High storage capacity in the Hopfield model with auto-interactions—stability analysis

Jacopo Rocchi1, David Saad1 and Daniele Tantari2

1 Nonlinearity and Complexity Research Group, Aston University, Birmingham B4 7ET, United Kingdom
2 Scuola Normale Superiore, Centro Ennio de Giorgi, Piazza dei Cavalieri 3, I-56100, Pisa, Italy

E-mail: j.rocchi@aston.ac.uk, d.saad@aston.ac.uk and daniele.tantari@sns.it

Received 5 June 2017, revised 19 September 2017
Accepted for publication 29 September 2017
Published 23 October 2017

Abstract
Recent studies point to the potential storage of a large number of patterns in the celebrated Hopfield associative memory model, well beyond the limits obtained previously. We investigate the properties of new fixed points to discover that they exhibit instabilities for small perturbations and are therefore of limited value as associative memories. Moreover, a large deviations approach also shows that errors introduced to the original patterns induce additional errors and increased corruption with respect to the stored patterns.

Keywords: recurrent neural networks, Hopfield model, retrieval ability, capacity

(Some figures may appear in colour only in the online journal)

1. Introduction
Hopfield models [1–3] are recurrent neural networks where connections between units form a fully connected symmetric network. They have been proposed as models of content addressable memories, i.e. systems that are able to retrieve memory items from partial information. Their introduction was inspired by the observation that in large physical systems, interactions between the elementary degrees of freedom generate collective phenomena, such low temperature magnetisation in Ising models. Following this idea, the stability of memories in systems of interacting neurons has been successfully described as an emergent property, instigated by the dynamics of neural network models [1, 2].

Any physical system whose dynamics is dominated by a number of locally stable states can act as a content addressable memory as long as these states can be controlled. The Hebbian...
rule [4] has played an important role in training the couplings between neurons such that a prescribed set of memories (binary configurations of the Hopfield model) become attractors of the dynamics. Hopfield pointed out [1] that the issue of pattern retrieval is non trivial and that retrieval performance falls rapidly as more patterns are introduced, and are incorporated in the couplings. This behaviour was first found in numerical simulations and then analysed utilising statistical mechanics methods and exploiting the analogy with spin glass models. For is the number of patterns, is the number of neurons and , it has been found [5, 6] that the critical value below which recovery is possible is approximately . Original studies considered the case where diagonal terms are not present, i.e. neural network models without auto-interactions. Subsequent studies [7, 8] considered also the problem with auto-interactions, but only recently it has been pointed out [9] that in this case an interesting regime can be found at . The probability that a given pattern is not a fixed point of the dynamics was studied and it has been shown that this probability is very small for very low , as expected, but surprisingly that there is another unexplored region at very large where this probability is very small as well. While the former and other intermediate regimes are well studied [6], the behaviour at was unexpected, since it implies that in this new regime the patterns are again fixed points of the dynamical equations. Moreover, it has been pointed out [9] that this new regime does not appear in the absence of diagonal interaction terms.

In this paper we study the stability of these fixed points. The relevance of this analysis comes from the fact that associative memories are useful when (in the regime where) they recover memories on the basis of similarity. In other words, the Hopfield model can be used to retrieve memories when starting the dynamics from a configuration similar, but not exactly equal to, a given pattern, we converge to one of the original patterns. Our analysis suggests that although training patterns are fixed points of the dynamics in the newly discovered regime [9], they are unstable, contrarily to the known regime of small values. In section 2 we introduce the model and its dynamics; while in section 3 we compute the probability of escaping the stored pattern when a small perturbation is introduced. In section 4 we complete the analysis by computing the typical number of errors made after one dynamical step, where errors are measured in terms of the Hamming distance between the initial configuration (one of the training vectors) and the dynamical configuration.

2. Dynamics of a neural network

The neural network model that we will consider in this work is a system of binary variables (neurons) , , interconnected by a symmetric network of synapses specified by the real coupling matrix . We will focus on the non-linear dynamical equations

\[
    s_i(t+1) = \text{sign}(x) \left( \sum_{j=1}^{N} J_{ij} s_j(t) \right),
\]

where the value represents the state of the neuron at time , which may be active, , or inactive, . The value depends on the state of the neurons at the previous time step, . These equations give rise to a dynamical process in the space of configurations, depending on the properties of the matrix but, as pointed out by Hopfield [1], a careful choice of may trap the dynamics in basins of attraction that correspond to a given random set of patterns (training vectors) where and . These patterns can be considered as memories that the system is able to retrieve and should be fixed points of equation (1). In the following we will focus on the case where the matrix is specified by the Hebbian rule [4].
\[ J_{ij} = \frac{1}{P} \sum_{\nu=1}^{P} \xi^\nu_i \xi^\nu_j, \]  

(2)

introduced in order to explain associative learning. In fact, for \( P = O(N) \), equation (2) can be obtained cumulatively from the successive application of the learning rule \( \Delta J_{ij} \propto \xi^\mu_i \xi^\mu_j \), specifying the change in the coupling between neurons when learning a given pattern \( \mu \) and describing the observation that the simultaneous activation of neurons \( i \) and \( j \) increased the coupling strength between them.

Retrieval of patterns is known to be possible only for a number of patterns that is a small fraction of that of the neurons \([1, 2, 5, 6]\). Diagonal interaction terms were not considered in the early works about Hopfield model for a physical reason: in the corresponding spin models the field of a variable is induced by the state of its neighbours, but not on its own; thus self interactions do not exist and \( J_{ii} = 0, \forall i \). Neural networks with diagonal terms have been studied in \([9]\) and a very interesting regime has been found for \( \alpha \gg 1 \). The probability \( p_V \) that a given pattern \( \xi^\mu \) is a not fixed point of the dynamics has been computed and has been shown to be very small for very low \( \alpha = P/N \), as expected, but surprisingly another region has been identified at the very large \( \alpha \) regime, where \( p_V \) is also very small. In the intermediate regime, \( p_V \) is large. The first and the intermediate regime are well known but the behaviour of \( p_V \) at \( \alpha \gg 1 \) was new and unexpected. The probability \( p_V \) that a random vector, not in the training set \( \{\xi^\nu\}_{\nu=1,...,P} \), is not a fixed point of the dynamics has also been studied. As expected, in the low \( \alpha \) regime, the probability \( p_V \) is close to 1 but, and as \( \alpha \) increases this probability vanishes. Thus, in the new, large \( \alpha \) regime, \( p_V \) is also very small. In other words, in this regime, both patterns \( \xi^\mu \) and random configurations are likely to be fixed point of the dynamics. This has a trivial interpretation by noticing that for very large \( P \gg N \), the interaction matrix \( J_{ij} \) defined in equation (2) tends to the unit matrix. Even if this result seems to invalidate the usefulness of this new regime, it was shown that the ratio \( p_V/p_V \) tends to a finite number, \( e \), in the large \( N \) limit and that real patterns have an higher probability of being fixed points of the model. In the next section we address the stability question of these patterns, making use of the same strategy used in \([9]\).

3. Stability of the fixed points

In order to study the stability of fixed points of the dynamical equations given in equation (1), we consider the case where one of the patterns is randomly perturbed. Since the patterns \( \xi^\mu \) are configurations of binary variables, a random perturbation is obtained by flipping the value of \( K \) sites. We denote by \( \mathcal{P} \) the set of perturbed sites and by \( \mathcal{L} \) the set of unperturbed variables and clearly, \( \mathcal{L} = \overline{\mathcal{P}} \). We can consider the equations

\[ s'_i = \text{sign} \left( \sum_{j=1}^{N} \sum_{\nu=1}^{P} \xi^\nu_j \xi^\nu_i s_j \right) \]  

(3)

and compute the probability to get back to \( \xi^\mu_i \) when the starting configuration is given by

\[ s_j = \xi^\mu_j I_{j,\mathcal{L}} - \xi^\mu_j I_{j,\mathcal{P}}, \]  

(4)

where \( I_{j,\mathcal{P}(\mathcal{L})} \) is one if \( j \in \mathcal{P}(\mathcal{L}) \), and zero otherwise. As in \([9]\) we focus on the one-step dynamics.

Let us first consider the case when \( i \notin \mathcal{P} \). After some elementary algebra, the argument of the sign function in the rhs of equation (3) becomes


\[(P + N - 1 - 2K)\xi^\mu + \sum_{j \neq \mu} \xi^\mu_j \xi^\nu_j \left[ \xi^\mu_j I_{j,C} - \xi^\nu_j I_{j,P} \right].\] (5)

The second term contains \((N - 1)(P - 1)\) uncorrelated terms of unitary variance. Using the central limit theorem we obtain for large \(N\)

\[s'_i = \text{sign} \left( \frac{(P + N - 1 - 2K)}{\sqrt{(N - 1)(P - 1)}} \xi^\mu + z \right),\] (6)

where \(z \sim \mathcal{N}(0,1)\) is drawn from a normalised Gaussian distribution. Clearly, if \(z\) were 0, \(s'_i = \xi^\mu_i\) and we recover the correct sign. The variable \(z\), induced by the arbitrary bit flips, is thus a destabilising term, that impacts on the rhs of equation (6). It is actually harmless as long as it does not change the sign of \(\xi^\mu_i\), thus we make a mistake on the value \(s'_i\) with a probability equal to

\[p^C = \int_{-\infty}^{\frac{(P + N - 1 - 2K)}{\sqrt{(N - 1)(P - 1)}}} P(z) dz = \frac{1}{2} \text{erfc} \left( \frac{(P + N - 1 - 2K)}{\sqrt{2(N - 1)(P - 1)}} \right).\] (7)

Analogously, when \(i \in \mathcal{P}\), the rhs of equation (3) becomes

\[(N - P + 1 - 2K)\xi^\mu + \sum_{j \neq \mu} \xi^\mu_j \xi^\nu_j \left[ \xi^\mu_j I_{j,C} - \xi^\nu_j I_{j,P} \right].\] (8)

Using a central limit argument one obtains

\[s'_i = \text{sign} \left( \frac{(N - P + 1 - 2K)}{\sqrt{(N - 1)(P - 1)}} \xi^\mu + z \right),\] (9)

where \(z\) is again drawn from a normalised Gaussian distribution. As in equation (7) we obtain the probability of making an error on one of the perturbed variables,

\[p^P = \int_{-\infty}^{\frac{(N - P + 1 - 2K)}{\sqrt{(N - 1)(P - 1)}}} P(z) dz = \frac{1}{2} \text{erfc} \left( \frac{(N - P + 1 - 2K)}{\sqrt{2(N - 1)(P - 1)}} \right).\] (10)

Notice that \(p^P\) and \(p^C\) differ in the sign in front of \(P\), indicating the contribution coming from the diagonal interaction components. This contribution is always aligned with the variable value and consequently decreases the error probability in unperturbed variables. In the limit of large \(N\) and \(P = \alpha N\), and finite \(K\) we obtain

\[p^C = \frac{1}{2} \text{erfc} \left( \frac{1}{\sqrt{2\alpha}} - \sqrt{\frac{\alpha}{2}} \right) \quad p^P = \frac{1}{2} \text{erfc} \left( \frac{1}{\sqrt{2\alpha}} + \sqrt{\frac{\alpha}{2}} \right).\] (11)

While \(p^C\) tends to zero as \(\alpha \to 0\) and \(\alpha \to \infty\), with a maximum at \(\alpha = 1\), the error probability for perturbed spins \(p^P\) is an increasing function of \(\alpha\), going from 0 to 1. This difference in stability between perturbed and unperturbed spins affects the overall stability of the original pattern. Since there are \(K\) perturbed variables and \(N - K\) unperturbed variables, the probability of failing to recover the original pattern \(\xi^\mu\) after a single step of parallel dynamics is

\[p^K = 1 - (1 - p^P)^K (1 - p^C)^{N-K}\] (12)

that, for \(K = 0\), becomes the probability \(p_0^P\) [9]. This probability can be plotted for different \(P\) values at a given \(N\). In figure 1(a) we plot \(p^K\) for \(N = 100\) and \(K = 1\) (blue) and \(K = 0\) (red), where the second case corresponds to the unperturbed case studied in [9]. We also performed
numerical simulations (dots) in systems of $N = 100$ variables for a different number of $P$, counting the number of times that taking one of the training vectors and perturbing $K = 1$ of its values we did not recover the original training vector, repeating this procedure $10^4$ times for different patterns at any value of $P$. (a) For $N = 100$ and (b) For $N = 500$.

4. Large deviations

In this section we compute the typical number of errors observed in $s'$ with respect to the original patterns, produced by applying equation (3) to the vector of dynamical variables $s$, which is specified by equation (4) for a given number $K$ of perturbed spins. Let us denote by $M_P$ the number of errors in the set of perturbed spins $P$ and by $M_L$ the number of errors in the set of unperturbed spins $L$ in $s$. While the probability of $M_P$ is given by

$$P_P(M_P) = \binom{K}{M_P} (p^P)^{M_P} (1 - p^P)^{K-M_P}, \quad (13)$$

the probability of $M_L$ is given by

$$P_L(M_L) = \binom{N-K}{M_L} (p^L)^{M_L} (1 - p^L)^{N-K-M_L}. \quad (14)$$

Let us denote the total number of error by $M = M_P + M_L$. The probability of $M$, the number of errors at the next time step, is given by
and we readily obtain
\begin{equation}
P(M) \propto \min(K,M) \sum_{M_P=\max(0,M+K-N)}^{\min(K,M)} \sum_{M_L=0}^{N-M_M} P_P(M_P) P_L(M_L) \delta_{M,M_P+M_L}
\end{equation}
and we readily obtain
\begin{equation}
P(M) \propto \min(K,M) \sum_{M_P=\max(0,M+K-N)} P_P(M_P) P_L(M - M_P).
\end{equation}

Since we are mostly interested in the large $N$ behaviour, we denote by $K = N \rho$, $M = N m$, $M_P = N m_P$ and $M_L = N m_L$, and use the Sterling’s formula for approximating the factorial of a large integer,
\begin{equation}
\lim_{N \to \infty} N! = \sqrt{2\pi N} \left( \frac{N}{e} \right)^N.
\end{equation}
Simple algebra leads to the expression

$$P(m) \propto \int_{\max(0,m+\rho-1)}^{\min(\rho,m)} \, dm \, e^{\phi(mP,m)} \approx e^{N\phi(m^*,m)}$$

(18)

where $m^*_P$ is the maximum of $\phi(m_P,m)$ over $m_P \in [\max(0,m+\rho-1),\min(\rho,m)]$ at a given $m$. The large deviation function of the probability $P(m)$ is given by $\phi(m^*_P,m) \equiv \psi(m)$ and its expression is

$$\psi(m) = \rho \log(\rho) + (1-\rho) \log(1-\rho)$$

$$+ m^*_P \log\left(\frac{p^c}{m^*_P}\right) + (\rho - m^*_P) \log\left(\frac{1-p^c}{\rho-m^*_P}\right)$$

$$+ (m-m^*_P) \log\left(\frac{1-p^c}{m-m^*_P}\right)$$

$$+ (1-\rho - m + m^*_P) \log\left(\frac{1-p^c}{1-\rho-m+m^*_P}\right).$$

(19)

Let us first consider the case $\rho = 0$, the case where the selected pattern is not perturbed at all. In figure 2(a) we plot

$$\omega(m) = \log(\log(-\psi(m) + 1) + 1)$$

(20)

for $N = 100, K = 0$ as a function of $\log(m + 1)$ for different choices of $P$. The maximum $m^*$ of $\psi(m)$ corresponds to a minimum of $\omega(m)$, where the double logarithm is chosen to emphasise the difference between different lines at large $P$ values. While at small $P$ values we find $m^* = 0$, i.e. the most likely value for $m$ is zero, corresponding to a non-increasing number of errors, as $P$ increases the probability of observing $m = 0$ decreases and for $P > 15$ (which corresponds roughly to $\alpha = 0.15$) we find a different minimum at $m > 0$, as can be seen in figure 2(b) and 3(a). We also observe that the probability of $Nm$ is sharply peaked in 0 in the low $P$ regime, while it is much broader in the large $P$ regime, even if it is clearly visible that the probability of observing $Nm = 0$ is negligible.

The behaviour of $m^*$ at $K = 10$ can be seen in figure 2(d) leading to a qualitatively similar behaviour of $\omega(m)$ shown in figure 2(c), where we observe that small values of $P$ are
dominated by $m^* = 0$, while as $P$ increases $m^*$ remains greater than zero. In other words, while the low $P$ regime leads to a recovery of the original pattern with probability 1 even in the case when we perturb $K = 10$ variables, the large $P$ regime does not. Notice in figure 2(d) that the value of $m^*$ is mainly given by $m^*_P$ with fluctuations of order $1/N$: errors are mainly induced by the set of perturbed spins that remain blocked because of the dominating diagonal terms.

Finally, to emphasize the sensitivity of fixed points to perturbations we plotted the $m^*$ values for a single pattern error $K = 1$ shown in figure 3(b). We observe the same qualitative behaviour of figure 2(b).

5. Conclusion

The discovery of fixed points in the Hopfield model at the large number of patterns limit raised new hopes for a high-capacity properties of the Hopfield model, especially in within the context of associative memories in neural networks. We examine the usefulness of the newly discovered fixed points by focussing on their ability to recover stored patterns on the basis of incomplete information. In other words, the ability to converge to the original pattern when starting from a configuration that is similar to, but not exactly equal to it. We study the stability properties of these fixed points and show that these fixed points are unstable with respect to small perturbations. We also investigate the typical number of errors made by the one step dynamics given in equation (1) and find that while this number is zero in the low storage regimes, it is not in the new large storage regime.

We notice that a simple statistical mechanical argument suggests that it is unlikely that the phase diagrams of an Hopfield model with auto-interactions differs from the phase diagram of an Hopfield model without auto-interactions. In fact, the partition functions of an Hopfield model with auto-interactions and that of an Hopfield model without auto-interactions differ from sub-leading terms in $N$ and so their thermodynamical properties have to be the same. Thus the presence of a new, unexplored, thermodynamical phase comprising multiple stable fixed points with non-vanishing basins of attraction at $\alpha \gg 1$ has to be ruled out. On the other hand, the non-equilibrium dynamics of the two systems may still be different and our analysis shows that the presence of auto-interactions does not change the description of the stability of patterns in the large $\alpha$ regime.

Finally, it has been drawn to our attention that among the previous attempts to stabilise patterns, the presence of a reinforcement term in the dynamics improves the retrieval performances \[10, 11\]. Acting as a sort of stabilizer, the effect of a reinforcement term is similar to a diagonal coefficient in the interaction matrix. It could be then potentially interesting to study the stability of the fixed points in the high-storage regime under a dynamics where a reinforcement term is considered.

Acknowledgment

Support from The Leverhulme Trust grant RPG-2013-48 is acknowledged. DT was supported by National Group of Mathematical Physics (GNFM-Indam). We wish to thank Pierfrancesco Urbani for insightful discussions.

ORCID iDs

Jacopo Rocchi  https://orcid.org/0000-0002-4535-9426
David Saad  https://orcid.org/0000-0001-9821-2623
References

[1] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities Proc. Natl Acad. Sci. 79 2554–8
[2] Hopfield J J 1984 Neurons with graded response have collective computational properties like those of two-state neurons Proc. Natl Acad. Sci. 81 3088–92
[3] Hopfield J J, Feinstein D I and Palmer R G 1983 ‘Unlearning’ has a stabilizing effect in collective memories Nature 304 158–9
[4] Hebb D O 1949 The Organization of Behavior: a Neuropsychological Theory (New York: Wiley)
[5] Amit D J, Gutfrreund H and Sompolinsky H 1985 Spin-glass models of neural networks Phys. Rev. A 32 1007
[6] Amit D J, Gutfrreund H and Sompolinsky H 1985 Storing infinite numbers of patterns in a spin-glass model of neural networks Phys. Rev. Lett. 55 1530
[7] Kabashima Y and Saad D 2001 The TAP approach to intensive and extensive connectivity systems Advanced Mean Field Methods: Theory and Practice ed M Opper and D Saad (Cambridge, MA: Massachusetts Institute of Technology)
[8] Mézard M 2017 Mean-field message-passing equations in the hopfield model and its generalizations Phys. Rev. E 95 022117
[9] Folli V, Leonetti M and Ruocco G 2017 On the maximum storage capacity of the hopfield model Frontiers Comput. Neurosci. 10 144
[10] Sen P and Chakrabarti B K 1992 Extended memory loading capacity of a neural network model with delayed time interaction Phys. Lett. A 162 327–30
[11] Maiti P K, Dasgupta P K and Chakrabarti B K 1995 Improved performance of the Hopfield and little neural network models with time delayed dynamics Int. J. Mod. Phys. B 9 3025–37