}\, A(\mathcal{T}))^2 \cdot R(\mathcal{T}), \end{aligned} \tag{44}$$

where $R(\mathcal{T}) := \operatorname{var}(A(\mathcal{T}))/(\mathbb{E}\, A(\mathcal{T}))^2 = \left(1 + \dfrac{c}{N_1}\right)^d - 1$, with $c := 1/a^2 - 1 \geq 0$. Now, one computes

$$0 \leq R(\mathcal{T}) = \left(1 + \frac{c}{N_1}\right)^d - 1 \leq e^{cd/N_1} - 1.$$

Thus, if $N_1 \to \infty$ such that $d = o(N_1)$ (i.e $d/N_1 \to 0$), then $R(\mathcal{T}) = o(1)$, and Chebychev's inequality gives

$$\mathbb{P}(|A(\mathcal{T}) - \mathbb{E}\, A(\mathcal{T})| \geq (1/2)\mathbb{E}\, A(\mathcal{T})) \leq 4R(\mathcal{T}) = o(1),$$

and the claim is proved. □

H.2 A Stronger Lower-Bound Via Chernoff

Let us now remove the troublesome requirement "$d = o(N_1)$" from Proposition H.1. First observe that, in the definition of S_i, we can further write $z_t = 2b_t - 1$, where b_t is Bernoulli with parameter $p = p(\theta) := 1-\theta/2 \in (1/2, 1]$. Thus, $S_i = \sum_{t \in G_i}(2b_t - 1) = 2B_i - N_1$, where $B_i := \sum_{t \in G_i} b_t \sim \operatorname{Bin}(N_1, p)$. By well-known concentration results [3], we have

$$\begin{aligned}\mathbb{P}(B_i \geq (1+t)N_1 p) &\leq e^{-\frac{t^2 p N_1}{2+t}}, \text{ for all } t > 0,\\ \mathbb{P}(B_i \leq (1-t)N_1 p) &\leq e^{-\frac{t^2 p N_1}{2}}, \text{ for all } 0 < t < 1.\end{aligned} \tag{45}$$

We deduce that

$$\begin{aligned}\mathbb{P}(S_i \geq (2p(1+t)-1)N_1) &\leq e^{-\frac{t^2 p N_1}{2+t}}, \text{ for all } t > 0,\\ \mathbb{P}(S_i \leq (2p(1-t)-1)N_1) &\leq e^{-\frac{t^2 p N_1}{2}}, \text{ for all } 0 < t < 1.\end{aligned} \tag{46}$$

Therefore: for any $t \in (0, a)$ and $i \in [d]$, it holds w.p $1 - e^{-t^2 p N_1/2}$ that

$$S_i \geq ((2p-1)-t)N_1 = (a-t)N_1,$$

where $a = a(\theta) := 2p - 1 = 1 - \theta \in (1/2, 1]$ as before. A union-bound over $i \in [d]$ then gives: w.p $1 - \delta(N_1) = 1 - de^{-t^2 p N_1/2}$ it holds that

$$A(\mathcal{T}) \geq (aN_1)^d (1-t/a)^d = (aN_1)^d ((1-t/a)^a)^{d/a} \geq (aN_1)^d e^{-td/a} = (ab(t)N_1)^d,$$

where $b(t) := e^{-t/a} \in (0, 1/e)$. Further taking $t = 1/2$ gives: w.p $1 - de^{-pN_1/8}$,

$$A(\mathcal{T}) \geq (abN_1)^d, \tag{47}$$

where $b = e^{-1/(2a)}$. Now, we want d to be as large as possible, and the RHS of the above to be as large as possible too. We can achieve by ensuring that $\delta(N_1) := e^{-pN_1/8 + \log d} \to 0$ in the limit $N_1 \to \infty$. To satisfy this constraint (perhaps non-optimally!) it suffices to take

$$N_1 \geq (9/p)\log d, \tag{48}$$

so that $\delta(N_1) \leq e^{-pN_1/72}$. We have proved the following.

Proposition H.2. *If $N_1 \geq (9/p)\log d$, then it holds w.p $1 - e^{-pN_1/72}$ that $A(\mathcal{T}) \geq (abN_1)^d$, where $a := 1-\theta$, $p := 1 - \theta/2$, and $b := e^{-1/(2a)}$.*

The following result is the last technical step towards the prove of Theorem 4.1.

Proposition H.3. *Fix a corruption level $\theta \in [0, 1/2)$ and let $N_1 \geq C \log N$, where $C \geq 73/p$. Then, for large N, it holds w.p $1 - o(1/N)$ that*

$$A(\mathcal{T}) \geq (abN_1)^d, \tag{49}$$

where $a := 1-\theta$, $p := 1-\theta/2$, and $b := e^{-1/(2a)}$.

Proof. Indeed, observe that $Ne^{-pN_1/72} = e^{-pN_1/72 + \log N} = e^{-(N_1 - (72/p)\log N)p/72} = o(1)$ if $N_1 \geq (73/p)\log N$. The result then follows from Proposition H.2 since $\log N \geq \log d$. □

Note that the constant 73 appearing in Proposition H.3 (and also in Theorem 4.3) has not been optimized an could potentially be made much smaller with a bit of more work.

H.3 Proof of Theorem 4.3

We are now ready to prove Theorem 4.3 proper. Fix a corruption level $\theta \in [0, 1)$, and let $y = y(\theta) \in \{\pm 1\}^N$ be a corrupt version of a memory $x^{(1)}$ which is formed by chosen a subset s_θ of size θN of neurons, uniformly at random and independently of the memories $x^{(1)}, \ldots, x^{(\mu)}$, and then setting

$$y_n = \begin{cases} x_n^{(1)}, & \text{if } n \in s_\theta,\\ -1, & \text{else.}\end{cases} \tag{50}$$

For any neuron $n \in [N]$, the signal term in (11) is given by

$$A_n := \sum_{s \in \partial_n \mathfrak{S}} x_s^{(1)} y_s. \tag{51}$$

Observe that $\partial_n \mathfrak{S}$ is precisely the collection of subsets of $[N]$ which contain exactly one element of each group of neurons G_i except the group which contains the neuron n. Applying Proposition H.3 with $\mathcal{T} = \partial_n \mathfrak{S}$, $d = k - 1$, and $A_n = A(\mathcal{T})$ one has $A_n \geq (abN_1)^{k-1} = (ab)^{k-1} d(n)$ w.p $1 - o(1/N)$ as soon as $N_1 \geq C \log N$, where $d(n) = N_1^{k-1}$, $a := 1 - \theta$ and $b := e^{-1/(2a)}$.

Reasoning analogously to (17), we see that

$$\begin{aligned}
\mathbb{P}(T(y) \neq x^{(1)}) &\leq \sum_{n=1}^{N} \mathbb{P}\left(Z_n \geq A_n\right) \\
&\leq \sum_{n=1}^{N} \left(\mathbb{P}\left(Z_n \geq (ab)^{k-1} d(n)\right) + o(1/N)\right) \\
&\leq \sum_{n=1}^{N} \exp\left(\frac{(ab)^{2(k-1)} d(n)^2}{2(M-1)d(n)}\right) + N \cdot o(1/N) \\
&= \sum_{n=1}^{N} \exp\left(\frac{(ab)^{2(k-1)} d(n)}{2(M-1)}\right) + o(1) \\
&\leq N \cdot \exp\left(-\frac{(a^2 b^2 N_1)^{k-1}}{2(M-1)}\right) + o(1) \\
&= \exp\left(-\frac{(a^2 b^2 N_1)^{k-1}}{2(M-1)} + \log N\right) + o(1),
\end{aligned} \tag{52}$$

To make the RHS go to zero in the limit $N \to \infty$, it suffices that $(a^2 b^2 N_1)^{k-1}/M \geq (2 + \gamma) \log N$, or equivalently,

$$M \leq \frac{(a^2 b^2 N_1)^{k-1}}{(2 + \gamma) \log N}$$

where γ is a positive constant.

In particular, taking $N_1 = C \log N$ and $k = N/N_1$, and then lower-bounding the logarithm of the function $f(N) := (a^2 b^2 C \log N)^{N/(C \log N) - 1}$ by $\Omega(N \log \log(N) / \log N)$ gives the result. □