LARGE ASSOCIATIVE MEMORY PROBLEM IN NEUROBIOLOGY AND MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense Associative Memories or modern Hopfield networks permit storage and reliable retrieval of an exponentially large (in the dimension of feature space) number of memories. At the same time, their naive implementation is non-biological, since it seemingly requires the existence of many-body synaptic junctions between the neurons. We show that these models are effective descriptions of a more microscopic (written in terms of biological degrees of freedom) theory that has additional (hidden) neurons and only requires two-body interactions between them. For this reason our proposed microscopic theory is a valid model of large associative memory with a degree of biological plausibility. The dynamics of our network and its reduced dimensional equivalent both minimize energy (Lyapunov) functions. When certain dynamical variables (hidden neurons) are integrated out from our microscopic theory, one can recover many of the models that were previously discussed in the literature, e.g. the model presented in "Hopfield Networks is All You Need" paper. We also provide an alternative derivation of the energy function and the update rule proposed in the aforementioned paper and clarify the relationships between various models of this class.

1 INTRODUCTION

Associative memory is defined in psychology as the ability to remember (link) many sets, called memories, of unrelated items. Prompted by a large enough subset of items taken from one memory, an animal or computer with an associative memory can retrieve the rest of the items belonging to that memory. The diverse human cognitive abilities which involve making appropriate responses to stimulus patterns can often be understood as the operation of an associative memory, with the "memories" often being distillations and consolidations of multiple experiences rather than merely corresponding to a single event.

The intuitive idea of associative memory can be described using a "feature space". In a mathematical model abstracted from neurobiology, the presence (or absence) of each particular feature i is denoted by the activity (or lack of activity) of a model neuron v_i due to being directly driven by a feature signal. If there are N_f possible features, there can be only at most N_f^2 distinct connections (synapses) in a neural circuit involving only these neurons. Typical cortical synapses are not highly reliable, and a cortical synapse stores no more than one or two bits of information. The description of a particular memory requires roughly N_f bits of information. Such a system can therefore store at most $\sim N_f$ unrelated memories. Simple artificial neural network models of associative memory (based on attractor dynamics of feature neurons and understood through an energy function describing this dynamics) exhibit this limitation even with precise synapses, with limits of memory storage to less than $\sim 0.14 N_f$ memories (Hopfield, 1982).

Situations often arise in which the number N_f is small and the desired number of memories far exceeds $\sim N_f$. For example, a small part of a high-resolution photograph may contain only 1000 pixels, but the number of describable "objects" which might occur in such a fragment is far larger. The only way to address this fundamental problem, and to increase the number of memories for a fixed N_f, is to introduce additional circuitry and additional synapses.

The starting point of this paper is a machine learning approach to associative memory based on an energy function and attractor dynamics in the space of N_f variables, called Dense Associative

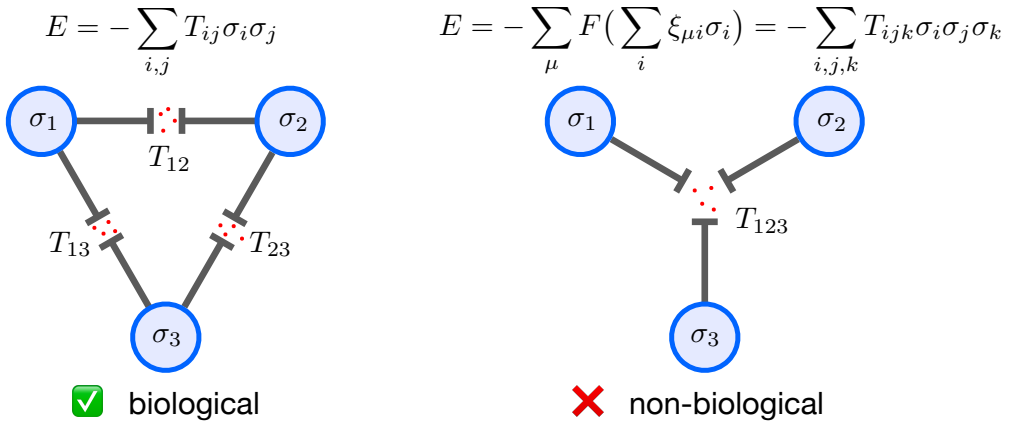

Figure 1: Two binary networks consisting of three neurons $\sigma_1, \sigma_2, \sigma_3 = \{\pm 1\}$. On the left is the classical Hopfield network (Hopfield, 1982) with the matrix $T_{ij} = \sum_\mu \xi_{\mu i}\xi_{\mu j}$ being the outer product of memory vectors (see section 2 for the definitions of notations). In this case the matrix T_{ij} is interpreted as a matrix of synaptic connections between cells i and j. On the right is a Dense Associative Memory network of (Krotov & Hopfield, 2016) with cubic interaction term $F(x) = x^3$. In this case the corresponding tensor $T_{ijk} = \sum_\mu \xi_{\mu i}\xi_{\mu j}\xi_{\mu k}$ has three indices, thus cannot be interpreted as a biological synapse, which can only connect two cells.

Memory (Krotov & Hopfield, 2016). This idea has been shown to dramatically increase the memory storage capacity of the corresponding neural network (Krotov & Hopfield, 2016; Demircigil et al., 2017) and was proposed to be useful for increasing robustness of neural networks to adversarial attacks (Krotov & Hopfield, 2018). Recently, an extension of this idea to continuous variables, called modern Hopfield network, demonstrated remarkably successful results on the immune repertoire classification (Widrich et al., 2020), and provided valuable insights into the properties of attention heads in Transformer architectures (Ramsauer et al., 2020).

Both Dense Associative Memories (Krotov & Hopfield, 2016) and modern Hopfield networks (Ramsauer et al., 2020), however, cannot describe biological neural networks in terms of true microscopic degrees of freedom, since they contain many-body interaction terms in equations describing their dynamics and the corresponding energy functions. To illustrate this point consider two networks: a conventional Hopfield network (Hopfield, 1982) and a Dense Associative Memory with cubic interaction term in the energy function (see Fig. 1). In the conventional network the dynamics is encoded in the matrix T_{ij}, which represents the strengths of the synaptic connections between feature neurons i and j. Thus, this network is manifestly describable in terms of only two-body synapses, which is approximately true for many biological synapses. In contrast, a Dense Associative Memory network with cubic energy function naively requires the synaptic connections to be tensors T_{ijk} with three indices, which are harder to implement biologically. Many-body synapses become even more problematic in situations when the interaction term is described by a more complicated function than a simple power.

Many-body synapses typically appear in situations when one starts with a microscopic theory described by only two-body synapses and integrates our some of the degrees of freedom (hidden neurons). The argument described above based on counting the information stored in synapses in conjunction with the fact that modern Hopfield nets and Dense Associative Memories can have a huge storage capacity hints at the same solution. The reason why these networks have a storage capacity much greater than N_f is because they do not describe the dynamics of only N_f neurons, but rather involve additional hidden neurons.

Thus, there remains a theoretical question: what does this hidden circuitry look like? Is it possible to introduce a set of hidden neurons with appropriately chosen interaction terms and activation functions so that the resulting theory has both large memory storage capacity (significantly bigger than N_f), and, at the same time, is manifestly describable in terms on only two-body synapses?

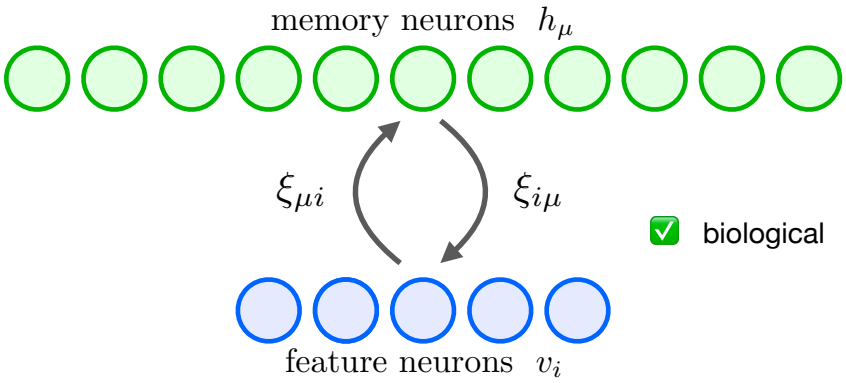

Figure 2: An example of a continuous network with $N_f = 5$ feature neurons and $N_h = 11$ complex memory (hidden) neurons with symmetric synaptic connections between them.

The main contributions of this current paper are the following. First, we extend the model of (Krotov & Hopfield, 2016) to continuous state variables and continuous time, so that the state of the network is described by a system of non-linear differential equations. Second, we couple an additional set of N_h "complex neurons" or "memory neurons" or hidden neurons to the N_f feature neurons. When the synaptic couplings and neuron activation functions are appropriately chosen, this dynamical system in $N_f + N_h$ variables has an energy function describing its dynamics. The minima (stable points) of this dynamics are at the same locations in N_f - dimensional feature subspace as the minima in the corresponding Dense Associative Memory system. Importantly, the resulting dynamical system has a mathematical structure of a conventional recurrent neural network, in which the neurons interact only in pairs through a two-body matrix of synaptic connections. We study three limiting cases of this new theory, which we call models A, B, and C. In one limit (model A) it reduces to Dense Associative Memory model of (Krotov & Hopfield, 2016) or (Demircigil et al., 2017) depending on the choice of the activation function. In another limit (model B) our model reduces to the network of (Ramsauer et al., 2020), which is equivalent to attention mechanism in Transformers. Finally we present a third limit (model C) which we call Spherical Memory model. To the best of our knowledge this model has not been studied in the literature. However, it has a high degree of symmetry and for this reason might be useful for future explorations of various models of large associative memory in machine learning.

2 Mathematical Formulation

In this section, we present a simple mathematical model in continuous time, which, on one hand, permits the storage of a huge number of patterns in the artificial neural network, and, at the same time, involves only pairwise interactions between the neurons through synaptic junctions. Thus, this system has the useful associative memory properties of the AI system, while maintaining conventional neural network dynamics and thus a degree of biological plausibility.

The spikes of action potentials in a pre-synaptic cell produce input currents into a postsynaptic neuron. As a result of a single spike in the pre-synaptic cell the current in the post-synaptic neuron rises instantaneously and then falls off exponentially with a time constant τ. In the following the currents of the feature neurons are denoted by v_i (which are enumerated by the latin indices), and the currents of the complex memory neurons are denoted by h_μ (h stands for hidden neurons, which are enumerated by the greek indices). A simple cartoon of the network that we discuss is shown in Fig.2. There are no synaptic connections among the feature neurons or the memory neurons. A matrix $\xi_{\mu i}$ denotes the strength of synapses from a feature neuron i to the memory neuron μ. The synapses are assumed to be symmetric, so that the same value $\xi_{i\mu} = \xi_{\mu i}$ characterizes a different physical synapse from the memory neuron μ to the feature neuron i. The outputs of the memory neurons and the feature neurons are denoted by f_μ and g_i, which are non-linear functions of the corresponding currents. In some situations (model A) these outputs can be interpreted as activation functions for the corresponding neurons, so that $f_\mu = f(h_\mu)$ and $g_i = g(v_i)$ with some non-linear functions $f(x)$

and $g(x)$. In other cases (models B and C) these outputs involve contrastive normalization, e.g. a softmax, and can depend on the currents of all the neurons in that layer. In these cases $f_\mu = f(\{h_\mu\})$ and $g_i = g(\{v_i\})$. For the most part of this paper one can think about them as firing rates of the corresponding neurons. In some limiting cases, however, the function $g(v_i)$ will have both positive and negative signs. Then it should be interpreted as the input current from a pre-synaptic neuron. The functions $f(h_\mu)$ and $g(v_i)$ are the only nonlinearities that appear in our model. Finally, the time constants for the two groups of neurons are denoted by τ_f and τ_h. With these notations our model can be written as

$$\begin{cases} \tau_f \frac{dv_i}{dt} = \sum\limits_{\mu=1}^{N_h} \xi_{i\mu} f_\mu - v_i + I_i \\ \tau_h \frac{dh_\mu}{dt} = \sum\limits_{i=1}^{N_f} \xi_{\mu i} g_i - h_\mu \end{cases} \tag{1}$$

where I_i denotes the input current into the feature neurons.

The connectivity of our network has the structure of a bipartite graph, so that the connections exist between two groups of neurons, but not within each of the two groups. This design of a neural network is inspired by the class of models called Restricted Boltzmann Machines (RBM) (Smolensky, 1986). There is a body of literature studying thermodynamic properties of these systems and learning rules for the synaptic weights. In contrast, the goal of our work is to write down a general dynamical system and an energy function so that the network has useful properties of associative memories with a large memory storage capacity, is describable only in terms of manifestly two-body synapses, and is sufficiently general so that it can be reduced to various models of this class previously discussed in the literature. We also note that although we use the notation v_i (v stands for visible neurons), commonly used in the RBM literature, it is more appropriate to think about v_i as higher level features. For example the input to our network can be a latent representation produced by a convolutional neural network or a latent representation of a BERT-like system rather than raw input data. Additionally, our general formulation makes it possible to use a much broader class of activation functions (e.g. involving contrastive or spherical normalization) than those typically used in the RBM literature. The relationship between Dense Associative Memories and RBMs have been studied in (Barra et al., 2018; Agliari & De Marzo, 2020).

It is possible to write down an energy function for the network (1), which is given by

$$E(t) = \Big[\sum_{i=1}^{N_f} (v_i - I_i) g_i - L_v \Big] + \Big[\sum_{\mu=1}^{N_h} h_\mu f_\mu - L_h \Big] - \sum_{\mu,i} f_\mu \xi_{\mu i} g_i \tag{2}$$

Here we introduced two Lagrangian functions $L_v(\{v_i\})$ and $L_h(\{h_\mu\})$ for the feature and the hidden neurons. They are defined through the following equations, so that derivatives of the Lagrangian functions correspond to the outputs of neurons

$$f_\mu = \frac{\partial L_h}{\partial h_\mu}, \quad \text{and} \quad g_i = \frac{\partial L_v}{\partial v_i} \tag{3}$$

With these notations expressions in the square brackets in (2) have a familiar from classical mechanics structure of the Legendre transform between a Lagrangian and an energy function. By taking time derivative of the energy and using dynamical equations (1) one can show (see appendix for details) that the energy monotonically decreases on the dynamical trajectory

$$\frac{dE(t)}{dt} = -\tau_f \sum_{i,j=1}^{N_f} \frac{dv_i}{dt} \frac{\partial^2 L_v}{\partial v_i \partial v_j} \frac{dv_j}{dt} - \tau_h \sum_{\mu,\nu=1}^{N_h} \frac{dh_\mu}{dt} \frac{\partial^2 L_h}{\partial h_\mu \partial h_\nu} \frac{dh_\nu}{dt} \leq 0 \tag{4}$$

The last inequality sign holds provided that the Hessian matrices of the Lagrangian functions are positive semi-definite.

In addition to decrease of the energy function on the dynamical trajectory it is important to check that for a specific choice of the activation functions (or Lagrangian functions) the corresponding energy is bounded from below. This can be achieved for example by using bounded activation function for the feature neurons $g(v_i)$, e.g. hyperbolic tangent or a sigmoid. Provided that the energy is bounded,

the dynamics of the neural network will eventually reach a fixed point, which corresponds to one of the local minima of the energy function[1].

The proposed energy function has three terms in it: the first term depends only on the feature neurons, the second term depends only on the hidden neurons, and the third term is the "interaction" term between the two groups of neurons. Note, that this third term is manifestly describable by two-body synapses - a function of the activity of the feature neurons is coupled to another function of the activity of the memory neurons, and the strength of this coupling is characterized by the parameters $\xi_{\mu i}$. The absence of many-body interaction terms in the energy function results in the conventional structure (with unconventional activation functions) of the dynamical equations (1). Each neuron collects outputs of other neurons, weights them with coefficients ξ and generates its own output. Thus, the network described by equations (1) is biologically plausible.

For the purposes of this paper we defined "biological plausiblity" as the absence of many-body synapses. It is important to note that there other aspects in which equations (1) are biologically implausible. For instance, this model assumes that the strengths of two physically different synapses $\mu \to i$ and $i \to \mu$ are equal. This assumption is necessary for the existence of the energy function, which makes it easy to prove the convergence to a fixed point. It can be relaxed in equations (1), which makes them even more biological, but, at the same time, more difficult to analyse.

Lastly, note that the memory patterns $\xi_{\mu i}$ of our network (1) can be interpreted as the strengths of the synapses connecting feature and memory neurons. This interpretation is different from the conventional interpretation, in which the strengths of the synapses is determined by matrices $T_{ij} = \sum_\mu \xi_{\mu i}\xi_{\mu j}$ (see Fig. 1), which are outer products of the memory vectors (or higher order generalizations of the outer products).

3 Effective Theory for Feature Neurons

In this section we start with the general theory proposed in the previous section and integrate out hidden neurons. We show that depending on the choice of the activation functions this general theory reduces to some of the models of associative memory previously studied in the literature, such as classical Hopfield networks, Dense Associative Memories, and modern Hopfield networks. The update rule in the latter case has the same mathematical structure as the attention mechanism in Transformers.

3.1 Model A. Dense Associative Memory Limit.

Consider the situation when the dynamics of memory neurons h_μ is fast. Mathematically this corresponds to the limit $\tau_h \to 0$. In this case the second equation in (1) equilibrates quickly, and can be solved as

$$h_\mu = \sum_{i=1}^{N_f} \xi_{\mu i} g_i \tag{5}$$

Additionally, assume that the Lagrangian functions for the feature and the memory neurons are additive for individual neurons

$$L_h = \sum_\mu F(h_\mu), \quad \text{and} \quad L_v = \sum_i G(v_i) \tag{6}$$

where $F(x)$ and $G(x)$ are some non-linear functions. In this limit we set $G(x) = |x|$. Since, the outputs of the feature neurons are derivatives of the Lagrangian (3), they are given by the sign functions of their currents,which gives a set of binary variables that are denoted by σ_i

$$\sigma_i = g_i = g(v_i) = \frac{\partial L_v}{\partial v_i} = Sign\big[v_i\big] \tag{7}$$

Since $G(v_i) = |v_i|$ the only term that survives in the first square bracket in equation (2) is the one proportional to the input current I_i. The first term in the second bracket of equation (2) cancels the

[1]There is also a border case possibility that the dynamics cycles without decreasing the energy (limit cycle), but this requires that the Hessian matrix in (4) has a zero mode everywhere along the trajectory. This border case possibility should be checked for a specific choice of the activation functions.

interaction term because of the steady state condition (5). Thus, in this limit the energy function (2) reduces to

$$E(t) = -\sum_{i=1}^{N_f} I_i \sigma_i - \sum_{\mu=1}^{N_h} F\Big(\sum_i \xi_{\mu i}\sigma_i\Big) \tag{8}$$

If there are no input currents $I_i = 0$ this is exactly the energy function for Dense Associative Memory from (Krotov & Hopfield, 2016). If $F(x) = x^n$ is a power function, the network can store $N_{\text{mem}} \sim N_f^{n-1}$ memories, if $F(x) = \exp(x)$ the network has exponential storage capacity Demircigil et al. (2017). If power $n = 2$ this model further reduces to the classical Hopfield network (Hopfield, 1982).

Lastly, for the class of additive models (6), which we call models A, the equation for the temporal evolution of the energy function reduces to

$$\frac{dE(t)}{dt} = -\tau_f \sum_{i=1}^{N_f} \Big(\frac{dv_i}{dt}\Big)^2 g(v_i)' - \tau_h \sum_{\mu=1}^{N_h} \Big(\frac{dh_\mu}{dt}\Big)^2 f(h_\mu)' \leq 0 \tag{9}$$

Thus, the condition that the Hessians are positive definite is equivalent to the condition that the activation functions $g(v_i)$ and $f(h_\mu)$ are monotonically increasing.

3.2 Model B. Modern Hopfield Networks Limit and Attention of Transformers.

Models B are defined as models having contrastive normalization in the hidden layer. Specifically we are interested in

$$L_h = \log\Big(\sum_\mu e^{h_\mu}\Big), \quad \text{and} \quad L_v = \frac{1}{2}\sum_i v_i^2 \tag{10}$$

so that L_v is still additive, but L_h is not. Using the general definition of the activation functions (3) one obtains

$$\begin{aligned} f_\mu &= \frac{\partial L_h}{\partial h_\mu} = \text{softmax}(h_\mu) = \frac{e^{h_\mu}}{\sum\limits_\nu e^{h_\nu}} \\ g_i &= \frac{\partial L_v}{\partial v_i} = v_i \end{aligned} \tag{11}$$

Similarly to the previous case, consider the limit $\tau_h \to 0$, so that equation (5) is satisfied. In this limit the energy function (2) reduces to (currents I_i are assumed to be zero)

$$E = \frac{1}{2}\sum_{i=1}^{N_f} v_i^2 - \log\Big(\sum_\mu \exp(\sum_i \xi_{\mu i} v_i)\Big) \tag{12}$$

This is exactly the energy function studied in (Ramsauer et al., 2020) up to additive constants (inverse temperature β was assumed to be equal to one in this derivation). Notice that we used the notations from (Krotov & Hopfield, 2016), which are different from the notations of (Ramsauer et al., 2020). In the latter paper the state vector v_i is denoted by ξ_i and the memory matrix $\xi_{\mu i}$ is denoted by the matrix $\mathbf{X^T}$.

Making substitutions (11) in the first equation of (1), using steady state condition (5), and setting input current $I_i = 0$ results in the following effective equations for the feature neurons, when the memory neurons are integrated out

$$\tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu} \text{softmax}\Big(\sum_{j=1}^{N_f} \xi_{\mu j} v_j\Big) - v_i \tag{13}$$

This is a continuous time counterpart of the update rule of (Ramsauer et al., 2020). Writing it in finite differences gives

$$v_i^{(t+1)} = v_i^{(t)} + \frac{dt}{\tau_f}\Big[\sum_{\mu=1}^{N_h} \xi_{i\mu} \text{softmax}\Big(\sum_{j=1}^{N_f} \xi_{\mu j} v_j^{(t)}\Big) - v_i^{(t)}\Big] \tag{14}$$

which for $dt = \tau_f$ reduces to

$$v_i^{(t+1)} = \sum_{\mu=1}^{N_h} \xi_{i\mu} \text{softmax}\Big(\sum_{j=1}^{N_f} \xi_{\mu j} v_j^{(t)} \Big) \tag{15}$$

This is exactly the update rule derived in (Ramsauer et al., 2020), which, if applied once, is equivalent to the familiar attention mechanism in Transformer networks.

The derivation of this result in (Ramsauer et al., 2020) begins with the energy function for a DAM model with exponential interactions $F(x) = exp(x)$. Then it is proposed to take a logarithm of that energy (with a minus sign) and add a quadratic term in the state vector v_i to ensure that it remains finite and the energy is bounded from below. While this is a possible logic, it requires a heuristic step - taking the logarithm, and makes the connection with Dense Associative Memories less transparent. In contrast, our derivation follows from the general principles specified by equations (1,2) for the specifically chosen Lagrangians.

It is also important to note, that the Hessian matrix for the hidden neurons has a zero mode (zero eigenvalue) for this limit of our model.

3.3 Model C. Spherical Memory.

Models C are defined as having spherical normalization in the feature layer. We are not aware of discussion of this class of models in the literature. Specifically,

$$L_h = \sum_\mu F(h_\mu), \quad \text{and} \quad L_v = \sqrt{\sum_i v_i^2} \tag{16}$$

so that L_h is additive, but L_v is not. Using the general definition of the activation functions (3) one obtains

$$\begin{aligned} f_\mu &= F'(h_\mu) \\ g_i &= \frac{\partial L_v}{\partial v_i} = \frac{v_i}{\sqrt{\sum_j v_j^2}} \end{aligned} \tag{17}$$

Legendre transform of L_v in this case vanishes. For this reason, the dynamical equations for feature neurons will not have a decay term. Equations (1) for model C are given by (I_i is assumed to be zero)

$$\begin{cases} \tau_f \frac{dv_i}{dt} = \sum\limits_{\mu=1}^{N_h} \xi_{i\mu} f(h_\mu) \\ \tau_h \frac{dh_\mu}{dt} = \sum\limits_{i=1}^{N_f} \xi_{\mu i} g_i - h_\mu \end{cases} \tag{18}$$

Taking the limit $\tau_h \to 0$ and excluding h_μ gives the effective energy

$$E(t) = -\sum_\mu F\Big(\sum_i \xi_{\mu i} \frac{v_i}{\sqrt{\sum_j v_j^2}} \Big) \tag{19}$$

and the corresponding effective dynamical equations

$$\tau_f \frac{dv_i}{dt} = \sum_\mu \xi_{i\mu} f\Big[\sum_j \xi_{\mu j} \frac{v_j}{\sqrt{\sum_k v_k^2}} \Big] \tag{20}$$

The Hessian matrix for the feature neurons has a zero mode (zero eigenvalue) for this limit of our model.

4 Discussion and Conclusions

We have proposed a general dynamical system and an energy function that has a large memory storage capacity, and, at the same time, is manifestly describable in terms of two-body synaptic

connections. From the perspective of neuroscience it suggests that Dense Associative Memory models are not just mathematical tools useful in AI, but have a degree of biological plausibility similar to that of the conventional continuous Hopfield networks (Hopfield, 1984). Compared to the latter, these models have a greater degree of psychological plausibility, since they can store the much larger number of memories that is necessary to explain memory-based animal behavior.

From the perspective of AI research our paper provides a conceptually grounded derivation of various associative memory models discussed in the literature, and relationships between them. We hope that the more general formulation, presented in this work, will assist in the development of new models of this class that could be used as building components of new recurrent neural network architectures.

APPENDIX

In this appendix we show a step by step derivation of the change of the energy function (2) under dynamics (1). Time derivative of the energy function can be expressed through time derivatives of the neuron's activities v_i and h_μ (the input current I_i is assumed to be time-independent in the calculation below). Using the definition of the functions f_μ and g_i in (3) one can obtain

$$\begin{aligned}
\frac{dE}{dt} &= \sum_{i,j} \big(v_i - I_i\big) \frac{\partial^2 L_v}{\partial v_i \partial v_j} \frac{dv_j}{dt} + \sum_{\mu,\nu} h_\mu \frac{\partial^2 L_h}{\partial h_\mu \partial h_\nu} \frac{dh_\nu}{dt} \\
&- \sum_{\mu,\nu} \frac{dh_\nu}{dt} \frac{\partial^2 L_h}{\partial h_\nu \partial h_\mu} \Big(\sum_i \xi_{\mu i} g_i \Big) - \sum_{i,j} \frac{dv_j}{dt} \frac{\partial^2 L_v}{\partial v_j \partial v_i} \Big(\sum_\mu \xi_{i\mu} f_\mu \Big) = \\
&- \sum_{i,j} \frac{dv_j}{dt} \frac{\partial^2 L_v}{\partial v_j \partial v_i} \Big[\sum_\mu \xi_{i\mu} f_\mu + I_i - v_i \Big] - \sum_{\mu,\nu} \frac{dh_\nu}{dt} \frac{\partial^2 L_h}{\partial h_\nu \partial h_\mu} \Big[\sum_i \xi_{\mu i} g_i - h_\mu \Big] = \\
&- \tau_f \sum_{i,j=1}^{N_f} \frac{dv_i}{dt} \frac{\partial^2 L_v}{\partial v_i \partial v_j} \frac{dv_j}{dt} - \tau_h \sum_{\mu,\nu=1}^{N_h} \frac{dh_\mu}{dt} \frac{\partial^2 L_h}{\partial h_\mu \partial h_\nu} \frac{dh_\nu}{dt} \leq 0
\end{aligned} \tag{21}$$

In the last equality sign the right hand sides of dynamical equations (1) are used to replace expressions in the square brackets by the corresponding time derivatives of the neuron's activities. This completes the proof that the energy function decreases on the dynamical trajectory described by equations (1) for arbitrary time constants τ_f and τ_h provided that the Hessians for feature and memory neurons are positive semi-definite.

REFERENCES

Elena Agliari and Giordano De Marzo. Tolerance versus synaptic noise in dense associative memories. *arXiv preprint arXiv:2007.02849*, 2020.

Adriano Barra, Matteo Beccaria, and Alberto Fachechi. A new mechanical approach to handle generalized hopfield neural networks. *Neural Networks*, 106:205–222, 2018.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural computation*, 30(12):3151–3167, 2018.

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

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

Michael Widrich, Bernhard Schäfl, Hubert Ramsauer, Milena Pavlović, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, et al. Modern hopfield networks and attention for immune repertoire classification. *arXiv preprint arXiv:2007.13505*, 2020.