DENSE HOPFIELD NETWORKS WITH HIERARCHICAL MEMORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider a 3-level hierarchical generative model for memories which are sampled and stored in a dense Hopfield network with polynomial activation. We analytically derive conditions for each level of this hierarchy to be locally stable – that is they are local energy maxima. We find that it takes only a polynomial amount of information to generalize beyond particular memories and even particular groups in the hierarchy. Our theory predicts the qualitative features a phase diagram in the number of memories, sharpness of the activation function (polynomial degree) for data from Fashion-MNIST.

1 INTRODUCTION

Understanding the structure of Hopfield networks could help us understand when they merely reproduce memorized data, and when they can generalize beyond what they have already seen. This question is closely related to the notion of capacity in generalized Hopfield models but is more subtle as it is insufficient to say that generalization happens precisely when we exceed a capacity threshold. In this work we show that intuition is true, under certain assumptions on the data.

Additionally Ambrogioni (2023) shows that diffusion models at zero temperature have the same energy landscape as a modern Hopfield network. This relationship implies that our studies here may contextualize the memorization/generalization behavior of models deployed at scale. While we study this generalization behavior in modern Hopfield networks, under a particular hierarchical model of data, we expect that our qualitative results should transfer with the appropriate modification in more general settings.

Data with latent hierarchical structure is very common. Because modern Hopfield networks have an energy function which depends only on the distance to all the memories (on the sphere) we may expect that generically only the clustering structure of can be memorized by Hopfield networks. While general diffusion models can exhibit much more complicated behavior, clustering is a universal property of data.

This motivates our attempt at understanding the following questions:

1. With hierarchically correlated memories, do dense Hopfield models memorize and remember patterns?
2. Can these Hopfield models recover generalized patterns from the underlying correlation structure?

While there is a role for understanding how the structure of a diffusion model impacts its memorization or generalization behavior here we focus primarily on how hierarchical features in the data can be learned, and how that is precisely related to memorization/forgetting in an exactly solvable model.

We note that Hopfield networks with correlated patterns have been studied in previous work Dotsenko (1986); Engel (1990); Agliari et al. (2013). However, these previous works all considered a quadratic activation function and we find that precisely because of the stronger activation function, interesting phenomena may occur.

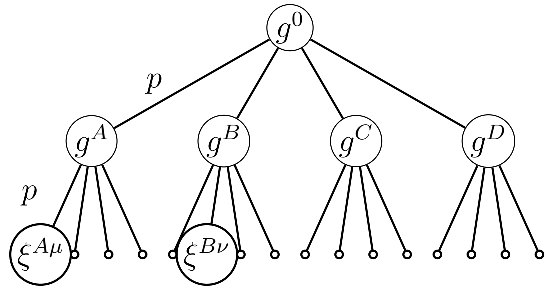

Figure 1: Schematic of the hierarchical memory structure we consider in this work. Here, $A, B, \dots \in \{1, \dots, G\}$ and $\mu, \nu, \dots \in \{1, \dots, K\}$. Only the ξ are encoded into the network.

2 Model Description

We consider a system of binary neurons, each of which is denoted by a variable σ_i which can take on values ± 1 (Hopfield, 1982). The state of the entire system is denoted $\boldsymbol{\sigma} \in \{0, 1\}^N$. A pattern to be stored, or memory, is denoted as $\boldsymbol{\xi}$, where the i^{th} index ξ_i is the state of the i^{th} neuron in the memory. We define the following as the energy function for the system:

$$E(\boldsymbol{\sigma}) = -\sum_{\mu=1}^{M} F(\boldsymbol{\xi}^{\mu} \cdot \boldsymbol{\sigma}), \tag{1}$$

where $F(x)$ is an activation function which here takes in as input the dot product between the memory and the current state of the system. Recovery of memories happens by performing local hill-climbing starting at a probe point $\boldsymbol{\sigma}^0$ until a fixed point (local maximum) is reached.

The original work by Hopfield (1982) utilized a quadratic activation function and Amit et al. (1985) found that the memories are reliably minima of the system as long as $M \leq \alpha N$, with $\alpha \approx 0.14$. For a higher density of memories, the "basins of attraction" surround the memory states begin interfering with one another and the memories cannot reliably be recovered. More recently, *dense Hopfield networks* Krotov & Hopfield (2016) were introduced as a generalization of the original idea in which the quadratic activation function is replaced with a higher order polynomial or an exponential function. Such functions induce a much greater energy penalty for a system being evolving away from a memory state and this effectively sharpens the energy wells of the memories and "pulls apart" closely correlated memories. Here, we will consider polynomial activation functions, i.e. $F(x) = x^n$.

In this work, we consider the ability of these dense Hopfield networks to recover *hierarchically correlated memories*. In particular, we imagine the memories correlated in a tree structure, such that groups of memories are derived from prototypes, and the prototypes are further derived from a singular root, see Fig. 1. We call the central root prototype g^0 the level 1 root prototype, and the prototypes underneath it are the level 2 prototypes g^A. Importantly, we initialize the network with only leaf memories $\xi^{A\mu}$.

We minimally model this system by initializing the level 1 root prototype g^0 as a random binary vector. We then generate G level 2 prototypes via the following:

$$g_i^A = \begin{cases} -g_i^0, & \text{with probability } p \\ g_i^0, & \text{with probability } 1-p \end{cases} . \tag{2}$$

In principle, the parameter p will be different for each of the $g^{A\mu}$ but for simplicity, we maintain a uniform correlation between all $g^{A\mu}$ and g^0. From these prototypes, the memories are generated via

$$\xi_i^{A,\mu} = \begin{cases} -g_i^A, & \text{with probability } p \\ g_i^A, & \text{with probability } 1-p \end{cases} . \tag{3}$$

for $\mu = 1, \ldots K$. With this structure of memories, each memory ξ is conditionally independent with every other memory within the same group. We again note that the prototypes are *not* encoded into the network as memories.

Memory Stability

We investigate the stability of memory retrieval in the dense Hopfield networks by initializing the state in a memory state, perturbing the system in an arbitrary direction, and determining whether the magnitude of fluctuations about the memory state is greater than the energy gap. Specifically, calculate the mean and variance of the following energy gap:

$$\Delta E = E(\boldsymbol{\sigma}) - E(\boldsymbol{\sigma} - 2\sigma_i \hat{\mathrm{e}}_i) \tag{4}$$

This measures the gap between the energy at a point $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}$ with the i^{th} coordinate flipped, and the variance allows us to upper-bound the probability that $\Delta E < 0$. This analysis yields the familiar linear memory capacity of the Hopfield model, and superlinear memory capacities of dense Hopfield networks.

3 Memory Retrieval

We first investigate the ability of the network to retrieve each of the individual encoded memories. We find in this case that the ratio of the variance to the squared mean of the energy gap (eq. 4) is

$$\left(\frac{\mathrm{var}\Delta E}{\mathbb{E}[\Delta E]^2}\right)_{\xi} \approx K^2 \cdot \frac{G^2 q^{4n}(q^{-2} - 1) + G q^{4n-3}}{(1 + K q^n + G K q^{2n})^2}. \tag{5}$$

Here, $q = (1 - 2p)^2$, and n is the power of the activation function (i.e. $F(x) = x^n$). The derivation of this and the following equations are presented in the Appendix B. From this equation, we observe that for any fixed value of n, as the number of memories within a group K is increased, the fluctuations become more and more prevalent. However, even a modestly large n can overwhelm this effect.

We may also calculate the statistics of the energy gap from a *prototype state*. For a level 2 prototype, we find

$$\left(\frac{\mathrm{var}\Delta E}{\mathbb{E}[\Delta E]^2}\right)_{g^A} \approx \frac{1}{K q^n}\left(\frac{1 + (G q^n) q^{n-2}}{(1 + G q^n)^2}\right) + \left(\frac{(G q^n) q^{n-2} + (G q^n)^2 \left(q^{-1} - 1\right)}{(G q^n + 1)^2}\right). \tag{6}$$

This interesting relation indicates that for very small n and finite G the second term will always be $O(1)$ and hence prototype states will not be stable when N is large. Once n becomes large enough so that $G q^n \equiv \epsilon \ll 1$ is small (remember $q \in [0, 1]$) then the second term becomes small and the network might remember the prototype states.

The first term of eq. (6) becomes approximately $(K q^n)^{-1} = G/(K\epsilon)$ for small ϵ. For stability this term also needs to be small, so we require that $G/(K\epsilon) = \epsilon$ or $K = G\epsilon^{-2}$ (ϵ has only a weak dependence on N). The level 2 prototypes become stable minima at $K = O(G\epsilon^{-2})$ *despite not being encoded into the network as memories*. Indeed we only need K *polynomially* large before we begin to see this kind of generalization so long as n is tuned carefully.

Finally, calculate the same quantity for the level 1 root probability:

$$\left(\frac{\mathrm{var}\Delta E}{\mathbb{E}[\Delta E]^2}\right)_{g^0} = \frac{1}{G}\left(\frac{1}{q} + \frac{1}{K q^2}\right). \tag{7}$$

In addition to the stability of the level 2 prototypes, the level 1 root prototype also remains stable, with stability growing with G and K. We interpret this as the memories within each group ”coalesce” into the prototype for each group, so that the entire system behaves akin to a single group, with the level 2 prototypes now forming the memories based upon the level 1 root prototype. Interestingly, this quantity shows no dependence on n, and only requires that the groups have at least some minimal correlation $q \neq 0$.

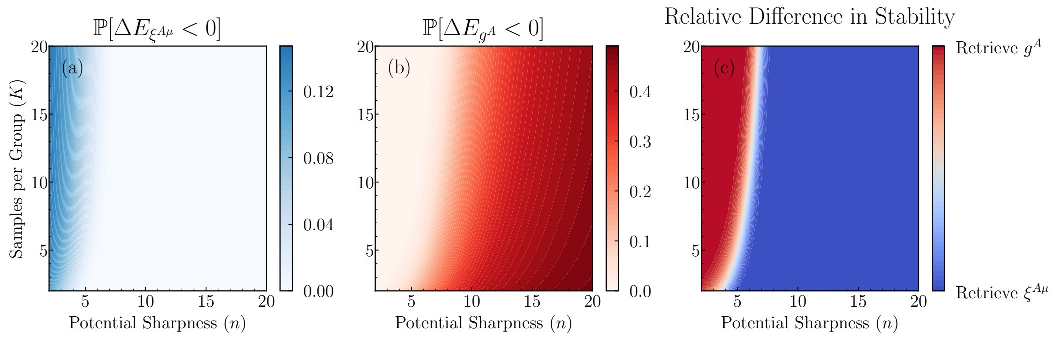

Figure 2: Failure probability for $\xi^{A\mu}$ (a) and g^A (b) calculated from eq. 8, with $q = 0.7$ and $G = 10$. Relative difference between these two (Eq. 9) is shown in (c).

With this ratios, we may calculate failure probabilities to remember each of these levels of memories, assuming that ΔE is Gaussian. That is, for $\xi^{A\mu},\ g^A,\ g^0$, we calculate

$$\mathbb{P}[\Delta E < 0] = \frac{1}{2}\text{erfc}\left(\frac{\mathbb{E}[\Delta E]}{\sqrt{2\text{var}\Delta E}}\right). \tag{8}$$

Here, $\text{erfc}(x)$ is the complementary error function. If this value is large, then the state is likely to not be stable minima in the energy landscape.

We plot this failure probability as a function of n and K in fig. 2. In (a) and (b), we observe the relationships on the independent variables discussed above. That is, the failure probability in remembering $\xi^{A\mu}$ increases with K but decreases rapidly with n. In (b), we observe that the network is better able to remember the intermediate level 2 prototypes at small n and larger K, but at larger n, the energy minima corresponding to the individual memories become well separated and the network ceases to be able to recall the level 2 prototypes. Note that at the chosesn model parameters, the failure probability of remembering the root protoype was approximately 0.

In order to further capture the behavior of the network, in fig. fig. 2(c), we plot the relative difference between the failure probability of remembering the level 3 memories and the level 2 prototypes:

$$\frac{\mathbb{P}[\Delta E(\xi^{A\mu}) < 0] - \mathbb{P}[\Delta E(g^A) < 0]}{\mathbb{P}[\Delta E(\xi^{A\mu}) < 0] + \mathbb{P}[\Delta E(g^A) < 0]}. \tag{9}$$

conditioned on one of them being unstable. We additionally assume that the events are disjoint (which is in approximate accord with fig. 2(a,b)) to simplify the denominator into a sum of probabilities. When this quantity is close to 1, the probability of *failing to remember* $\xi^{A\mu}$ is much larger than that of failing to remember g^A. As a result, the system is likely to evolve towards g^A. Conversely, if this quantity is close to -1, then the probability of failing to remember g^A is much larger than that of failing to remember $\xi^{A\mu}$. Thus, in this regime, the system is likely to remain in a level 3 memory state.

4 Experiments

The model of data we rely on in this manuscript aims to model hierarchy, while assuming that every level in the hierarchy is related to the one above it via isotropic, and independent link variables. Real data will violate these assumptions so to ensure that our modeling assumptions are not fine-tuned with respect to real data we consider a Hopfield model on various subsets of Fashion-MNIST (Xiao et al., 2017). This dataset is composed of 10 classes. These classes also have non-trivial overlaps, so we might initially model the latent structure of this data as a tree of the type shown fig. 1, composed of sixty-thousand leaf nodes, ten nodes above those with g^A corresponding to prototypes of the ten classes, and some small amount of structure between these and the root node. For further details about the experimental setup see appendix A.

We chose this dataset in part because all the images are centered, with the same rotation, so our results will not be confounded by those symmetries typically present in images. In more complex

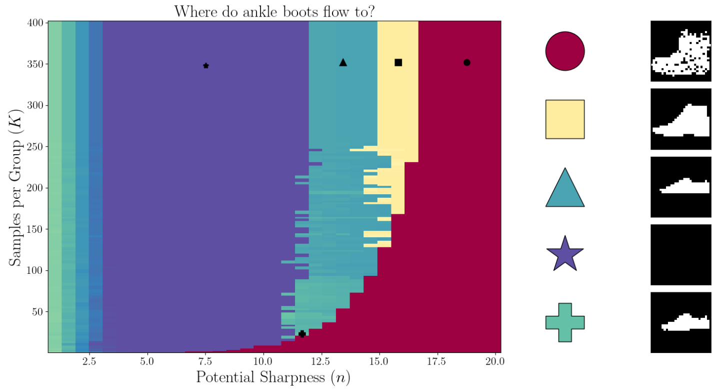

Figure 3: A phase diagram depicting the final location of the flow initialized at an ankle boot image. The colors depict Hamming distance from the original image where red (the circle in the legend) is exactly zero. We sample the final image at five points and show them on the right with corresponding shape and color legend. We see four well-separated phases, and a regime on the left corresponding to the root (purple star) fixed point slowly moving as n decreases. At the lower triple point (teal plus) we can see the shoe prototype taking on some features of the bootprototype (e.g. a lift at the front of the shoe) but remains largely consistent with the blue triangle shoe prototype.

settings these symmetries may result in further structure (see for example the analysis by Kamb & Ganguli (2024) which takes translation symmetry into account) which we do not aim to describe here.

Additionally these experiments will show to what extent our nearest-neighbor calculations agree with global properties of the energy landscape. We expect to see transitions as soon as one direction becomes unstable, but our theory only suggests that the memory flows to the basin of attraction formed by its parent. We will be able to test this hypothesis as well.

In fig. 3 we see that the leaf prototype (ankle boot) is stable for large n and small K as expected by our calculation, with a stability frontier which looks qualitatively similar to that shown in fig. 2. Additionally we see that for small n, we do not need very large K to remember higher-order prototypes (yellow square, blue triangle, green cross from fig. 3), and that for K too large we simply remember the root prototype.

We see an additionally interesting phenomenon that several different prototypes are remembered based on the value of n, with more complicated prototypes requiring larger K to be resolved, and are resolved at larger n. The stability criterion for a small second term in eq. (6) implies that $n = \log_{1/q}(G_{\text{eff}})$, and G_{eff} ought to be larger for more fine-grained prototypes as they correspond to a larger effective number of groups. Similarly $K = G_{\text{eff}}\epsilon^{-2}$ has to be larger.

5 Conclusions

In this work, we have examined dense Hopfield networks in the presence of hierarchically correlated memories. We find that as a function of the number of correlated patterns and the activation function, there are interesting regimes of generalization, where the network remembers states corresponding to patterns which are higher up in the correlation structure of the memories, despite not being encoded into the network outright. We interpret this as a form of generalization and notice that we only require polynomial data to generalize for an appropriate potential F.

This work may be extended in numerous directions. First, the case of exponential activation functions is being currently explored by the authors. From a different perspective, the statistical physics of these models would be interesting to explore given other recent work Lucibello & Mézard (2024). The connection to diffusion models as well as the attention mechanism in transformers would be worth exploring as well.

REFERENCES

Elena Agliari, Adriano Barra, Andrea De Antoni, and Andrea Galluzzi. Parallel retrieval of correlated patterns: from hopfield networks to boltzmann machines. *Neural Netw*, 38:52–63, Feb 2013. ISSN 1879-2782 (Electronic); 0893-6080 (Linking). doi: 10.1016/j.neunet.2012.11.010.

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *arXiv preprint arXiv:2309.17290*, 2023.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

Viktor S. Dotsenko. Hierarchical model of memory. *Physica A: Statistical Mechanics and its Applications*, 140(1):410–415, 1986. ISSN 0378-4371. doi: https://doi.org/10.1016/0378-4371(86)90248-7. URL `https://www.sciencedirect.com/science/article/pii/0378437186902487`.

A Engel. Storage of hierarchically correlated patterns. 23(12):2587, 1990. doi: 10.1088/0305-4470/23/12/034. URL `https://dx.doi.org/10.1088/0305-4470/23/12/034`.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL `https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554`.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Phys. Rev. Lett.*, 132:077301, Feb 2024. doi: 10.1103/PhysRevLett.132.077301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.132.077301`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

A Data and Model Preparation

Fashion-MNIST is an image dataset with pixels taking on values in $[0, 255]$. To match the setting of our calculations we first rescale the range to $[-1, 1]$ and then dither the images, choosing either ± 1 with probabilities so that the average pixel value matches the original pixel value.

We then consider two tuneable parameters, the number of elements per group K, and the power for the potential n where $F(x) = \mathrm{sign}(x)|x|^n$. We don't consider G as a tuneable parameter because of the limited dynamic range (1-10).

B Stability Criterion Derivations

In this appendix we derive stability criterion for the retrieval of prototypes within the hierarchical memory structure. These stability criterion are derived for the dense associative memories with polynomial activation.

B.1 Level 3 Memory Retrieval Stability Criterion

To derive the stability criterion for retrieval of a level 3 memory, we begin again with the gap to an excitation from a memory state, $\xi^{B\nu}$:

$$\Delta E = 2 \sum_{A} \sum_{\mu \in A} \sum_{k \text{ odd}} \binom{n}{k} (\xi_i^{A\mu} \xi_i^{B\nu}) \left(\sum_{j \neq i}^{N} \xi_j^{A\mu} \xi_j^{B\nu} \right)^{n-k} . \tag{10}$$

The expectation value to be evaluated is

$$\mathbb{E}[\Delta E] = 2 \sum_{A} \sum_{\mu \in A} \sum_{k \text{ odd}} \binom{n}{k} \mathbb{E}\left[(\xi_i^{A\mu} \xi_i^{B\nu}) \right] \mathbb{E}\left[\left(\sum_{j \neq i}^{N} \xi_j^{A\mu} \xi_j^{B\nu} \right)^{n-k} \right] . \tag{11}$$

Terms contributing to this expectation value are

1. $A = B,\ \mu = \nu$. $M = 1$.

$$\begin{aligned} T_1 &= 2 \sum_{k \text{ odd}} \binom{n}{k} \mathbb{E}\left[\left(\sum_{j \neq i}^{N} \xi_j^{B\nu} \xi_j^{B\nu} \right)^{n-k} \right] \\ &\approx 2n(N-1)^{n-1} \end{aligned}$$

2. $A = B,\ \mu \neq \nu$. $M = K - 1$.

$$\begin{aligned} T_2 &= 2 \sum_{k \text{ odd}} \binom{n}{k} \mathbb{E}\left[(\xi_i^{B\mu} \xi_i^{B\nu}) \right] \mathbb{E}\left[\left(\sum_{j \neq i}^{N} \xi_j^{B\mu} \xi_j^{B\nu} \right)^{n-k} \right] \\ &\approx 2n(N-1)^{n-1} q^n. \end{aligned}$$

3. $A \neq B$. $M = (G-1)K$.

$$\begin{aligned} T_3 &= 2 \sum_{k \text{ odd}} \binom{n}{k} \mathbb{E}\left[(\xi_i^{A\mu} \xi_i^{B\nu}) \right] \mathbb{E}\left[\left(\sum_{j \neq i}^{N} \xi_j^{A\mu} \xi_j^{B\nu} \right)^{n-k} \right] \\ &\approx 2n(N-1)^{n-1} q^{2n}. \end{aligned}$$

The full mean is then

$$\mathbb{E}[\Delta E] \approx 2n(N-1)^{n-1} \times \left(1 + (K-1)q^n + G(K-1)q^{2n}\right). \tag{12}$$

Next, we calculate the second moment.

$$\mathbb{E}[(\Delta E)^2] = 4 \sum_{A,A'} \sum_{\mu\in A,\mu'\in A'} \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{A'\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{A'\mu'}\xi_j^{B\nu}\right)^{n-k'}\right]. \tag{13}$$

Terms contributing to this are the following:

1. $A = A' = B,\ \mu = \mu' = \nu$. $M = 1$.

$$T_1 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{B\nu}\xi_i^{B\nu}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{B\nu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{B\nu}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}.$$

2. $A = A' = B,\ \mu = \mu' \neq \nu$. $M = K-1$.

$$T_2 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{B\mu}\xi_i^{B\mu}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{B\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{B\mu}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{2n-2}$$

3. $A = A' = B,\ \mu \neq \mu' = \nu$ (or $\mu' \neq \mu = \nu$). $M = 2(K-1)$.

$$T_3 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{B\mu}\xi_i^{B\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{B\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{B\nu}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{n}.$$

4. $A = A' = B$, with μ, μ', and ν distinct. $M = (K-1)(K-2)$.

$$T_4 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{B\mu}\xi_i^{B\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{B\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{B\mu'}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{2n-1}.$$

5. $A = A' \neq B, \mu = \mu'$. $M = (G-1)K$.

$$T_5 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{A\mu}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{4n-4}.$$

6. $A = A' \neq B,\ \mu \neq \mu'$. $M = (G-1)K(K-1)$.

$$T_6 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{A\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{A\mu'}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{4n-3}.$$

7. $A \neq A' = B,\ \mu' = \nu$ (or $A' \neq A = B$). $M = 2(G-1)K$.

$$T_7 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{B\nu}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{B\nu}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{2n}.$$

8. $A \neq A' = B,\ \mu' \neq \nu$ (or $A' \neq A = B$). $M = 2(K-1)(G-1)K$.

$$T_8 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{B\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{B\mu'}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{3n-1}$$

9. A, A', and B distinct. $M = (G-1)(G-2)K^2$

$$T_9 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{A'\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_j^{A\mu}\xi_j^{B\nu}\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_j^{A'\mu'}\xi_j^{B\nu}\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{4n-2}$$

Putting these all together, the full second moment, taking $N, K, G \gg 1$, is

$$\mathbb{E}[(\Delta E)^2] = 4n^2N^{2n-2}\left[1 + Kq^{2n-2} + 2Kq^n + K^2q^{2n-1} + KGq^{4n-4} + GK^2q^{4n-3} + 2KGq^{2n} + 2K^2Gq^{3n-1} + K^2G^2q^{4n-2}\right] \tag{14}$$

From this, we obtain the ratio of the variance to the squared mean:

$$\frac{\text{var}\Delta E}{\mathbb{E}[\Delta E]^2} = \frac{G^2K^2q^{4n}(q^{-2}-1) + 2GK^2q^{3n}(q^{-1}-1) + GK^2q^{4n-3} + GKq^{4n-4} + K^2q^{2n}(q^{-1}-1) + Kq^{2n-2}}{(1 + Kq^n + GKq^{2n})^2} \tag{15}$$

To simplify this arduous expression, we neglect terms in the numerator which are lower than quadratic in K, as since the denominator is quadratic in K, these terms will vanish in the large K limit. Furthermore, terms which contain powers of q smaller than $4n$ vanish faster than the others, so we neglect these as well (which results in only small qualitative or visible changes to the figures and calculated metrics). This yields

$$\approx K^2 \cdot \frac{G^2q^{4n}(q^{-2}-1) + Gq^{4n-3}}{(1 + Kq^n + GKq^{2n})^2} \tag{16}$$

This is eq. 5 in the main text.

B.2 Prototype Retrieval

Next, we focus on the ability of the dense Hopfield network to recover the higher level prototypes within the tree, that is, the level 2 memories as well as the root level 1 memory. We begin with the expression for the change in energy upon perturbing a prototype state. We denote the prototype states as follows: g^0 will refer to the root level 1 prototype. $g^A,\ A \in \{1, \ldots, G\}$ will refer to one of the G level 2 prototypes. The energy gap to perturbing a level 2 prototype is

$$\Delta E = \sum_A \sum_{\mu\in A} F\left[\xi_i^{A\mu}g_i^B + \sum_{j\neq i}^{N} \xi_j^{A\mu}g_j^B\right] - F\left[-\xi_i^{A\mu}g_i^B + \sum_{j\neq i}^{N} \xi_j^{A\mu}g_j^B\right] \tag{17}$$

$$= 2\sum_A \sum_{\mu\in A} \sum_{k \text{ odd}} \binom{n}{k} (\xi_i^{A\mu}g_i^B)^k \left(\sum_{j\neq i}^{N} \xi_j^{A\mu}g_j^B\right)^{n-k}. \tag{18}$$

We now evaluate the expectation value of the above. In order to evaluate terms such as $\mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_j^{A\mu} g_j^B\right)^{n-k}\right]$, we will make use of the fact that at large N, inner products involving a random vector concentrate around the their mean. We may therefore neglect fluctuations in such quantities and make such approximations as $\mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_j^{A\mu} g_j^B\right)^{n-k}\right] \sim \mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_j^{A\mu} g_j^B\right)\right]^{n-k}$.

This expectation may be separated into a sum over cases. We enumerate the cases and their multiplicity here. In each of these, will we approximate the combinatorial sum with its largest term (leading order in N). Finally, we define note that $q \equiv (1-2p)^2$.

1. $A = B$, with multiplicity $M = K$.

$$\begin{aligned} T_1 &= 2 \sum_{k \text{ odd}} \binom{n}{k} \mathbb{E}\left[\xi_i^{B\mu} g_i^B\right] \mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_j^{B\mu} g_j^B\right)^{n-k}\right] \\ &\approx 2n(N-1)^{n-1} q^{n/2} \end{aligned}$$

2. $A \neq B$, with multiplicity $M = (G-1)K$.

$$\begin{aligned} T_2 &= 2 \sum_{k \text{ odd}} \binom{n}{k} \mathbb{E}\left[\xi_i^{A\mu} g_i^B\right] \mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_j^{A\mu} g_j^B\right)^{n-k}\right] \\ &\approx 2n(N-1)^{n-1} q^{3n/2}. \end{aligned}$$

With these expressions, we obtain for the expectation value of the energy gap

$$\mathbb{E}[\Delta E] \approx 2n(N-1)^{n-1} q^{n/2} K \left(1 + (G-1)q^n\right). \tag{19}$$

We will require the squared mean:

$$\mathbb{E}[\Delta E]^2 \approx 4n^2(N-1)^{2n-2} q^n K^2 \left(1 + 2(G-1)q^n + (G-1)^2 q^{2n}\right) \tag{20}$$

Now we require the second moment.

$$(\Delta E)^2 = 4 \sum_{A,A} \sum_{\mu\in A, \mu'\in A'} \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu} \xi_i^{A'\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_i^{A\mu} g_i^B\right)^{n-k} \left(\sum_{j\neq i}^N \xi_i^{A'\mu'} g_i^B\right)^{n-k'}\right] \tag{21}$$

We enumerate the cases in order to take the expectation value:

1. $A = A' = B,\ \mu = \mu'$. $M = K$.

$$\begin{aligned} T_1 &= 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_i^{B\mu} g_i^B\right)^{2n-k-k'}\right] \\ &\approx 4n^2(N-1)^{2n-2} q^{n-1} \end{aligned}$$

2. $A = A' = B,\ \mu \neq \mu'$. $M = K(K-1)$.

$$\begin{aligned} T_2 &= 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{B\mu} \xi_i^{B\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^N \xi_i^{B\mu} g_i^B\right)^{n-k} \left(\sum_{j\neq i}^N \xi_i^{B\mu'} g_i^B\right)^{n-k'}\right] \\ &\approx 4n^2(N-1)^{2n-2} q^n. \end{aligned}$$

3. $A \neq A' = B$ (or symmetrically $A' \neq A = B$). $M = 2K^2(G-1)$.

$$T_3 = 4 \sum_{k,k' \text{ odd}}^{n} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{B\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_i^{A\mu} g_i^B\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_i^{B\mu'} g_i^B\right)^{n-k}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{2n}.$$

4. $A = A' \neq B,\ \mu = \mu'$. $M = (G-1)K$.

$$T_4 = 4 \sum_{k,k'\text{odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_i^{A\mu} g_i^B\right)^{2n-k-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{3n-3}$$

5. $A = A' \neq B, \mu \neq \mu'$. $M = (G-1)K(K-1)$.

$$T_5 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{A\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_i^{A\mu} g_i^B\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_i^{A\mu'} g_i^B\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{3n-2}.$$

6. A, A', and B all distinct. $M = (G-1)(G-2)K^2$.

$$T_6 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu}\xi_i^{A'\mu'}\right] \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_i^{A\mu} g_i^B\right)^{n-k} \left(\sum_{j\neq i}^{N} \xi_i^{A'\mu'} g_i^B\right)^{n-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2}q^{3n-1}.$$

With these sub-expressions and approximating $N, G, K \gg 1$, the full second moment is

$$\mathbb{E}[(\Delta E)^2] = 4n^2N^{2n-2}\left[Kq^{n-1} + K^2q^n + 2K^2Gq^{2n} + GKq^{3n-3} + K^2Gq^{3n-2} + K^2G^2q^{3n-1}\right]. \tag{22}$$

From this and the mean we derived above, we obtain the ratio of the variance to the second moment:

$$\frac{\text{var}\Delta E}{\mathbb{E}[\Delta E]^2} = \frac{K^{-1}q^{n-1} + GK^{-1}q^{3n-3} + Gq^{3n-2} + G^2q^{3n}\left(q^{-1}-1\right)}{q^n + 2Gq^{2n} + G^2q^{3n}} \tag{23}$$

$$= \frac{1}{Kq^n}\left(\frac{1+Gq^{2n-2}}{1+2Gq^n+G^2q^{2n}}\right) + q^{2n}\left(\frac{Gq^{-2} + G^2\left(q^{-1}-1\right)}{1+2Gq^n+G^2q^{2n}}\right) \tag{24}$$

$$= \frac{1}{Kq^n}\left(\frac{1+(Gq^n)q^{n-2}}{(1+Gq^n)^2}\right) + \left(\frac{(Gq^n)q^{n-2} + (Gq^n)^2\left(q^{-1}-1\right)}{(Gq^n+1)^2}\right). \tag{25}$$

This is eq. 6 in the main text.

Now we derive the case of the level 1 root memory g^0. We take the expectation value of Eq. 17 with $g^B \to g^0$. The expectation value of the energy gap above the root prototype is

$$\mathbb{E}[\Delta E] \approx 2n(N-1)^{n-1} \cdot GKq^n. \tag{26}$$

For the second moment, there are only three types of terms which contribute to the summation. Using eq. 21, with $g^B \to g^0$, we obtain

1. $A = A',\ \mu = \mu'$. $M = GK$.

$$T_1 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k}\binom{n}{k'} \mathbb{E}\left[\left(\sum_{j\neq i}^{N} \xi_i^{A\mu} g_i^0\right)^{2n-k-k'}\right]$$
$$\approx 4n^2(N-1)^{2n-2} \cdot q^{2n-2}.$$

2. $A = A',\ \mu \neq \mu'$. $M = GK(K-1)$.

$$T_2 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k} \binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu} \xi_i^{A\mu'}\right] \mathbb{E}\left[\left(\sum_{j \neq i}^{N} \xi_i^{A\mu} g_i^0\right)^{n-k} \left(\sum_{j \neq i}^{N} \xi_i^{A\mu'} g_i^0\right)^{n-k'}\right]$$
$$\approx 4n^2 (N-1)^{2n-2} \cdot q^{2n-1}$$

3. $A \neq A'$. $M = G(G-1)K^2$.

$$T_3 = 4 \sum_{k,k' \text{ odd}} \binom{n}{k} \binom{n}{k'} \mathbb{E}\left[\xi_i^{A\mu} \xi_i^{A'\mu'}\right] \mathbb{E}\left[\left(\sum_{j \neq i}^{N} \xi_i^{A\mu} g_i^0\right)^{n-k} \left(\sum_{j \neq i}^{N} \xi_i^{A'\mu'} g_i^0\right)^{n-k'}\right]$$
$$\approx 4n^2 (N-1)^{2n-2} \cdot q^{2n}.$$

With the simplifying assumptions of $N, K, G \gg 1$, we write the second moment:

$$\mathbb{E}[(\Delta E)^2] = 4n^2 (N-1)^{2n-2} (GKq^{2n-2} + GK^2 q^{2n-1} + G^2 K^2 q^{2n}). \tag{27}$$

Subsequently, we obtain the ratio of the variance to the squared mean which is eq. 7:

$$\frac{\text{var}\Delta E}{\mathbb{E}[\Delta E]^2} = \frac{1}{G}\left(\frac{1}{q} + \frac{1}{Kq^2}\right). \tag{28}$$