A BIOLOGICALLY PLAUSIBLE DENSE ASSOCIATIVE MEMORY WITH EXPONENTIAL CAPACITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Krotov and Hopfield (2021) proposed a biologically plausible two-layer associative memory network with memory storage capacity exponential in the number of visible neurons. However, the capacity was only linear in the number of hidden neurons. This limitation arose from the choice of nonlinearity between the visible and hidden units, which enforced winner-takes-all dynamics in the hidden layer, thereby restricting each hidden unit to encode only a single memory. We overcome this limitation by introducing a novel associative memory network with a threshold nonlinearity that enables distributed representations. In contrast to winner-takes-all dynamics, where each hidden neuron is tied to an entire memory, our network allows hidden neurons to encode basic components shared across many memories. Consequently, complex patterns are represented through combinations of hidden neurons. These representations reduce redundancy and allow many correlated memories to be stored compositionally. Thus, we achieve much higher capacity: exponential in the number of hidden units, provided the number of visible units is sufficiently larger than the number of hidden neurons. Exponential capacity arises because all binary states of the hidden units can become stable memory patterns with an appropriately chosen threshold. Moreover, the distributed hidden representation, which has much lower dimensionality than the visible layer, preserves class-discriminative structure, supporting efficient nonlinear decoding. These results establish a new regime for associative memory, enabling high-capacity, robust, and scalable architectures consistent with biological constraints.

1 INTRODUCTION

Associative memory networks are a class of attractor models in which the system can recall stored memories from their incomplete or noisy versions via recurrent dynamics (Krotov et al. (2025)). In such models, memories are conceptualized as the stable fixed points of the network dynamics. The number of fixed points determines the storage capacity of the network, and significant efforts have been made to construct networks with sufficiently high storage capacity to explain human memory.

The classical Hopfield network, a leading model for associative memory, has a storage capacity that scales linearly with the number of neurons in the network (Hopfield (1982)). Dense associative memories, sometimes also referred to as modern Hopfield networks (Krotov and Hopfield (2016)), are promising modifications of the classical Hopfield model. By incorporating higher-order interactions (e.g., interactions that are quadratic, rather than linear, in the input to a neuron), they achieve a storage capacity that scales super-linearly with the number of neurons. There are many possible choices for the energy function in this class of models. For instance, the power interaction vertex leads to the power-law scaling of the capacity (Baldi and Venkatesh (1987); Gardner (1987); Abbott and Arian (1987); Horn and Usher (1988); Chen et al. (1986); Krotov and Hopfield (2016)). More sophisticated shapes of the energy function result in exponential storage capacity in the number of neurons, while maintaining large basins of attraction (Demircigil et al. (2017); Lucibello and Mézard (2024)).

The naive implementation of Dense Associative Memory models, however, relies on synaptic interactions that are challenging to implement broadly in biological circuits. In particular, these models require nonlinear interactions among synapses. While several biological mechanisms could in principle support restricted forms of higher-order interactions, such as astrocytic processes, dendritic computations, or distributed neurotransmitter effects, these remain limited in scope and dictate strong

constraints on the possible shape of the energy landscape (Burns and Fukai (2023); Kozachkov et al. (2025); Kafraj et al. (2025)). The implementation of Dense Associative Memory introduced by Krotov and Hopfield (2021) does not suffer from these limitations, as it relies only on standard synaptic interactions. In this architecture, the visible neurons correspond to features of the patterns, while the hidden neurons serve as auxiliary computational elements that mediate complex interactions. Higher-order interactions among visible neurons emerge by selecting appropriate activation functions for the hidden neurons.

Nevertheless, the two-layer implementation of Krotov and Hopfield has two key limitations. First, the storage capacity is at most linear in the number of hidden neurons (Krotov (2021); Krotov and Hopfield (2021)). This is unsatisfactory from the perspective of information storage – one would like to store as much information as possible while utilizing only a small number of neurons. Second, at inference time, the network demonstrates a winner-takes-all behavior. This means that the asymptotic fixed point that the network converges to corresponds to a single hidden neuron being activated, while the rest of the hidden neurons are inactive. This behavior results in grandmother-like representations for hidden neurons, as opposed to distributed representations, which are more efficient at storing information.

Our work tackles these two limitations. Specifically, we present a novel implementation of Dense Associative Memory that achieves exponential storage capacity in the number of hidden neurons. This is accomplished with a simple yet critical change: we use a threshold activation function that does not enforce winner-takes-all dynamics. The threshold activation enables distributed memory representations–multiple hidden neurons can be active for a memory, and each hidden neuron can participate in multiple memories. As a result, all possible binary patterns of hidden neuron states become stable fixed points, enabling the network to store exponentially many memories, including highly correlated ones. Beyond high capacity, the hidden layer of the network is low-dimensional compared to the visible layer, yet it produces structured representations that preserve class-discriminative information, with memories sharing components represented close together in the hidden activity space. We establish this result through both theoretical analysis and numerical simulations, and show that the resulting fixed points also possess large basins of attraction.

Our model is closely related to the framework recently proposed by Chandra et al. (2025), which combines multiple Dense Associative Memory modules to produce a distributed code for the visible neurons. Each module performs a winner-takes-all operation similar to Krotov and Hopfield (2021), so only a single hidden neuron is active per module. By combining several modules, though, they achieve exponential storage capacity. However, we show that multiple modules are unnecessary: exponential capacity can be achieved with a single module, provided the activation function is chosen appropriately.

Beyond its biological motivation, our work also connects to a growing body of research on Dense Associative Memories in machine learning. Notably, it has been shown that Dense Associative Memory closely corresponds to the attention mechanism in transformer architectures Ramsauer et al. (2021); Hoover et al. (2023a), offering a principled framework for viewing the transformer's attention and feedforward computations as steps in a global energy minimization process. Complementary research has demonstrated that generative diffusion models, widely used in high-quality image generation, also exhibit associative memory behavior Hoover et al. (2023b); Ambrogioni (2024); Pham et al. (2025). Further studies have expanded the model's functionality: for instance, Chaudhry et al. (2023) examined its ability to store and retrieve long sequences; Burns and Fukai (2023) introduced higher-order simplicial interactions; and Dohmatob (2023); Hoover et al. (2025) proposed alternative energy functions that also support exponential storage capacity. Our results contribute to this line of work by showing how exponential storage capacity can be achieved within a biologically plausible two-layer framework, thereby bridging theoretical neuroscience with modern machine learning architectures.

In the following sections, we first formally define the model and its dynamics and derive the optimal threshold analytically for a network with fixed weights. We then present a theoretical analysis of storage capacity and basins of attraction, showing that the network exhibits large basins of attraction, making recall robust to substantial noise in the visible inputs. Next, we introduce a learning rule for storing real, correlated memories, enabling compositional memory storage, and present numerical experiments on MNIST and CIFAR-10 that demonstrate high-capacity recall, structured hidden representations, and robustness to noise. Finally, we conclude by discussing the biological plausibility

of the network and potential directions for extending the model to incorporate additional constraints and more realistic neuronal properties.

2 MODEL

In this section, we present our model and demonstrate that its storage capacity scales exponentially with the number of hidden neurons, meaning that all possible binary patterns of hidden neurons are stable fixed points.

We first define the dynamics of the system as follows:

$$\tau_v \frac{dv_i}{dt} = -v_i + \frac{1}{\sqrt{N_h}} \sum_{\mu=1}^{N_h} \xi_{i\mu} \Theta(h_\mu - \theta) \tag{1a}$$

$$\tau_h \frac{dh_\mu}{dt} = -h_\mu + \frac{\sqrt{N_h}}{N_v} \sum_{i=1}^{N_v} \xi_{\mu i} v_i \,, \tag{1b}$$

where $\Theta(\cdot)$ is the standard Heaviside step function:

$$\Theta(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ 1 & \text{if } z > 0 \,. \end{cases} \tag{2}$$

The parameter θ will be chosen to ensure the stability of all binary patterns in the hidden layer.

The network consists of N_v visible neurons (the v_i) and N_h hidden neurons (the h_μ), arranged in a bipartite architecture, i.e., without lateral connections within either layer.

Synaptic connections between visible neuron i and hidden neuron μ are reciprocal and randomly drawn from a standard normal distribution:

$$\xi_{\mu i} = \xi_{i\mu} \sim \mathcal{N}(0, 1). \tag{3}$$

The scaling factors in front of the sums are chosen purely for convenience, as they simplify subsequent expressions. Additionally, for simplicity, our theoretical analysis and experiments assume a Heaviside step function; however, Appendix F shows that this assumption is not strictly required.

2.1 STORAGE CAPACITY

To determine the storage capacity, we'll first focus on the fixed points of the dynamics given in Eq. (1). Defining

$$s_\mu \equiv \Theta(h_\mu - \theta) \,, \tag{4}$$

it is straightforward to show that in steady state, s_μ satisfies

$$s_\mu = \Theta\left(\sum_{\nu=1}^{N_h} J_{\mu\nu} s_\nu - \theta \right) \tag{5}$$

where

$$J_{\mu\nu} \equiv \frac{1}{N_v} \sum_{i=1}^{N_v} \xi_{\mu i} \xi_{i\nu} \,. \tag{6}$$

Equation (5), with the weight matrix given in Eq. (6), is very close to the classical Hopfield model; the only difference is that in the classical model, the $\xi_{\mu i}$ are binary, whereas in our model they're Gaussian. However, the classical Hopfield model works in the regime $N_v < N_h$, with memory storage possible only if $N_v < 0.138 N_h$(Amit et al. (1985)). Here, though, we'll consider a very different regime: $N_v \gg N_h$. In this limit, $J_{\mu\nu}$ approaches the identity matrix (Marchenko and Pastur

(1967)), which completely decouples the hidden neurons. Assuming the threshold, θ, is chosen correctly, this leads immediately to exponential storage capacity.

Exponential capacity clearly holds in the limit $N_v \to \infty$. What happens when N_v is finite? We show in Appendix A.1 that

$$J_{\mu\nu} = \delta_{\mu\nu} + \frac{\zeta_{\mu\nu}}{\sqrt{N_v}} \tag{7}$$

where the $\zeta_{\mu\nu}$ are independent, zero-mean, unit-variance Gaussian random variable,

$$\zeta_{\mu\nu} \sim \mathcal{N}(0,1)\,, \tag{8}$$

and here and in what follows $\delta_{\mu\nu}$ is the Kronecker delta. Thus, Eq. (5) may be written

$$s_\mu = \Theta\left(s_\mu + \frac{1}{\sqrt{N_v}} \sum_{\nu=1}^{N_h} \zeta_{\mu\nu} s_\nu - \theta\right)\,. \tag{9}$$

Because the $\zeta_{\mu\nu}$ are independent, the second term in parentheses, which we denote q_μ, scales as (see Eq. (22) in Appendix A.1)

$$|q_\mu| \equiv \left|\frac{1}{\sqrt{N_v}} \sum_{\nu=1}^{N_h} \zeta_{\mu\nu} s_\nu\right| \sim \sqrt{\frac{1}{N_v} \sum_{\nu=1}^{N_h} s_\nu^2} \le \sqrt{\frac{N_h}{N_v}} \tag{10}$$

where the second inequality follows because s_ν is either 0 or 1.

If we set $\theta = 1/2$, in the limit $N_v \gg N_h$ Eq. (9) typically has two solutions: one at $s_\mu = 0$ and one at $s_\mu = 1$. In fact, the probability that there is only one solution is the probability that $(2s_\mu - 1)q_\mu < -1$; in Appendix D (see in particular Eq. (34)) we show that

$$P_{\text{no flip}} = \Phi\left(\sqrt{\frac{N_v}{4N_a}}\right)^{N_h} \tag{11}$$

where N_a is the number of active hidden neurons. Thus, even when N_v is only about ten times larger than N_h, and the threshold is not exactly 1/2, there are approximately 2^{N_h} fixed points.

There are exponentially many fixed points, but are they stable? To answer that, we need to do stability analysis. Combining Eq. (1) with the definitions of s_μ, Eq. (4), and $J_{\mu\nu}$, Eq. (6), we have

$$h_\mu = \sum_\nu J_{\mu\nu} s_\nu\,. \tag{12}$$

Since $J_{\mu\nu}$ is approximately the identity matrix, we see that at equilibrium h_μ is close to either 0 or 1. Thus, because our nonlinearity is a step function, its derivative vanishes at equilibrium, guaranteeing the stability of the fixed points (Appendix A.2). Consequently, when we solve Eq. (1), we expect to see 2^{N_h} stable fixed points in the regime $N_v \gg N_h$. This prediction is consistent with numerical simulations, as can be seen in Figure 1a.

2.2 Basins of Attraction

Although the fixed points are stable, that still leaves the question: how big are the basins of attraction? We'll assume that noisy input enters the network via the visible units, and initially all the h_μ are zero. How far from the fixed points can the input be and still be recalled perfectly?

Assuming the noise is additive, the initial values of the visible and hidden neurons are,

$$v_i(0) = \frac{1}{\sqrt{N_h}} \sum_{\mu=1}^{N_h} \xi_{i\mu} \Theta(h_{\mu,\text{target}} - \theta) + \epsilon_i^v \tag{13a}$$

$$h_\mu(0) = 0 \tag{13b}$$

where $h_{\mu,\text{target}} = 1$ if neuron μ encodes the target memory, and 0 otherwise (motivated by the fact that h_μ is close to either 0 or 1 at the fixed points; see Eq. (12)).

To reach the target fixed point in both the hidden and visible unit space, the hidden neurons must evolve to their target values, $h_{\mu,\text{target}}$, before the visible units change much. That requires the visible units to evolve much more slowly than the hidden units, which we can guarantee by setting $\tau_h \ll \tau_v$ (see Appendix G). With this condition, at a time t satisfying $\tau_h \ll t \ll \tau_v$, $h_\mu(t)$ reaches equilibrium while $v_i(t)$ is still approximately equal to $v_i(0)$. Using Eq. (1a) with $dv_i/dt = 0$ along with Eq. (6), that equilibrium is given by

$$h_\mu(t) = \sum_\nu J_{\mu\nu}\Theta(h_{\nu,\text{target}} - \theta) + \frac{\sqrt{N_h}}{N_v}\sum_{i=1}^{N_v}\xi_{\mu i}\epsilon_i^v + \mathcal{O}(t/\tau_v)\,. \tag{14}$$

Using Eq. (7), we see that the first term is $\Theta(h_{\mu,\text{target}} - \theta) + \mathcal{O}(\sqrt{N_h/N_v})$. And the second term scales as $\sigma_v\sqrt{N_h/N_v}$ where σ_v^2 is the variance of the noise. Thus, so long as

$$\text{Var}[\epsilon] \ll \frac{N_v}{N_h}\,, \tag{15}$$

$h_\mu(t)$ will be close to its target value when $t \ll \tau_v$. Since $v_i(t)$ is close to its target value at that time, it will stay close, and asymptotically the target pattern will be recovered. Given that $N_v \gg N_h$, ϵ_i^v can be very large without affecting recall. Thus, the basin of attraction is very large (see Figure 1b and Appendix D).

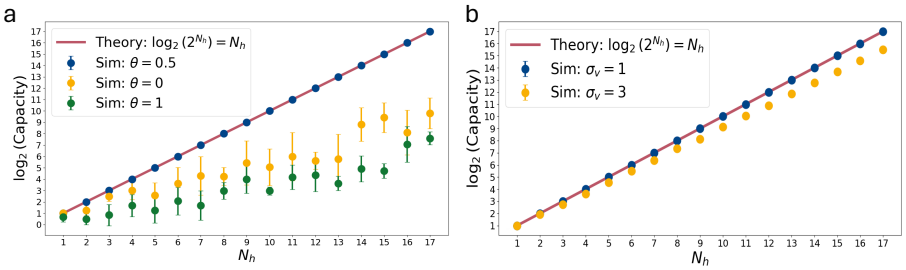

Figure 1: Capacity versus the number of hidden units, N_h, with $N_v = 100N_h$ and $\tau_v = 20\tau_h$. (a) Capacity for different thresholds, θ. The highest storage capacity is achieved when the threshold is set to its optimal theoretical value , $\theta = 0.5$. (b) The effect of noise in the visible layer (ϵ_i^v in Eq. (13a)), shown for different noise variances, demonstrates the large basin of attraction of the fixed points.

2.3 Biological Plausibility

Compared to Krotov and Hopfield (2021), our model exhibits greater biological plausibility in several respects.

The activation function used here is local and keeps neuron activity within a biologically realistic range. In contrast, in Krotov and Hopfield (2021), Model A is not biologically plausible because the hidden neuron activities can grow to unrealistically large values as a consequence of the power-law activation, which does not reflect realistic neural firing. Models B and C, on the other hand, rely on non-local activation functions, softmax and spherical normalization, respectively, which are biologically implausible unless additional mechanisms are assumed.(see Appendix C for more details).

Although our theoretical analysis focuses on symmetric weights and a global threshold for all neurons for simplicity, these assumptions are not restrictions of the model. Experimentally, we show that networks with asymmetric weights and heterogeneous neuron-specific thresholds also achieve stable recall. Allowing asymmetric weights is important because exact symmetry is rarely observed in biological neural circuits, yet memory networks can remain robust even without it. Similarly, heterogeneous thresholds capture the variability in neuron excitability across real neurons and demonstrate that stable memory dynamics do not require finely tuned, uniform parameters. Together, these features indicate that our model better reflects realistic neural mechanisms while retaining associative memory functionality. Figure 9 in Appendix E shows representative recall

examples for networks with asymmetric weights and heterogeneous thresholds that stored MNIST and CIFAR-10 images, respectively, using a learning rule similar to that discussed in Subsection 3.1, with the key differences being the absence of a symmetry restriction on the weights and the allowance of heterogeneous thresholds.

3 RESULTS

3.1 LEARNING RULE

So far we have focused on storage capacity with fixed synaptic weights. A natural next step is to understand how these weights can be learned. In this section, we introduce a learning rule that reflects *compositional learning*: a small number of simple, reusable components can be combined to form complex patterns, and conversely, complex patterns can be decomposed into simpler components.

In the steady-state, visible activity in Eq. (1) can be expressed as

$$\mathbf{v} = \frac{1}{\sqrt{N_h}} \sum_{\mu=1}^{N_h} \boldsymbol{\xi}_\mu s_\mu, \tag{16}$$

where $\boldsymbol{\xi}_\mu$ is the μ-th column of $\xi \in \mathbb{R}^{N_v \times N_h}$, i.e., $(\boldsymbol{\xi}_\mu)_i = \xi_{i\mu}$. If only hidden neuron μ is active, the visible state equals $\boldsymbol{\xi}_\mu$. A visible memory is thus called *basic* if it corresponds to a single active hidden neuron, and *complex* if it is formed by the activation of multiple hidden neurons, i.e. a composition of several basic memories.

The goal of learning is to find a synaptic weight matrix ξ and a threshold θ such that a set of target memories $\{\mathbf{v}_m \in \mathbb{R}^{N_v}\}_{m=1}^{M}$ approximately correspond to stable fixed points of the network dynamics, with $M \gg N_h$ (e.g., MNIST or CIFAR-10). This is achieved using the following optimization procedure,

$$(\xi, \theta) = \arg\min_{\xi,\theta} \sum_{m=1}^{M} \left\| \mathbf{v}_m - \frac{1}{\sqrt{N_h}} \sum_{\mu=1}^{N_h} \boldsymbol{\xi}_\mu \, \Theta\left(\frac{\sqrt{N_h}}{N_v} \boldsymbol{\xi}_\mu^\top \mathbf{v}_m - \theta\right) \right\|^2, \tag{17}$$

where s_μ is replaced by its target steady-state value. This learning rule is identical to the one proposed in Radhakrishnan et al. (2020). We used Xavier initialization for the weights and approximated the threshold function Θ with a sharp sigmoid to allow gradient-based training.

3.2 EXPERIMENTS

Figure 2 shows recall results after storing 60,000 MNIST digits with $N_v = 784$ and $N_h = 50$. Despite the high correlation among patterns, the network learns 57913 unique minima corresponding to the 60,000 stored images. Variants of the same digit produce hidden representations that are distinct yet partially overlapping, and the recalled visible states remain recognizable.

Figure 3a shows the learned basic memories for the MNIST dataset, which correspond to the columns of ξ. As shown in Figure 3b, these basic memories are nearly orthogonal, consistent with Eq. (7).

To evaluate the generality of the proposed learning rule beyond MNIST, we applied the same procedure to the CIFAR-10 dataset. In this case, $N_v = 3072\,(3 \times 32 \times 32)$, and to compensate for the increased complexity of this dataset, we used a network with 500 hidden neurons, compared to 50 hidden neurons for MNIST. Figure 4 presents examples of the cues alongside their recalls. These results show that the network is able to reconstruct interpretable outputs from the learned representations, despite storing a large number of complex memories (50,000) and significantly violating the condition $N_v \gg N_h$. Importantly, these images are highly correlated, yet the network produces 49982 unique stable minima corresponding to the stored memories, with each memory representation being both stable and interpretable.

The learned basic memories for CIFAR-10 images are shown in Figure 5a. They form a more heterogeneous set, yet remain nearly orthogonal, as shown in Figure 5b.

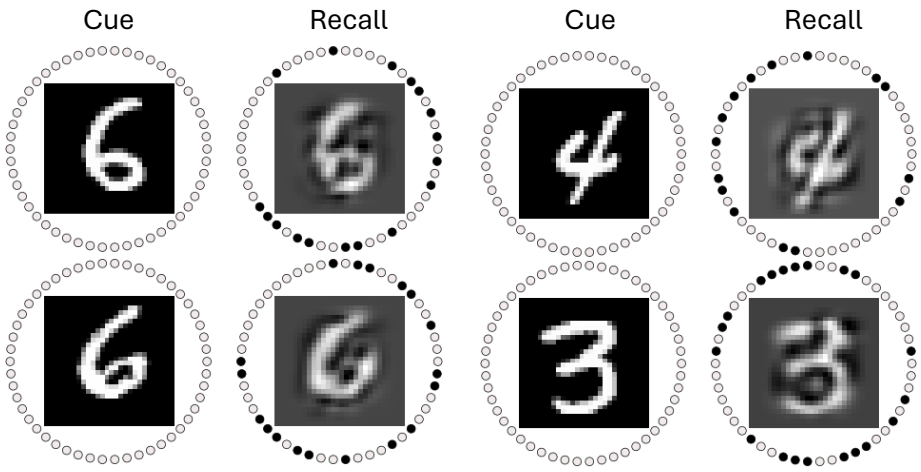

Figure 2: Examples of recall in a network with 50 hidden neurons that memorized 60,000 MNIST images. Hidden neurons are shown on the ring, and visible neurons are visualized as two-dimensional images. On the ring, black indicates high activity, and white indicates low activity. Highly correlated images of every digit, for instance, the digit 6 shown here, converge to unique but overlapping hidden representations.

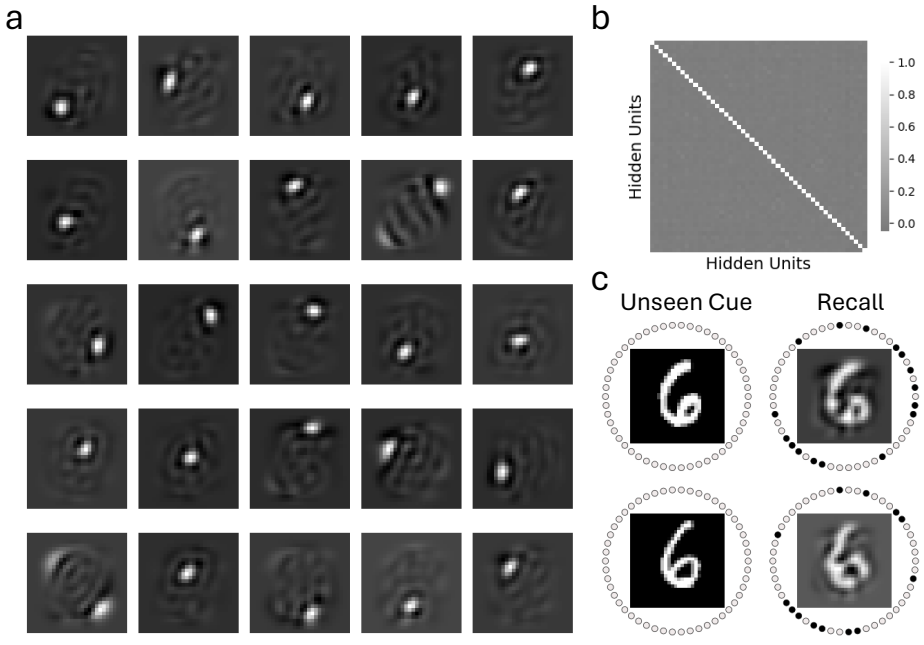

Figure 3: a) 25 (out of 50) columns of the learned weight matrix, for MNIST images, which serve as basic memories, are shown as two-dimensional images. b) Correlation matrix of the basic patterns, which correspond to the hidden units. c) The network generalizes compositionally, associating unseen cues with interpretable fixed points.

The network learns an effective threshold of $\theta = 0.21$ for MNIST and $\theta = 0.43$ for CIFAR-10. Note that the statistics of the learned base memories in MNIST and CIFAR-10 differ from one another and from the normal distribution assumed in the theory, which explains the difference between the learned thresholds.

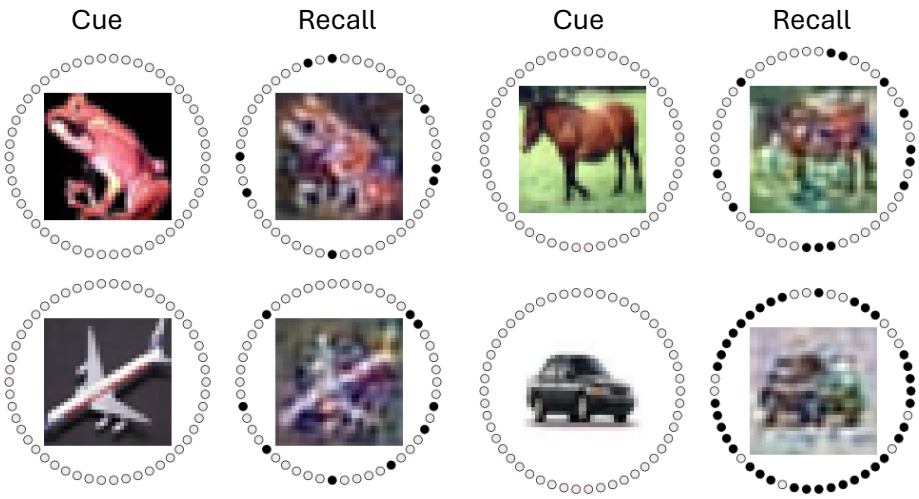

Figure 4: Examples of recall in a network with 500 hidden neurons that memorized 50,000 CIFAR-10 images. Hidden neurons are arranged on a ring (50 out of 500), while visible neurons are shown as two-dimensional images. On the ring, black indicates high activity, and white indicates low activity.

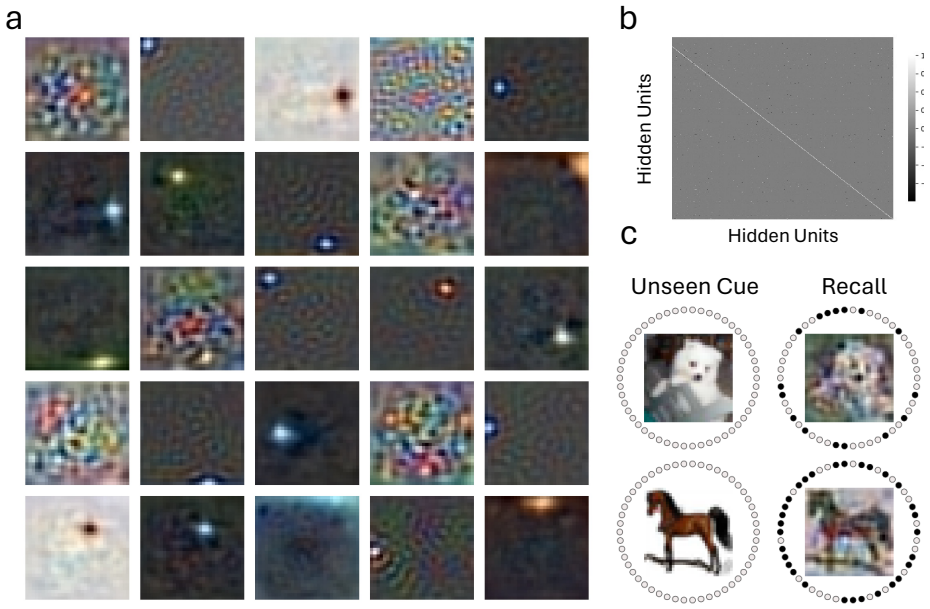

Figure 5: a) 25 (out of 500) columns of the learned weight matrix for CIFAR-10 images, which serve as basic memories, are shown as images. b) Correlation matrix of the basic patterns, which correspond to the hidden units. c) The network generalizes compositionally, associating unseen cues with interpretable fixed points.

For this system to function as an associative memory, new "unseen" cues should converge toward the approximately correct fixed points. For example, cues related to dogs should end up near fixed points associated with dogs, not horses. As shown in Figure 3c and Figure 5c, this behavior is indeed observed. The network converges to the nearest minimum of the energy landscape relative to the cue.

Importantly, when the learned basic memories are expressive enough, this nearest minimum can correspond to a stable representation that is very close to the unseen cue itself rather than to a memorized pattern. In other words, the network not only memorizes but also generalizes: the learned

basic memories shape the energy landscape so that unseen inputs are mapped to stable attractors that capture their distinctive features.

For example, in Figure 3c, when two unseen images of the digit "6" are presented, the network converges to two distinct attractors that preserve the distinguishing details of each input while still sharing overlapping components in the hidden layer that identify them as class "6". This illustrates an advantage of our model: it supports both memorization and generalization through its learned basic components.

To quantitatively evaluate this behavior, we trained nonlinear classifiers on the recalled representations of the stored images, and tested them on the recalled representations of unseen images (Convolutional neural network (CNN) for visible representations and multilayer perceptron (MLP) for hidden representations). For comparison, we trained CNNs directly on the original stored images and tested them on the original unseen images as well. This allows us to assess the classifiability of the recalled visible representations by the associative memory network relative to the original memories. (see Appendix B for details of the classifiers)

For the MNIST dataset, classification accuracy is high for both the hidden and visible representations. This is a desirable property, as it indicates that the lower-dimensional hidden representations still preserve strong class discriminability. In MNIST, the raw pixel space itself carries strong class structure: two images of the same digit are highly correlated and closer to each other than images of different digits. Consequently, the hidden neurons retain this information almost perfectly, having structured and meaningful encodings in which correlated memories are represented close together and remain classifiable.

For CIFAR-10, classification accuracy is high for the visible representations but low for the hidden ones. This difference arises because, in CIFAR-10, two images of the same class (for example, dogs) are not necessarily correlated in raw pixel space, so the hidden layer, which is a nonlinear transformation of those pixels, does not exhibit a clear class structure. The classifier used for the visible representations in this analysis is a CNN, which, when trained on raw pixel data, first learns a nonlinear transformation that maps images of the same class close together in a learned feature space while separating images from different classes. After this transformation, classification is performed using a linear decision boundary in that space. This ability to internally build such class-specific representations explains why classification accuracy remains higher for the visible neurons.

Overall, the high performance of the visible representations for both MNIST and CIFAR-10 demonstrates that the recalled representations remain highly class-discriminative and that the associative memory preserves the essential structure of the data, enabling *compositional generalization* to unseen examples.(for the CIFAR10 dataset, the classification accuracy can be increased by scaling up the associative memory, including N_h, the epoch size, the optimization step, and the number of training samples.)

Representation	**MNIST Accuracy**	**CIFAR-10 Accuracy**
Recalled Hidden Patterns	95%	40%
Recalled Visible Patterns	98%	56%
Original Images	99%	88%

Table 1: Classification test accuracy of nonlinear classifiers trained and tested on recalled hidden and visible representations, as well as on the original images for comparison, for MNIST and CIFAR-10 datasets. High accuracy on visible representations for both datasets demonstrates that the recalled representations remain highly class-discriminative, while the lower accuracy on hidden representations for CIFAR-10 reflects the lack of strong class structure in raw pixel space.

4 Conclusion

This work introduces a novel Dense Associative Memory Krotov and Hopfield (2021) that achieves exponential storage capacity in the number of hidden neurons, overcoming the limitations of previous two-layer models. By using a threshold activation function, with a theoretically derived threshold, the network supports distributed hidden representations, allowing each hidden neuron to participate in

multiple memories. This enables compositional storage of complex and correlated patterns, reducing redundancy while maintaining robust retrieval.

Specifically, the network achieves exponential capacity, 2^{N_h}, using only $N_h N_v$ parameters. In contrast, previous two-layer implementations were limited to a maximum capacity of N_h (Krotov and Hopfield (2021)). As a result, the number of memories per weight grows as $\frac{2^{N_h}}{N_h N_v} \approx 2^{N_h}$, while in previous implementations it is at best $\frac{1}{N_v}$. Even for complex datasets such as MNIST and CIFAR-10, networks with only 50 and 500 hidden units, respectively, were able to store tens of thousands of highly correlated memories and associate the vast majority of them with unique minima, whereas previous models could not store more memories than the number of hidden units (see AppendixC).

Beyond storing exponentially many memories, the network is also able to generalize to novel inputs. This behavior arises because the hidden layer encodes a set of basic memories that can be flexibly composed to represent previously unseen patterns, associating them with distinct minima rather than forcing retrieval toward the nearest stored pattern, while still producing meaningful, class-consistent representations. Our results are consistent with biological principles of feature learning, as embodied in hierarchical predictive coding models, which detect novelty and generalize by recombining features across successive levels of abstraction (Li et al. (2025); Salvatori et al. (2021)). This mechanism further illustrates that learning the underlying compositional structure of naturalistic data enables a biological associative memory to effectively support both memorization and generalization.

The model is biologically grounded, relying solely on standard pairwise synapses and a local activation function. We also provide evidence that it achieves stable recall even in the presence of asymmetric weights and heterogeneous neuronal thresholds. Moreover, the hidden layer forms low-dimensional representations that preserve class-discriminative information, placing memories with shared components close together in activity space. This structured organization supports efficient nonlinear decoding.

Overall, this work establishes a new regime for associative memory that combines high capacity, robust recall, compositional and interpretable representations, and biological plausibility. It provides a theoretical foundation for scalable memory systems that bridge neuroscience models with modern machine learning architectures.

Future work will focus on developing a biologically plausible learning rule and on examining the model's capacity under additional biological constraints, including sparse connectivity and adherence to Dale's law.

REFERENCES

Laurence F Abbott and Yair Arian. Storage capacity of generalized networks. *Physical review A*, 36 (10):5091, 1987.

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024.

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.

Pierre Baldi and Santosh S Venkatesh. Number of stable points for spin-glasses and neural networks of higher orders. *Physical Review Letters*, 58(9):913, 1987.

Thomas F Burns and Tomoki Fukai. Simplicial hopfield networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Sarthak Chandra, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete. Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 2025.

Hamza Chaudhry, Jacob Zavatone-Veth, Dmitry Krotov, and Cengiz Pehlevan. Long sequence hopfield memory. *Advances in Neural Information Processing Systems*, 36:54300–54340, 2023.

HH Chen, YC Lee, GZ Sun, HY Lee, Tom Maxwell, and C Lee Giles. High order correlation model for associative memory. In *AIP Conference Proceedings*, volume 151, pages 86–99, 1986.

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, 2017.

Elvis Dohmatob. A different route to exponential storage capacity. In *Associative Memory $\{\backslash\&\}$ Hopfield Networks in 2023*, 2023.

Elizabeth Gardner. Multiconnected neural network models. *Journal of Physics A: Mathematical and General*, 20(11):3453, 1987.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. *Advances in neural information processing systems*, 36:27532–27559, 2023a.

Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories. In *Associative Memory $\{\backslash\&\}$ Hopfield Networks in 2023*, 2023b.

Benjamin Hoover, Zhaoyang Shi, Krishnakumar Balasubramanian, Dmitry Krotov, and Parikshit Ram. Dense associative memory with epanechnikov energy. *arXiv preprint arXiv:2506.10801*, 2025.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79:2554–2558, 1982.

D Horn and M Usher. Capacities of multiconnected memory models. *Journal de Physique*, 49(3): 389–395, 1988.

Mohadeseh Shafiei Kafraj, Dmitry Krotov, Brendan A Bicknell, and Peter E Latham. A biologically plausible associative memory network. In *New Frontiers in Associative Memories*, 2025.

Leo Kozachkov, Jean-Jacques Slotine, and Dmitry Krotov. Neuron-astrocyte associative memory. *Proceedings of the National Academy of Sciences of the United States of America*, 122(21): e2417788122, 2025.

Dmitry Krotov. Hierarchical associative memory. *arXiv preprint arXiv:2107.06446*, 2021.

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

Dmitry Krotov and John J Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021.

Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. Modern methods in associative memory. *arXiv preprint arXiv:2507.06211*, 2025.

T Ed Li, Mufeng Tang, and Rafal Bogacz. Predictive coding model detects novelty on different levels of representation hierarchy. *Neural computation*, 37(8):1373–1408, 2025.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024.

VA Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.(NS)*, 72(114):4, 1967.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory. *arXiv preprint arXiv:2505.21777*, 2025.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117, 2020.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Tommaso Salvatori, Yuhang Song, Yujian Hong, Lei Sha, Simon Frieder, Zhenghua Xu, Rafal Bogacz, and Thomas Lukasiewicz. Associative memories via predictive coding. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 3874–3886, 2021.

A Technical Appendices and Supplementary Material

A.1 Distributional Properties of $\zeta_{\mu\nu}$

We define the matrix elements

$$\zeta_{\mu\nu} = \frac{1}{\sqrt{N_v}} \sum_{i=1}^{N_v} \xi_{\mu i}\xi_{i\nu}, \quad \mu \neq \nu, \tag{18}$$

where $\xi_{\mu i}$s are randomly drawn from a standard normal distribution.

Each product $\xi_{\mu i}\xi_{i\nu}$ is a zero-mean random variable, since $\xi_{\mu i}$ and $\xi_{i\nu}$ are independent with zero mean.

By the central limit theorem, the sum of these N_v independent terms converges in distribution to a Gaussian. Specifically,

$$\sqrt{N_v}\left(\frac{1}{N_v}\sum_{i=1}^{N_v}\xi_{\mu i}\xi_{i\nu} - \mathbb{E}[\xi_{\mu i}\xi_{i\nu}]\right) \xrightarrow{d} \mathcal{N}(0,1), \tag{19}$$

given that $\mathbb{E}[\xi_{\mu i}\xi_{i\nu}] = 0$, and $\text{Var}[\xi_{\mu i}\xi_{i\nu}] = 1$,

$$\zeta_{\mu\nu} \xrightarrow{d} \mathcal{N}(0,1). \tag{20}$$

Now consider the random variable q_μ defined as:

$$q_\mu = \frac{1}{\sqrt{N_v}} \sum_{\nu=1}^{N_h} \zeta_{\mu\nu} s_\nu, \tag{21}$$

This is a random variable with respect to the index μ with s_ν fixed. Its variance is given by

$$\text{Var}\left[\frac{1}{\sqrt{N_v}}\sum_{\nu=1}^{N_h}\zeta_{\mu\nu}s_\nu\right] = \frac{1}{N_v}\sum_{\nu=1}^{N_h} s_\nu^2 \text{Var}[\zeta_{\mu\nu}] = \frac{1}{N_v}\sum_{\nu=1}^{N_h} s_\nu^2 \tag{22}$$

where we used the fact that the $\zeta_{\mu\nu}$ are independent random variables with mean 0 and variance 1.

A.2 Stability of the fixed points

The stability of fixed points is determined by the Jacobian of the system. Grouping the variables into $(\mathbf{v}, \mathbf{h})$, corresponding to the visible and hidden units respectively, the Jacobian has the block structure

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{vv} & \mathbf{A}_{vh} \\ \mathbf{A}_{hv} & \mathbf{A}_{hh} \end{bmatrix}. \tag{23}$$

For the diagonal blocks, consider first the visible units. We have

$$\frac{\partial \dot{v}_i}{\partial v_j} = \begin{cases} -1, & j = i, \\ 0, & j \neq i, \end{cases} \quad \Rightarrow \quad \mathbf{A}_{vv} = -\mathbf{I}_{N_v}, \tag{24}$$

where $\mathbf{I}_{N_v}$ is the $N_v \times N_v$ identity matrix. Similarly, for the hidden units,

$$\frac{\partial \dot{h}_\mu}{\partial h_\nu} = \begin{cases} -1, & \nu = \mu, \\ 0, & \nu \neq \mu, \end{cases} \quad \Rightarrow \quad \mathbf{A}_{hh} = -\mathbf{I}_{N_h}, \tag{25}$$

where $\mathbf{I}_{N_h}$ is the $N_h \times N_h$ identity matrix.

For the off-diagonal blocks, the derivative of the Heaviside step function in Eq. (2) is zero almost everywhere,

$$\Theta'(z) = 0, \qquad z \neq 0. \tag{26}$$

Therefore, away from threshold crossings ($h_\mu \neq \theta$) in Eq. (1a),

$$\frac{\partial \dot{v}_i}{\partial h_\mu} = 0 \quad \Rightarrow \quad \mathbf{A}_{vh} = \mathbf{0}. \tag{27}$$

The hidden dynamics depend explicitly on the visible variables:

$$\frac{\partial \dot{h}_\mu}{\partial v_i} = \xi_{\mu i}, \quad \Rightarrow \quad \mathbf{A}_{hv} = \big(\xi_{\mu i}\big). \tag{28}$$

Putting everything together, the Jacobian is lower-triangular,

$$\mathbf{A} = \begin{bmatrix} -\mathbf{I}_{N_v} & \mathbf{0} \\ \mathbf{A}_{hv} & -\mathbf{I}_{N_h} \end{bmatrix}. \tag{29}$$

The eigenvalues of a triangular matrix are its diagonal entries, which in this case are all equal to -1. Hence, all fixed points of the dynamics are stable.

B Details of the Classifiers

Table 2: Architecture and Parameters of the CNN Classifier

Block	Layer Type	Channels / Filters	Kernel / Pool	Activation
Conv Block 1	2 × Conv2D + BatchNorm2D	$3 \to 32$	3×3, MaxPool(2)	ReLU
Conv Block 2	2 × Conv2D + BatchNorm2D	$32 \to 64$	3×3, MaxPool(2)	ReLU
Conv Block 3	2 × Conv2D + BatchNorm2D	$64 \to 128$	3×3, MaxPool(2)	ReLU
Flatten	–	Auto-computed (f)	–	–
Fully Connected 1	Linear	$f \to 128$	–	ReLU
Fully Connected 2	Linear	$128 \to N_{classes}$	–	Softmax

Table 3: Architecture and Parameters of the MLP Classifier

Layer	Type	Dimensions / Units	Activation
Input	Linear	Input dimension = d	–
Hidden Layer 1	Linear	$d \to 256$	ReLU
Hidden Layer 2	Linear	$256 \to 128$	ReLU
Output Layer	Linear	$128 \to N_{classes}$	Softmax

C Comparison with previous Dense Associative Memory networks

Krotov and Hopfield (2021) proposed a two-layer associative memory defined as

$$\tau_v \frac{dv_i}{dt} = -v_i + \sum_{\mu=1}^{N_h} w_{i\mu} f(h_\mu), \tag{30a}$$

$$\tau_h \frac{dh_\mu}{dt} = -h_\mu + \sum_{i=1}^{N_v} w_{\mu i} g(v_i), \tag{30b}$$

where $w_{\mu i} = w_{i\mu}$ and each $\mathbf{w}_\mu \in \mathbb{R}^{N_v}$ represents a stored memory. Three choices for the nonlinearities f and g were introduced in Krotov and Hopfield (2021), summarized in Table 4.

We evaluate their recall performance (Models A, B, and C) together with our proposed model. Model A is tested using $f(h_\mu) = h_\mu^5$ (A i) and $f(h_\mu) = h_\mu^{10}$ (A ii).

Despite differences, all three models operate under the same effective mechanism: during recall, a single hidden neuron becomes strongly active while the others remain suppressed, allowing only one memory per hidden neuron.

In contrast, our nonlinearity produces a fundamentally different recall regime. Multiple hidden neurons remain active simultaneously, enabling each hidden neuron to encode multiple stored patterns. This results in a dramatic increase in storage capacity: with only fifty hidden neurons, our model successfully stores all MNIST images (60,000) with high recall accuracy. By comparison, Models A, B, and C are limited to 50 memories and often fail to recall reliably (e.g., Model A i and Model C), as shown in Table 4.

And from a biological perspective, the nonlinearity used in Model A is not plausible, because the power-law activation causes hidden neuron activity to reach unrealistically high values during recall. Models B and C also rely on non-local activation functions, which would require additional circuit mechanisms to implement. In contrast, our model maintains bounded activity, and the nonlinearity is fully local.

Model	$f(h_\mu)$	$g(v_i)$	N_h	# Stored Memories	Recall Performance
A (i)	h_μ^5	$\text{sign}(v_i)$	50	50	12%
A (ii)	h_μ^{10}	$\text{sign}(v_i)$	50	50	84%
B	$\frac{e^{h_\mu}}{\sum_\nu e^{h_\nu}}$	v_i	50	50	90%
C	h_μ^5	$\frac{v_i}{\sqrt{\sum_j v_j^2}}$	50	50	2%
Our model	$\Theta(h_\mu - \theta)$	v_i	50	**60,000**	**98%**

Table 4: Nonlinearities used in the Dense Associative Memory models from Krotov and Hopfield (2021) and in our model, and a comparison of their recall performance. Recall performance is the percentage of recalled digits that are classified correctly.

D THE RATIO BETWEEN THE NUMBER OF VISIBLE NEURONS AND HIDDEN NEURONS

From Eqs. (10) and (14), and setting $\theta = 0.5$, we have

$$s_\mu(t) = \Theta\left(s_{\mu,\text{target}} + q_{\mu,\text{target}} + \frac{\sqrt{N_h}}{N_v} \sum_{i=1}^{N_v} \xi_{\mu i} \epsilon_i^v + \mathcal{O}(t/\tau_v) - 0.5 \right), \tag{31}$$

where $q_{\mu,\text{target}}$ is a normally distributed random variable with variance $\frac{N_a}{N_v}$ (see 22), where N_a is the number of active hidden neurons. The third term in the parentheses has variance $\sigma_v^2 N_h / N_v$. Thus, in the limit $\tau_v \gg \tau_h \gg t$, we can approximate

$$s_\mu(t) = \Theta\Big(s_{\mu,\text{target}} + z_{\mu,\text{target}} - 0.5 \Big), \tag{32}$$

where

$$z_{\mu,\text{target}} \sim \mathcal{N}\big(0, \sigma_z^2\big), \quad \sigma_z^2 = \frac{N_a + \sigma_v^2 N_h}{N_v}. \tag{33}$$

Consequently, the probability of having no bit flip is

$$P_{\text{no flip}} = \Phi(1/2\sigma_z)^{N_h} \tag{34}$$

where Φ is the cumulative normal function. This probability rapidly approaches 1 as the σ_z becomes small. In this regime, both the stability of the fixed point and correct recall are ensured.

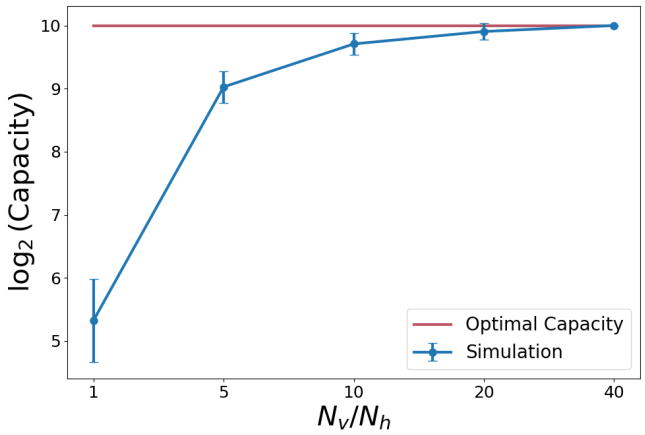

Figure 6: Capacity versus the ratio between visible and hidden neurons, for a fixed value of $N_h = 10$.

And, to determine how many hidden neurons are required for real world memories with diverse statistics, note that memories from an N_v-dimensional space are recalled within an at most N_h-dimensional subspace spanned by the N_h basic memories defined by the hidden to visible weights. If this subspace is not expressive enough, the reconstructed images will not be recognizable, particularly for complex datasets such as CIFAR-10.

In summary, the number of hidden neurons must be sufficiently smaller than the number of visible neurons to guarantee stable recall, but it must also be large enough to represent the statistical structure of the stored memories so that the reconstructions remain recognizable.

Figure 7 and Figure 8 show representative recall examples on MNIST and CIFAR-10, for networks with an insufficient number of hidden neurons ($N_h = 16$).

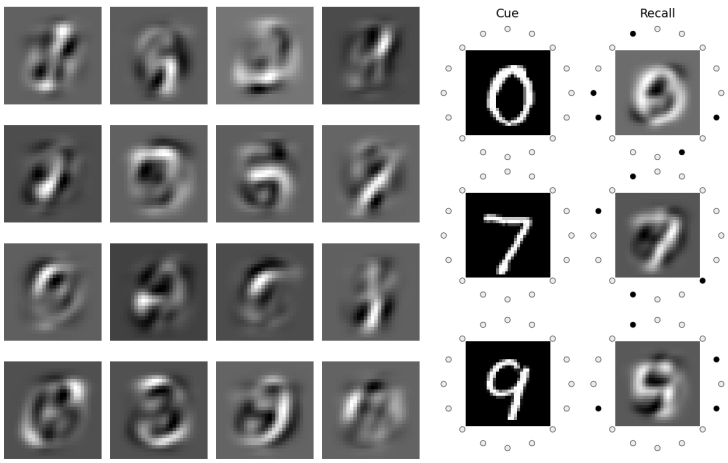

Figure 7: Learned basic memories (columns of learned weight matrix) and representative recall examples for MNIST, for a network with an insufficient number of hidden neurons ($N_h = 16$).

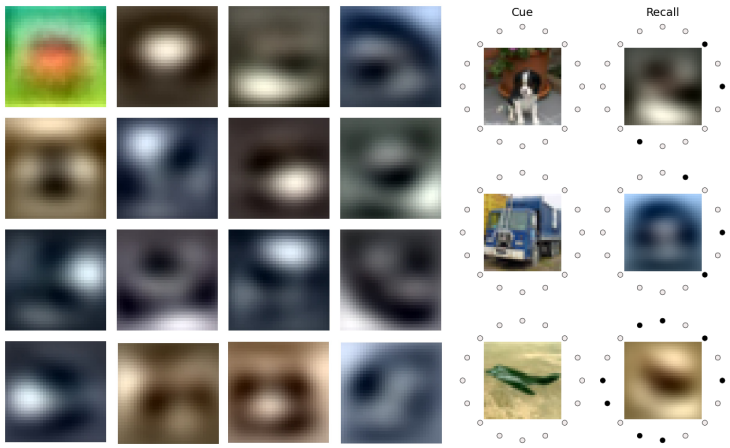

Figure 8: Learned basic memories (columns of learned weight matrix) and representative recall examples for CIFAR-10, for a network with an insufficient number of hidden neurons ($N_h = 16$).

E Asymmetric weights and heterogeneous thresholds

Figure 9 shows representative recall examples for networks with asymmetric weights and heterogeneous neuron thresholds on both MNIST and CIFAR-10.

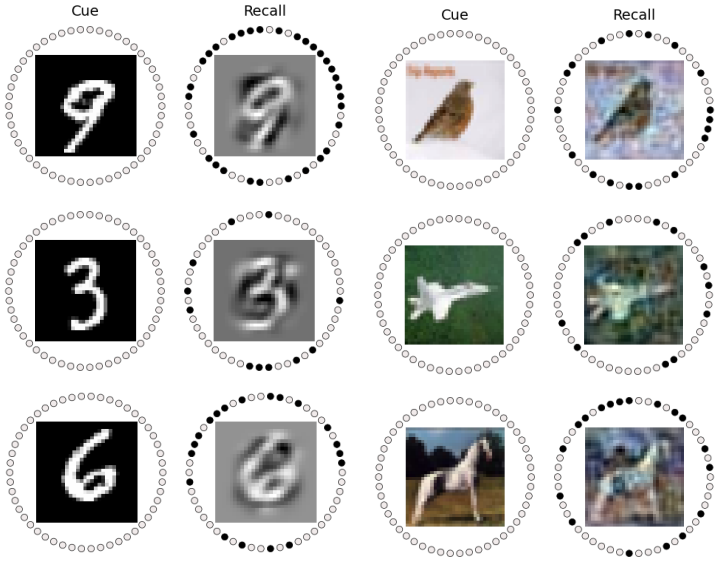

Figure 9: Representative recall examples for networks with asymmetric weights and heterogeneous neuron thresholds. On the left are examples of cue and recall for a network that stored 60,000 MNIST images with 50 hidden neurons, and on the right are examples for a network that stored 50,000 CIFAR-10 images with 500 hidden neurons.

F Sigmoid Activation Functions

In both the theory and the experiments, we used the Heaviside step function for the activation of hidden neurons. A smooth step function was used only during optimization.

However, the Heaviside function is not the only valid choice of activation. In fact, any sigmoid function that has three intersections with the identity line as showin in Figure 10a where the middle

intersection at $1/2$ is an unstable fixed point of (9), and the intersections near 0 and 1 are stable fixed points, is a valid activation function that guarantees stable recall as shown in .

We demonstrated this experimentally by showing that recall remains perfect even when a smooth sigmoid is used in place of the Heaviside function, as shown in Figure 10.

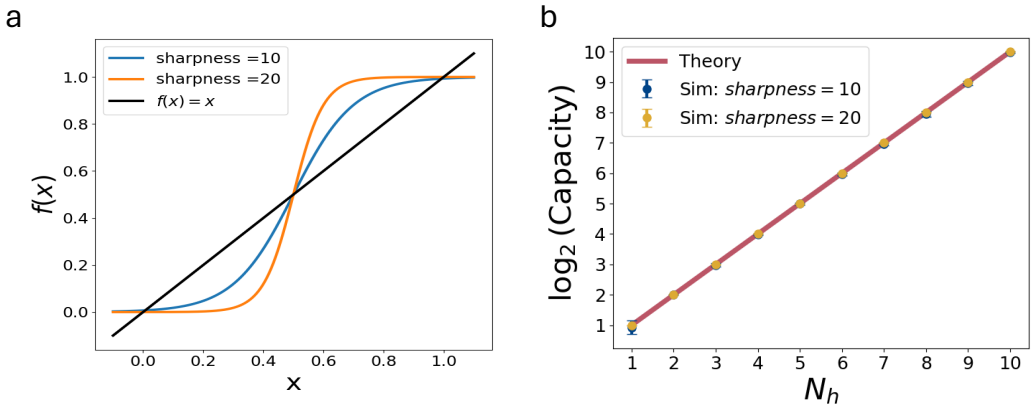

Figure 10: a) Sigmoid nonlinearities with different sharpness, each having three intersections with the identity line. b) Both nonlinearities produce stable recall, and the capacity remains exponential.

G THE TIME CONSTANT OF HIDDEN AND VISIBLE NEURONS

Figure 11 shows the impact of the time constant ratio between visible and hidden neurons on recall performance. As discussed in Section 3.2 on the basin of attraction, the visible neurons must be sufficiently slower than the hidden neurons because, during the cue, only the visible neurons receive input while the hidden neurons are initialized to zero. The visible neurons therefore need to evolve slowly enough to allow the hidden neurons to reach their correct steady state before the visible pattern changes significantly.

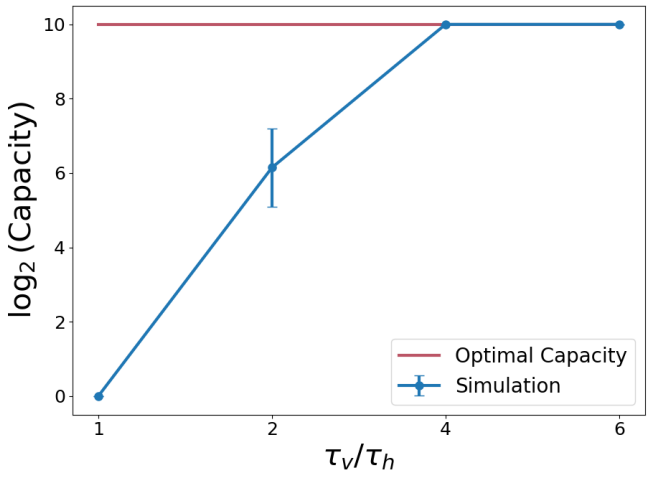

Figure 11: Capacity versus the ratio between the time constant of visible and hidden neurons for $N_h = 10$ and $N_v = 100N_h$.