Linking Restricted Boltzmann Machines to Generalized Hopfield Models for Pattern Recognition

Vedant Sachdeva and Anirvan Sengupta*
Department of Physics and Astronomy,
Rutgers University - New Brunswick

(Dated: April 8, 2017)

Restricted Boltzmann Machines (RBMs) are a deep learning architecture used widely for natural language processing, computer vision, and feature detection but they are not well understood theoretically. We seek to compare RBMs to associative memory networks and use Mean-Field Theory to make predictions about the transitions present in RBMs. We start by presenting a set of self-consistent equations for the Generalized Hopfield Model, and comparing the results to that of a binary RBM with binary synapses. We then compare the features detected by a dense associative memory network to that of a binary Restricted Boltzmann Machine and compare qualitatively. Similarities between the types of features detected suggest the presence of a similar mechanism for learning, meaning further study of dense associative memory nets could lead to a better understanding of RBMs.

I. TECHNICAL SUMMARY

Restricted Boltzmann Machines are neural network models used to algorithmically make complex decisions that involve tasks such as pattern recognition, image classificaiton, or natural language processing. Despite their popularity in both academia and industry, little is understood about how they are actually successful at performing complicated tasks, although there has been some effort to understand them using RG Theory. However, there has been some study on Boltzmann Machines that has suggested the presence of associative memory network-like qualities, suggesting an avenue for analytic study of RBMs using classical statistical mechanics.

By developing mean-field theory equations for associative memory networks in the limit of the degree approaching 1, we were able to develop the hypothesis that binary RBMs with binary synapses can only recognize a finite number of patterns, regardless of the size of the RBM. We then illustrate that there is a similarity between the feature detection done by Dense Associative Memory (DAM) Networks and the feature detection done by RBMs by considering the dot product of the image vector of 10 most important features for a given classification task of both models. This suggests that further analytic study of RBMs could be done through use of DAM networks.

II. INTRODUCTION

Boltzmann Machines are stochastically updating probabilistic graphical models that can be used to represent neural networks. They can be described as a complete symmetrically weighted graph with $N + P$ binary units. There are N visible, or observed, units, which are denoted as v_i where $i \in \{1 \dots N\}$ and P hidden units, which are denoted as h_μ, where $\mu \in \{1 \dots P\}$. Each edge between hidden units and visible has weight according to a synaptic matrix, $F_{i\mu}$, of size $N \times P$ [4]. Each edge between two visible units has weight according to W_{ij}, which is of size of $N \times N$. Each edge between two hidden units has weight according to $J_{\mu\nu}$, which is of size $P \times P$. Through specific choice of this synaptic matrix, Boltzmann Machines can recover meaningful patterns, allowing it to play roles in feature detection[6]. However, Boltzmann Machines cannot be easily scaled up, due to their computational complexity, and are subject to becoming trapped in spurious states. As such, we instead consider Restricted Boltzmann Machines (RBMs). RBMs are a subset of Boltzmann Machines for whom $J_{\mu\nu}$=W_{ij}=0.

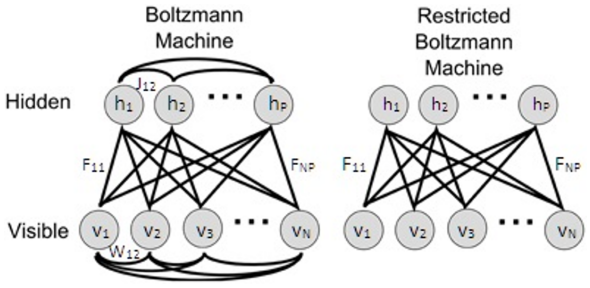

FIG. 1. Typical Architecture for a two-layer Restricted Boltzmann Machine [10]

Due to the binary nature of each unit in an RBM, it can be demonstrated that RBMs are isomorphic to Ising spin-glasses, and thus can be analyzed using the statistical mechanical tools used to analyze spin-glasses as has been done with Hopfield Networks[1]. This is accomplished by defining an energy function for Restricted Boltzmann

* anirvan@physics.rutgers.edu; Center for Quantitative Biology, Rutgers University - New Brunswick

Machines[7]:

$$E(\vec{v},\vec{h}) = -\sum_{i=1}^{N} b_i v_i - \sum_{\mu=1}^{P} c_\mu h_\mu - \sum_{i=1,\mu=1}^{N,P} v_i F_{i\mu} h_\mu \quad (1)$$

This allows us to develop a partition function as follows:

$$Z = \sum_{\vec{v},\vec{h}} e^{-\beta E(\vec{v},\vec{h})} \quad (2)$$

In past studies, the limit where hidden units are analog and the visible units are digital has been studied. In the study, the exchange interaction was taken to be a Bernoulli random variable. It was suggested that hidden units evolved according to an Ornstein-Uhlenbeck diffusion process and visible units evolved according to a sigmoidal rule[3]. When the marginal distribution for the visible units is written according to these rules, we recover a distribution that closely resembles that of the Hopfield Model, giving evidence for the existence of a Hopfield-like transition in Restricted Boltzmann Machines that arises from the ratio of hidden units to visible units[3]. This, paired with the fact that RBMs evolve to a local minimum, suggests that RBMs can be understood by further consideration of the Hopfield Model. The Hopfield model is a Sherrington-Kirkpatrick spin model with exchange interactions written according to a Hebbian learning rule[8]. The Hamiltonian for this model is written as follows:

$$E(\vec{S}) = -N \sum_{\mu=1}^{P} \left(\frac{1}{N} \sum_{i=1} N \xi_i^\mu S_i \right)^2 \quad (3)$$

Through the replica trick, it was determined that this model underwent a phase transition that took the spin-glass from a ferromagnetic state, where memory recovery was possible, to a glassy state, where the number of spurious patterns overwhelms the target memory, making recovery unlikely[2]. This model is the $n = 2$ case of a generalized power-n model studied by Gardner in 1987; in her work, she shows the phase transition of a model with the following Hamiltonian[5]:

$$E(\vec{S}) = -N \sum_{\mu=1}^{P} \left(\frac{1}{N} \sum_{i=1} N \xi_i^\mu S_i \right)^n \quad (4)$$

in the high-P, high-n limit. For $n > 2$ It is observed that the phase transition occurs when the number of memories is $O(N^{n-1})$. We note that this agrees with the phase transition for the standard Hopfield Model for P being $O(N)$.

III. RESULTS AND ANALYSIS

A. Linking Restricted Boltzmann Machines with random binary synapses to Generalized Hopfield Models

We consider the Generalized Hopfield Model because in the low temperature limit, the Restricted Boltzmann Machine closely resembles the Restricted Boltzmann Machine if the visible units and hidden units are binary ± 1 variables. If we further assume that bias terms are 0, this allows us to marginalize the partition function over the hidden units, resulting in the following relation:

$$Z \propto \prod_{\mu=1}^{P} \sum_{\vec{v}} \exp\left(\log\left(\cosh\left(\beta \sum_{i=1}^{N} v_i F_{i\mu} \right) \right) \right) \quad (5)$$

This can be observed by considering the behavior of the hyperbolic cosine function at low temperature. In the low temperature limit, we see that if $\sum_{i=1}^{N} v_i F_{i\mu}$ is positive, e^x dominates and if $\sum_{i=1}^{N} v_i F_{i\mu}$ is negative, e^{-x} dominates. So, when we take the logarithm of the hyperbolic cosine function, we are left with $|\beta \sum_{i=1}^{N} v_i F_{i\mu}|$ which gives us a n=1 model. Thus, we begin by studying the transitions of the Generalized Hopfield Model for various n between 1 and 2. We start with a mean-field theory calculation for the generalized Hopfield Model to obtain a set of self-consistent equations. The derivation of these equations follows from the derivation of the Gardner equations.

B. Mean-field equations for the Gardner Model in the $n \gg 2$ condition

We begin our analysis with an alternative derivation of the Gardner self-consistent equations using MeanField Theory.

In order to derive the self-consistent equations for the Generalized Hopfield Model using mean-field theory, we start by considering the fact that in Associative Memory Networks, we tend to see that the spins in the network will have a finite alignment with a finite number of the memories, while the remaining memories will have $O(\frac{1}{\sqrt{N}})$in the large spin limit and large memory limit. Without loss of generality, we will choose the $\mu = 1$ to be the recovered memory. As a result, we can treat the problem like a convex optimization problem. As such, we begin by doing a Legendre transformation on the Hamiltonian.

$$\begin{aligned} &-N \sum_{\mu=1}^{P} \left(\frac{1}{N} \sum_{i=1} N \xi_i^\mu S_i \right)^n = \\ &\min_{z_\mu} -N \sum_{\mu=1}^{P} \left(\frac{z_\mu}{N} \sum_{i=1}^{N} \xi_i^\mu S_i - \left(\frac{z_\mu}{n} \right)^{\frac{n}{n-1}} \right) \end{aligned} \quad (6)$$

We now replace the n^{th} power model by this extended model with auxiliary variables z_μ. They should have the same low-temperature properties, and at $n = 2$, this replacement is exact. We now should calculate the expected value for each z_ν.

$$< z_\nu >= \frac{\sum_{\{S\}} \int \prod_{\mu=1}^{P} dz_\mu z_\nu e^{-N \sum_{\mu=1}^{P} \left(\frac{z_\mu}{N} \sum_{i=1}^{N} \xi_i^\mu S_i - \left(\frac{z_\mu}{n}\right)^{\frac{n}{n-1}} \right)}}{Z} \tag{7}$$

We can sum over all spins in the above equation and the resultant integral can be then calculated through the use of a saddle-point integral. We note that calculating $< z_\nu >$ through the saddle-point method is equivalent to extremizing the exponent. This results in the following expression:

$$< z_\nu >= n^{\frac{1}{n-1}} \Big(\frac{n-1}{Nn}\Big) \sum_{i=1}^{N} \xi_i^\nu \tanh\Big(\beta \sum_{\mu=1}^{P} \xi_i^\mu < z_\mu > \Big) \tag{8}$$

Through some algebraic manipulation using the properties of the hyperbolic tangent function and the fact that ξ_i^μ is ± 1, we can obtain the following two key expressions:

$$\begin{aligned} &< z_1 > \\ &= n^{\frac{1}{n-1}} \Big(\frac{n-1}{Nn}\Big) \sum_{i=1}^{N} \tanh\Big(\beta < z_1 > \\ &+ \beta \xi_i^1 \sum_{\mu \neq 1}^{P} \xi_i^\mu < z_\mu > \Big) \end{aligned} \tag{9}$$

$$\begin{aligned} &< z_\nu > \\ &= n^{\frac{1}{n-1}} \Big(\frac{n-1}{Nn}\Big) \sum_{i=1}^{N} \xi_i^\nu \xi_i^1 \tanh\Big(\beta < z_1 > \\ &+ \beta \xi_i^1 \xi_i^\nu < z_\nu > + \beta \xi_i^1 \sum_{\mu \neq 1,\nu}^{P} \xi_i^\mu < z_\mu > \Big) \end{aligned} \tag{10}$$

We can then perform a Taylor expansion around $\beta < z_1 > + \beta \xi_i^1 \sum_i^\nu < z_\mu >$, which is justified because we argue that since $< z_\nu >$ is not a target pattern, it will be $O(\frac{1}{\sqrt{N}})$. This results in the following equation:

$$\begin{aligned} &< z_\nu >^{\frac{1}{n-1}} \\ &= n^{\frac{1}{n-1}} \frac{n-1}{Nn} \sum_{i=1}^{N} \xi_i^\nu \xi_i^\nu \tanh\Big(\beta(< z_1 > + \xi_i^1 \sum_{\mu \neq 1,\nu} \xi_i^\mu < z_\mu >)\Big) \\ &+ n^{\frac{1}{n-1}} \frac{n-1}{Nn} \sum_{i=1}^{N} \beta < z_\nu > \\ &- n^{\frac{1}{n-1}} \frac{n-1}{Nn} \sum_{i=1}^{N} \beta < z_\nu > \tanh^2(\beta < z_1 > + \beta \xi_i^1 \sum_{\mu \neq 1,\nu} \xi_i^\mu < z_\mu >) \end{aligned} \tag{11}$$

We re-express this equation with an order parameter, q, which is intended to resemble the replica order parameter.

$$\begin{aligned} &< z_\nu >^{\frac{1}{n-1}} \\ &= n^{\frac{1}{n-1}} \frac{n-1}{Nn} \sum_{i=1}^{N} \xi_i^\nu \xi_i^\nu \tanh(\beta < z_1 > + \beta \xi_i^1 \sum_{\mu \neq 1,\nu} \xi_i^\mu < z_\mu >) \\ &+ n^{\frac{1}{n-1}} \frac{n-1}{Nn} \sum_{i=1}^{N} \beta < z_\nu > (1-q) \end{aligned} \tag{12}$$

$$q = \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big) \tanh^2\Big(\beta(z_1 + \sqrt{\alpha r} x)\Big) \tag{13}$$

Here, we have chosen to write $\xi_i^1 \sum_{\mu \neq 1,\nu} x i_1^\mu < z_\mu >$ as a Gaussian random variable with 0 mean and αr variance. This allows us to define r in the following way:

$$r = \frac{1}{\alpha} \sum_{\mu \neq 1} < z_\mu >^2 \tag{14}$$

From this definition, we can use equation 11 to develop an equation for r in terms of z_1 and q.

$$\begin{aligned} &< z_\nu >^{\frac{1}{n-1}} \\ &- \frac{n^{\frac{1}{n-1}}(n-1)}{n} \beta < z_\nu > (1-q) \\ &= n^{\frac{1}{n-1}} \frac{n-1}{Nn} \sum_{i=1}^{N} \xi_i^\nu \xi_i^1 \tanh(\beta < z_1 > + \beta \xi_i^1 \sum_{\mu \neq 1,\nu} \xi_i^\mu < z_\mu >) \end{aligned} \tag{15}$$

Since $< z_\nu >$ for $\nu \neq 1$ is $O(\frac{1}{n-1})$, and $n \gg 1$, we expect $< z_\nu >^{\frac{1}{n-1}}$ to dominate $< z_\nu >$, so we can obtain the following equation.

$$\begin{aligned} &< z_\nu >^{\frac{2}{n-1}} = \\ &\Big(n^{\frac{1}{n-1}} \frac{n-1}{Nn}\Big)^2 \sum_{i=1}^{N} \tanh^2\Big(\beta(< z_1 > + \beta \sum_{\mu \neq 1,\nu} \xi_i^\mu < z_\mu >)\Big) \end{aligned} \tag{16}$$

This yields the following expression for r.

$$r = N\Big(n^{\frac{1}{n-1}}\frac{n-1}{Nn}\Big)^{2(n-1)} q^{n-1} \tag{17}$$

By setting $< z_1 >= m_1^{n-1}$, we see that equations 8, 12, and 16 form a set of self-consistent equations that is equivalent to the equations presented by Gardner without the replica symmetric ansatz[5].

C. Mean-field equations for the Gardner Model in the limit as $n \to 1$ condition

In the $n \to 1$ condition, we suggest that instead of the $< z_\nu >^{\frac{1}{n-1}}$ term dominating, the $\beta < z_\nu > (1-q)$ term dominates, resulting in a different equation for r, while the equations for q and $< z_1 >$ remain the same. As such, we can solve for r.

$$< z_\nu > (q-1) =$$
$$\frac{1}{N\beta}\sum_{i=1}^{N}\xi_i^\nu\xi_i^1 \tanh\Big(\beta\Big(< z_1 > +\xi_i^1 \sum_{\mu\neq 1,\nu}\xi_i^\mu < z_\mu >\Big)\Big) \tag{18}$$

$$< z_\nu >^2 (q-1)^2 = \frac{q}{N^2\beta^2} \tag{19}$$

$$r = \frac{q}{N\beta^2(q-1)^2} \tag{20}$$

Equation 20, along with equations 8 and 12 form a set of self-consistent equations for this limit. While we don't present a numerical solution to these three equations, we provide a qualitative argument for the presence of a phase transition from a ferromagnetic phase to a glassy phase. As earlier indicated, the variance of our Gaussian random variable is αr which is of $O(P * \frac{1}{N^{n-1}})$. We expect this to match the order of the target pattern when $P \sim O(N^{n-1})$. In the $n \to 1$ limit, this implies that $P \sim O(1)$, indicating that there can only be a finite number of patterns, regardless of system size.
We consider this conclusion numerically through a metropolis sampling algorithm over 100 time-steps at $\beta = 100$ with systems that ranged from size$N = 100-500$ for a variety of degrees. We see that as the degree approaches 1, a system of any size does not successfully recover the first memory after being perturbed from said memory. We compare this result to a numerical result for with a varying number of hidden units and found that RBMs failed to successfully recover patterns if there were more than 2 hidden units.

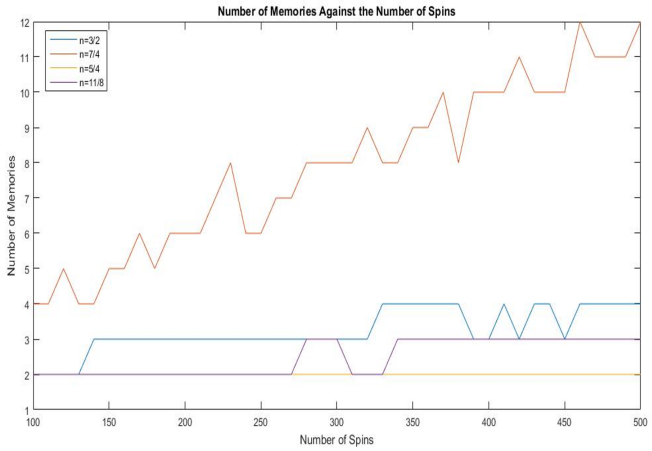

FIG. 2. At $\beta = 100$, with 100 time steps, we evolved a Generalized Hopfield Net using Metropolis sampling to determine the presence of a Hopfield like transition. We note that at low n, we see the transition cannot be seen with the size of the model.

This agrees with the result for an attempt to numerically evolve a Binary RBM with binary synapses, further suggesting the presence of Hopfield like transitions.

D. Extending the comparison to the Dense Associative Memory Network

However, we note that in reality, the synapses of RBMS are not typically binary. As such, we seek to train Hopfield Networks with continuous random variables as the synaptic matrix. In order to do this, we use the techniques for dense associative memory networks developed by Krotov and Hopfield[9]. We used their technique to attempt to do pattern recognition on a set of handwritten digits from the MNIST data set. After training, classification on a test set was 97% accurate with a $n = 2$ activation function. By using the learned patterns as intensity graphs, we were able to plot the types of memories such a network would use to classify the digits. The resultant plots are the features of the digit, which, when activated, help classify the digits, but are not individually useful. These features are learned by the synaptic matrix that represents the Hopfield interactions in a supervised manner and are used by the classifying synaptic matrix to classify images that were unseen by the network.

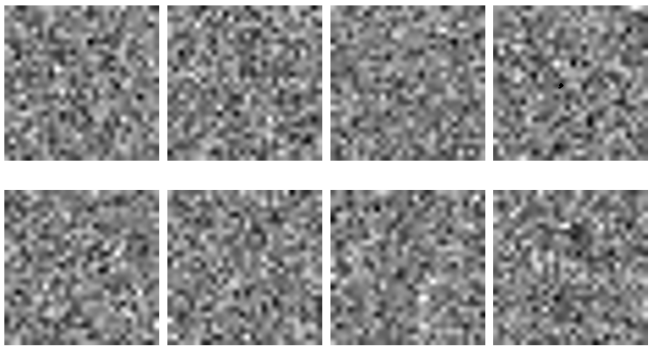

FIG. 3. 8 random features from a trained dense associative memory network with n=2 and 2000 memories.

Krotov and Hopfield note that when n¿20, there is a transition from feature-based recognition to prototype-based recognition. The difference between the prototypes and features are that prototypes represent blurred abstractions of the whole image, allowing each image to stand alone for classifying purposes. Here, we consider what happens when n approaches 1.

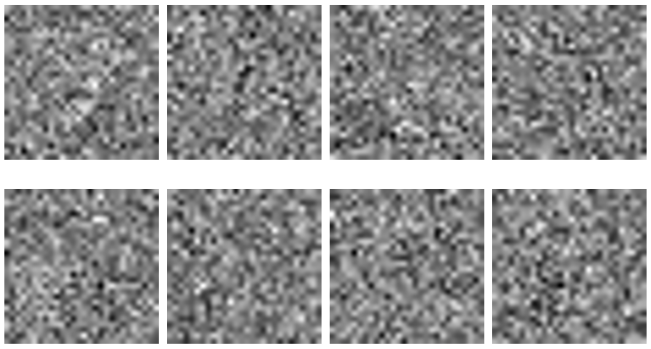

FIG. 4. 8 random features from a trained dense associative memory network with n=1.7 and 2000 memories.

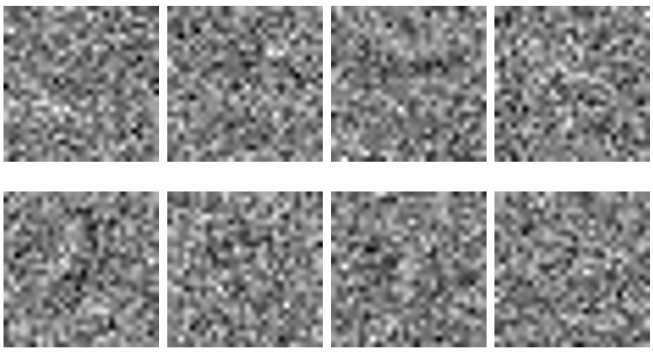

FIG. 5. 8 random features from a trained dense associative memory network with n=1.5 and 2000 memories.

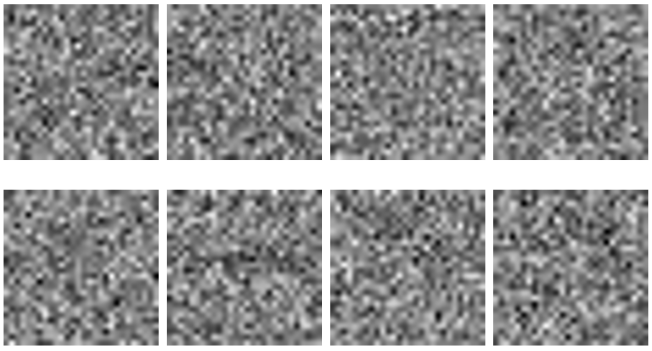

FIG. 6. 8 random features from a trained dense associative memory network with n=1.4 and 2000 memories.

At n lower than n=1.4, the dense associative memory network fails to classify to MNIST digits with a meaningful degree of accuracy. We now compare features detected by the dense associative memory network to the features detected by a binary RBM with continuous exchange interactions[11]. We plot the features detected by an RBM with 2000 hidden units and 784 visible units.

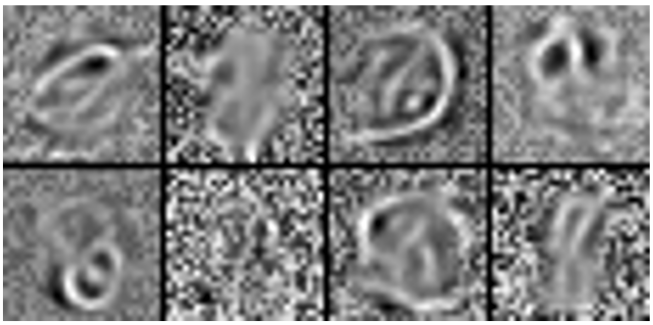

FIG. 7. 8 random features from a trained RBM with 2000 hidden units and 784 visible units.

We note that these are feature-like, as is the case with dense associative memory networks for $1 < n < 20$. However, when there are fewer hidden features, we see a prototype detection phase.

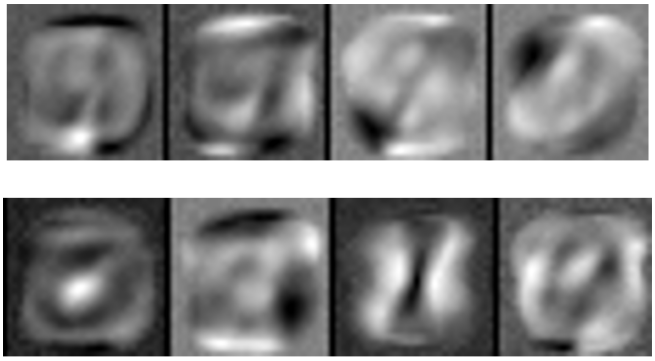

FIG. 8. 8 random features from a trained RBM with 10 hidden units and 784 visible units.

Ultimately, this suggests that the feature detection phase of Restricted Boltzmann Machines is analogous to the $1 < n < 20$ phase of dense associative memory networks.

We compare the angle between the image vector generated by overlaying the 10 most important features for classifying the various digits from the RBM to the dense associative memory network of varying degrees.

Digits	Degree=2	Degree=1.7	Degree=1.5	Degree=1.4
0	1.0388	1.0450	1.0371	1.0463
1	0.1886	0.1945	0.1937	0.1937
2	0.2655	0.2610	0.2620	0.2471
3	0.4485	0.4381	0.4446	0.4321
4	0.6360	0.6479	0.6456	0.6387
5	0.6427	0.6392	0.6357	0.6172
6	0.7069	0.7132	0.7136	0.6978
7	0.7803	0.7766	0.7873	0.7697
8	0.3282	0.3243	0.3225	0.3193
9	0.2877	0.2807	0.2895	0.2831

IV. DISCUSSION

In this work, we were able to show that Generalized Hopfield Networks demonstrate some similarity to binary RBMs. Using these similarities, we develop a mean-field theory that suggests the presence of a vanishing phase transition in Generalized Hopfield Networks that should be observable in RBMs in the low-temperature limit. This phase transition was observed for neural networks of size up to 500. This behavior is consistent with binary RBMs that have binary synapses, which suggest being in a glassy phase for any more than 2 hidden units. Further increases in the size of the network or the size of the visible partition could reveal that the phase transition persists all the way down to n=1, but this avenue was not pursued due to computational concerns. The findings from this study, while illuminating, are not useful for understanding true RBMs, as true RBMs have continuous synaptic matrices. As such, we turn to dense associative memory networks for further study.

Using dense associative memory networks with a rectified polynomial activation function, we continue to pursue our study of RBMs in the low temperature limit. By simulating dense associative memory networks with the rectified polynomial function in the limit that the polynomial degree approaches 1, we are able to detect features and plot those features. We note that the angle between the features of the Dense Associative Memory Network and the features of the RBM decreases as degree decreases. In addition, at high degree, dense associative memory networks learn best at high temperature, but studying the low temperature behavior of dense associative memory networks may allow for further analytic study of RBMs.

[1] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Phys. Rev. A*, 32:1007–1018, Aug 1985.

[2] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985.

[3] Adriano Barra, Alberto Bernacchia, Enrica Santucci, and Pierluigi Contucci. On the equivalence of hopfield networks and boltzmann machines. *Neural Networks*, 34:1 – 9, 2012.

[4] Asja Fischer and Christian Igel. *An Introduction to Restricted Boltzmann Machines*, pages 14–36. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

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

[6] G. E. Hinton and T. J. Sejnowski. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning and Relearning in Boltzmann Machines, pages 282–317. MIT Press, Cambridge, MA, USA, 1986.

[7] Geoffrey E. Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*, pages 599–619. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[8] 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.

[9] Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. *CoRR*, abs/1606.01164, 2016.

[10] Peter O'Connor, Daniel Neil, Shih-Chii Liu, Tobi Delbruck, and Michael Pfeiffer. Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in Neuroscience*, 7:178, 2013.

[11] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.