A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks

A thesis submitted during 2022 to the University of Hyderabad in partial fulfillment of the award of

Doctor of Philosophy

in

Cognitive Science

by

S Suchitra

Centre for Neural and Cognitive Sciences
School of Medical Sciences

University of Hyderabad
P.O. Central University, Gachibowli,
Hyderabad - 500 046
Telangana
India

December 2022

Declaration

I, **S Suchitra** hereby declare that the work presented in this thesis entitled "**A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks**" submitted by me under the supervision of **Prof. Vipin Srivastava** has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma. Keeping with the general practice, due acknowledgements have been made wherever the work described is based on other investigations.

I also declare that this thesis is free from plagiarism. A report on plagiarism statistics from the University Librarian is enclosed.

Date: 19-12-2022

S Suchitra
09CCPC01

UNIVERSITY OF HYDERABAD
Centre for Neural and Cognitive Sciences, School of Medical Sciences

---

CERTIFICATE

This is to certify that the thesis titled **"A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks"** submitted by **S. Suchitra** (Registration No. **09CCPC01**) in partial fulfilment of the requirements for the award of Doctor of Philosophy in Cognitive Science in the Centre for Neural and Cognitive Science under School of Medical Sciences in a bona fide work carried out by her under my supervision and guidance.

The thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for the award or any degree or diploma.

Prof. Ramesh Kumar Mishra
Head, Centre for Neural and Cognitive Sciences

Head
Centre for Neural & Cognitive Sciences
University of Hyderabad
Hyderabad-500 046. INDIA.

Prof. Vipin Srivastava
(Supervisor)
School of Physics
University of Hyderabad
Hyderabad - 500046
India

Prof. Geeta K. Vemuganti,
Dean, School of Medical Sciences

सकाय अध्यक्ष
Dean
चिकित्सा विज्ञान संकाय
School of Medical Sciences

UNIVERSITY OF HYDERABAD

Centre for Neural and Cognitive Sciences,

School of Medical Sciences

CERTIFICATE

This is to certify that the thesis entitled **"A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks"** submitted by **S. Suchitra** bearing Registration Number **09CCPC01** in partial fulfilment of the requirements for award of Doctor of Philosophy in the Centre for Neural and Cognitive Science under School of Medical Sciences is a bonafide work carried out by her under my supervision and guidance.

This thesis is free from Plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for award of any degree or diploma.

Further, the student has the following publications before submission of the thesis for adjudication and has produced evidence for the same in the form of acceptance letter or the reprint in the relevant area of his research:

1. **Sampath, S.**, & Srivastava, V., On stability and associative recall of memories in attractor neural networks. *PloS ONE*, 15(9), e0238054, 2020.
   Chapter of dissertation where this publication appears: **3**

2. Srivastava, V., & **Sampath, S.**, Could the brain function mathematically? *Neurology and Neuroscience Research*, 1(1):4, 2018.

3. Srivastava, V., & **Sampath, S.**, Chapter Fourteen- Cognition of Learning and Memory: What Have Löwdin's Orthogonalizations Got to Do With That?, *Advances in Quantum Chemistry*, 74, 299-319, 2017.
   Chapter of dissertation where this publication appears ***5***

4. Srivastava, V., **Sampath, S.**, & Parker, D. J. . Overcoming Catastrophic Interference in Connectionist Networks Using Gram-Schmidt Orthogonalization. *PloS ONE*, 9(9), e105619, 2014.

Chapter of dissertation where this publication appears ***4***

and has made presentations in the following international conferences:

1. **Sampath, S.** & Srivastava, V., On basins of attraction in attractor neural networks, APS (American Physical Society) March Meeting 2015; San Antonio, Texas (U.S.A.).

2. Srivastava, V. & **Sampath, S.**, Mathematical modeling of cognitive learning and memory, 13th International Conference on Cognitive Modeling (ICCM 2015); Groningen (The Netherlands).

Further, the student has passed the following courses towards fulfilment of coursework requirement for Ph.D. :

S.No.	Course code	Title of the course	Credits	Result
1	PC900	Formal and Computational Approaches to Cognition	4	Pass
2	PC901	Empirical Bases of Cognition	4	Pass
3	PC902	Language, Philosophy and Cognition	4	Pass
4	PC903	Dissertation oriented readings	4	Pass
5	PC926	Statistics and Research Methodology	4	Pass
6	PC931	Computational Intelligence	2	Pass
7	PC932	Cellular and Molecular Neuroscience	2	Pass
8	CO811	Reading Course "Information Processing in the Nervous System - I"	4	Pass
9	CO812	Reading Course "On the Role of Cerebellum in Motor Function and Cognition"	4	Pass

Prof. Vipin Srivastava
Supervisor

Prof. Ramesh Kumar Mishra
Head, CNCS

Prof. Geeta K. Vemuganti
Dean, SMS

हैदराबाद विश्वविद्यालय

University of Hyderabad

Certificate of Title

Enrolment No: 09CCPC01

Name of the Scholar: S Suchitra

Course of Study: Ph.D. Cognitive Science

Title of the Thesis: A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks

Name of the supervisor: Prof. Vipin Srivastava

Department/School: Centre for Neural and Cognitive Sciences, School of Medical Sciences

Dated: 23-12-2022

Controller of Examinations

Acknowledgements

At this major juncture in my academic life, I would like to express my feeling of gratitude towards many people who have helped me directly or indirectly in finishing my PhD thesis.

First and foremost I thank my PhD supervisor, Prof. Vipin Srivastava who introduced me to the challenging field of Cognitive Science, and guided me into the area of theoretical and computational cognitive neuroscience. I am extremely grateful to him for the immense amount of patience through the years, and for his constant guidance, encouragement and support.

I am indebted to Dr. Joby Joseph for his time and many invaluable discussions. I thank the faculty members Prof. Prajit K. Basu, Prof. S. Bapi Raju, Dr. Kiranmayi S. Bapi and Prof. D. Vasanta for their courses which gave me an idea of the scope of and diversity within the field of Cognitive Science. I am grateful to the other members of my Doctoral Committee, Prof. S. Bapi Raju and Prof. Samrat L. Sabat, and Prof. Ramesh Kumar Mishra, Head of CNCS for their guidance and support. I also thank Dr. Sudipta Saraswati for discussions and advice. I am also extremely grateful to Prof. Geeta K. Vemuganti for her timely help and support.

I thank the University of Hyderabad (UoH) for providing a stimulating academic environment for carrying out my research, and amidst a lush green campus. I am grateful to the Centre for Neural and Cognitive Sciences (CNCS) for providing me with the necessary facilities and infrastructure to carry out my research work. I also thank the staff of CNCS Lakshmi, Shalini, Ramchandar, Sarada, Keerthi and also Mr. Srinivas (School of Physics) and Mr. Panikkar (Library) for all their help. Thanks are also due to the Centre for Modelling, Simulation and Design (CMSD) for letting me use their resources for part of my research work.

To my friends and colleagues at CNCS: thank you. Special thanks to Shalini for being a great friend and for the encouragement and positive words when I needed them. Thank you, Shilpi, Tony, Anuj, Natha and Neelkanth, for your friendship and the memories. Sivaraju, Sandhya, Ankit and Abhilash- thank you for your invaluable help and for all the legwork on my behalf when I could not be there in person to do it myself. My batchmates Kiran, Venkat, Srinivas and Pavan- thank you for the company at the beginning of my journey into Cognitive Science. I will forever

be grateful to you, Rakesh and Jigar, for all the discussion sessions on coursework, codes and research work, as well as Life in general.

Special thanks to my best friend Malati, for being with me through the decades, boosting my confidence and giving me reality checks as needed, and for never losing your faith in me. Thanks also to Fathima and Lalitha for refraining from asking about the status of my thesis and pointing out their restraint in doing so. Thank you; Nikita, Mohanaselvi, Maunika, Manvi, Radhika, Vaishali, Uvashree, for being the friends and support system every woman needs. Bhavya and Praveen – from a shared love of languages to all the new adventures, I am forever thankful to you. To the friends I bonded with over a love of books, humour and puns – Nisha, Nandini, Ashwin – thank you.

I am also extremely grateful to my husband, Abhishake for discussions, comments, help and unwavering support. I also thank him for taking on the lion's share of dreary chores to help me, for putting up with my mood swings and for laughing at my silly jokes and stories. I am also grateful to my daughter Ananya for being a calm and cheerful baby, for entertaining herself and for letting others entertain her, for sleeping in intervals long enough for me to finish writing and editing my thesis and for being my latest teacher.

Most importantly, none of this would have been possible without the love, patience and support of my family. I am indebted to my father Sampath for his unwavering support and understanding and my mother Radha, for her level-headed words and encouragement when I needed them. I am also grateful to them for their major contribution to childcare, taking care not just of my daughter, but also theirs. I thank my brother Srikanth, sister-in-law Nisha and nephew Abhinav for being the pick-me-up I needed, and for entertaining me with their made-up songs.

It takes a village to raise a child - I thank my in-laws Asit, Shyamali and Chhanda for being part of the village raising mine, and for providing me with the time and peace of mind to focus on my thesis and research. I thank my extended family (especially my (late) grandparents S.K. Chary and Kamala) for their company, the laughter and unconditional love and support.

Last but not the least, I would like to thank UoH and the Department of Science and Technology (DST) for financial support during part of my Ph.D.

- Suchitra

List of Publications

■ **Sampath, S.** & Srivastava, V., On stability and associative recall of memories in attractor neural networks. *PloS ONE*, 15(9), e0238054, 2020.

■ Srivastava, V. & **Sampath, S.**, Could the brain function mathematically? *Neurology and Neuroscience Research*, 1(1):4, 2018.

■ Srivastava, V. & **Sampath, S.**, Chapter Fourteen- Cognition of Learning and Memory: What Have Löwdin's Orthogonalizations Got to Do With That?, *Advances in Quantum Chemistry*, 74, 299-319, 2017.

■ Srivastava, V., **Sampath, S.**, & Parker, D. J. . Overcoming Catastrophic Interference in Connectionist Networks Using Gram-Schmidt Orthogonalization. *PloS ONE*, 9(9), e105619, 2014.

List of Conferences

■ **Sampath, S.** & Srivastava, V., On basins of attraction in attractor neural networks, APS (American Physical Society) March Meeting 2015; San Antonio, Texas (U.S.A.).

■ Srivastava, V. & **Sampath, S.**, Mathematical modeling of cognitive learning and memory, 13th International Conference on Cognitive Modeling (ICCM 2015); Groningen (The Netherlands).

ABSTRACT

How the brain forms memories and associations between them has been a long-standing question in psychology and neuroscience. Computational models, in particular, the Hopfield model[1] with Hebbian learning[2] have been shown to be useful in studying the mechanisms underlying learning and memory in the brain. However, the Hopfield model suffers from a severe limitation in terms of the amount of information it can store before it experiences the so-called catastrophic blackout resulting in an abrupt and complete loss of information. A solution to the capacity problem by invoking an orthogonalization scheme in the Hopfield model has been proposed earlier[3] and demonstrated to be effective in increasing the memory capacity of the network. In this thesis, we first analyze the post-synaptic potential in detail to show how the Gram-Schmidt orthogonalization scheme helps in overcoming the memory catastrophe and enhances memory capacity. We then address some fundamental issues related to memory stability and the associative character of the network. We also define mathematically the terms retrieval, recognition and recall in order to list out in exact terms the conditions required for pattern stability and hence for the efficient functioning of an associative memory network. Apart from the sequential Gram-Schmidt orthogonalization procedure, in this thesis we also study the effects of invoking the democratic Symmetric and Canonical schemes due to Löwdin[4, 5, 6] on the dynamics of the Hopfield network. We have also attempted to situate our studies in the context of cognition, to understand how the results of the study relate to biological learning and memory.

Contents

List of Figures

List of Tables

To my family

Chapter 1

Introduction

A refrain from a song or the scent of a meal can evoke memories, the sound of a passing train could remind you of your first train journey. Seeing a face could trigger memories of familiarity, especially if they bear some resemblance to someone you know or have come across. How does this happen? How do we learn and form memories? How does seeing or hearing something remind us of something else? What makes the brain capable of forming associations between the various things we learn or remember?

These questions have long been the focus of research across domains including psychology, neuroscience and computation. Computational modelling of observed cognitive phenomena using principles from mathematics, physics and computer science have proven to be a valuable tool in furthering our understanding of cognitive mechanisms. In this thesis, we try to address some of the questions mentioned above using a mathematical model.

In the decades since it was first proposed, the Hopfield model[1] has been studied extensively in the context of learning and memory. The model is a multi-nodal fully-connected network which can memorize and recover information represented by vectors with binary components. The model employs the learning principle postulated by Hebb[2]. The information thus stored become points of minimum energy. Not only that, each memorized vector is associated with a set of vectors that collectively form its basin of attraction. This property of the network makes the vectors attractors, which is why such networks are also known as attractor neural networks.

The appeal of the Hopfield model lies in its simplicity, though the network is still capable of reproducing many observed complex phenomena. In spite of its limitations, the model remains relevant to our attempts at studying the mechanisms underlying learning and memory.

The brain is capable of not just storing information, but also categorizing and classifying information and forming associations between them. This aspect of the functionality of the brain is captured well by the Hopfield model using principles from physics (see for instance, [7] or [8]). The model acts not just as a storage system, but also as an associative memory. However, the storage capacity of the model is severely limited, and is just a small fraction of its size. Moreover, beyond this low limit, there is a sudden and total loss of memory, a phenomenon referred to as the memory catastrophe or blackout.

Implementation of orthogonalization in the Hopfield model has been proposed earlier[3] as a solution to the capacity problem. It has been demonstrated to be effective in increasing the memory capacity of the network and in overcoming the memory catastrophe.

Orthogonalization refers to the transformation of a set of linearly independent vectors into an orthonormal basis, a set whose component vectors are all perpendicular to each other and normalized, that is, of uniform length. It has been explored extensively in physics, chemistry and mathematics. While orthogonalization may initially appear completely unrelated to cognition, it may not be so – we have shown that orthogonalization might be a part of the mechanisms underlying learning and memory, and may also be biologically plausible.

The Gram-Schmidt orthogonalization scheme plays a role in the capacity of the brain to learn, store and distinguish between information [15, 16]. The information is first perceived through the sense organs and then received by the brain, and may include sights, sounds and/or smells. We claim argue that the brain is capable of the process of orthogonalization and can differentiate new from the previously memorized information [24]. By comparing new incoming information with what is already present, the brain is, in essence, performing the process of orthogonalization. The brain may thus possess the physiological architecture and mechanisms required for orthogonalization. This idea is part of our hypothesis that the functioning of the brain might be, by and large, mathematical in nature.

We have shown earlier how when the brain compares different information and identifies similarities and differences between them, it is, in fact, carrying out the process of Gram-Schmidt orthogonalization [15, 16]. Adopting the Hopfield model[1] as our base , we have seen that when the model brain uses Gram-Schmidt orthogonalization in the memorization process, it compares the new information coming in with what has been memorized previously and checks how similar or different they are. Following orthogonalization, what the model brain memorizes is the result of the comparison, the similarities and differences the model brain has established, rather than the information in its entirety. A significant result of the model is the substantial increase in the memory capacity of the network. Another interesting and remarkable highlight of the model is that in spite of the network memorizing the orthogonalized versions, or the similarities and differences between the incoming information, the model brain is capable of recognizing the complete input information when it is presented to the network. In other words, the network can recover the presented input information with perfect accuracy. The network is also capable of associative recall of the information, that is, each input information has a basin of attraction around it, indicating that the information is content addressable [7].

We hence argue that the brain might have become capable of performing orthogonalization along the lines of the Gram-Schmidt scheme early during the evolutionary process, though the mathematical procedure was invented relatively recently, about a century ago.

1.1 Objectives of the thesis

In this thesis, we first show analytically how the proposed orthogonalization scheme provides a solution to the catastrophic blackout problem, apart from increasing the memory capacity of the network. To do so, we present a detailed analysis of the post-synaptic potential. We also show how the scheme addresses the stability-plasticity dilemma.

However, a deeper probe into the robustness of the enhanced memory capacity raises some fundamental questions related to pattern stability and the associative character of the network, which we will try to address. Pattern stability forms the crux of an efficient storage system, ensuring that the information stored in the system can be recovered. While the concept of pattern stability in the Hopfield model can be

understood in straightforward terms, the results of our study show that this notion may not be universal in its definition. Different criteria have been used to term a pattern stable or unstable (refer to [9] and [8]for stability criteria). We list out all the conditions stable patterns must satisfy, and also define them mathematically.

This brings us to re-examine what we mean by terms such as retrieval, recognition and recall. While these terms are often used interchangeably in the literature, their meanings differ across domains. For consistency, we define each term precisely in our context. We also express them in mathematical terms, and discuss how our definitions compare to their counterparts in other fields such as psychology or computation.

1.2 Evaluating network efficacy

We evaluate the efficacy of the Hopfield model in terms of the following:

- the memory capacity of the network,
  that is, the amount of information that can be stored in a network of a certain size;
- the sizes of the basins of attraction of the memories,
  or, the maximum distance between a memory and a pattern within its basin of attraction and hence associated with the memory;
- the network dynamics,
  i.e., the effect of increasing memory loads on the basins of attraction,
- the energy landscape,
  or how the energies of the patterns are affected as more patterns are added to the memory store;
  and
- the effect of correlations between the patterns on the behaviour of the network.

For comparison, we analyze the performance of our model with Gram-Schmidt orthogonalization along the same lines.

Our results indicate that implementing orthogonalization in the Hopfield model not only increases the memory capacity, but also makes the network a more effective

associative memory in comparison with the Hopfield network. Moreover, the Gram-Schmidt scheme also automatically endows the network with many desirable traits in the network performance parameters, such as making the basins of attraction large and uniform, for instance.

Apart from the Gram-Schmidt orthogonalization procedure which is sequential in nature, we also implement two democratic orthogonalization schemes in the Hopfield network. Symmetric and Canonical schemes proposed by Löwdin[4, 5, 6] orthogonalize sets of vectors, independent of the sequence in which the vectors are arranged. We highlight some remarkable properties of the schemes and analyze the behaviour of the Hopfield model invoking these two schemes.

Finally, we interpret what our results could mean in cognitive terms. We also discuss the possible biological circuitry that could realize the proposed schemes.

1.3 Organization of the thesis

This thesis is a study on the effects of various orthogonalization schemes in the Hopfield model. The thesis is organized into various chapters as follows: Chapter 2 presents an overview of the Hopfield model with Hebbian learning, while Chapter 3 discusses GS orthogonalization as a solution to the memory catastrophe. A more in-depth study of the scheme is presented in Chapter 4, with emphasis on pattern stability and the associative character of the network. Chapter 5 looks at Löwdin orthogonalization schemes and their relevance to cognition. The thesis concludes with a summary and a discussion on future research avenues in Chapter 6.

Chapter 2

An overview of the Hopfield model

In this chapter, we present a brief overview of the Hebb-Hopfield model which forms the base upon which we have built the models used in our studies.

2.1 Introduction

We adopt the Hopfield model[1] as our base for our studies. As the model follows the Hebbian learning principle[2], we also refer to the Hopfield model as Hebb-Hopfield or H-H model hereafter. In this thesis, we propose and elaborate upon some schemes that can be implemented in the basic Hopfield network to improve its performance and cognitive relevance, while also taking biological plausibility into account. We will now outline the model and discuss the behaviour and functioning of the network.

2.2 Learning and memorization in the Hopfield network

The Hopfield network[1, 7] is a system of N nodes, called neurons, which are connected to all other neurons without self-connections. They are connected to each other through synapses which are characterized by their nature and weights (also referred to as strengths or efficacies). The synapses can be excitatory or inhibitory in nature and are characterized by efficacies that change in response to new information being memorized by the network. Information is presented to the network in the

form of patterns. Each pattern $\boldsymbol{\xi}^{(\mu)}$ is an $N-$dimensional vector whose components $\xi_i^{(\mu)}$ are ± 1 with equal probability. $+1$ denotes an active or firing neuron while -1 denotes one which is inactive or non-firing. The information is thus represented as a pattern of activities on the neurons given by

$$\boldsymbol{\xi}^{(\mu)} = \{+1, +1, -1, \ldots, +1\}. \tag{2.1}$$

Here we use ± 1 for mathematical convenience, but conversion to $0 - 1$ (0=non-firing/1=firing) notation can be achieved easily as follows:

$$\xi_i^{(\mu)} = 2n_i - 1, \tag{2.2}$$

where ξ_i and n_i represent $-1/+1$ and $0/1$ units respectively[19] (also see [7]).

When an information is to be memorized by the network, it follows the learning rule constructed following Hebb's hypothesis[2] and written as:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \sum_{\substack{i=1 \\ i \neq j}}^{N} \xi_i^{(\mu)} \xi_j^{(\mu)}. \tag{2.3}$$

Here $\xi_i^{(\mu)}$ and $\xi_j^{(\mu)}$ give the activities of the i^{th} and j^{th} neurons in the μ^{th} pattern and J_{ij} is the weight of the synapse between the two neurons i and j. For the neuron i, $\xi_i^{(\mu)} \cdot \xi_i^{(\mu)} = 1$. The addition of new patterns leads to cumulative changes in the synaptic weights.

The activities of the remaining $N - 1$ neurons in pattern ν projecting onto neuron i via J_{ij}'s give rise to a local field potential(LFP) or the post-synaptic potential (PSP) on the neuron i given by:

$$h_i^{(\nu)} = \sum_{\substack{i=1 \\ i \neq j}}^{N} J_{ij} \xi_j^{(\nu)}. \tag{2.4}$$

The neurons are connected with all the neurons except themselves and so the self-connection terms in eq.(2.3) are explicitly made zero. See [1] and also [19] for a discussion of the mathematical reason for this.

2.3 The concept of pattern stability

Patterns stored in the network following the learning prescription in eq.(2.3) form fixed points in the network, as they minimize an energy function, or Hamiltonian, given by,

$$H = -\frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N} J_{ij}\xi_i^{(\mu)}\xi_j^{(\mu)}. \tag{2.6}$$

This condition is a prerequisite for pattern stability and can be verified from the following simple analysis (following [19]). For symmetric connections, the Hamiltonian is given by

$$\begin{aligned} H &= C - \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N} J_{ij}\xi_i^{(\mu)}\xi_j^{(\mu)}, \\ H &= C - \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N} h_i^{(\mu)}\xi_i^{(\mu)}, \end{aligned} \tag{2.7}$$

[1] We know from eq.(2.4) that

$$\begin{aligned} h_i^{(\nu)} &= \sum_{j=1}^{N} J_{ij}\xi_j^{(\nu)}, \\ &= J_{ii}\xi_i^{(\nu)} + \sum_{\substack{j=1\\ j\neq i}}^{N} J_{ij}\xi_j^{(\nu)}, \end{aligned}$$

and so,

$$h_i^{(\nu)}\xi_i^{(\nu)} = J_{ii}\xi_i^{(\nu)}\xi_i^{(\nu)}\xi_i^{(\nu)} + \sum_{\substack{j=1\\ j\neq i}}^{N} J_{ij}\xi_j^{(\nu)}\xi_i^{(\nu)}. \tag{2.5}$$

The first term will always be positive, as $\xi_i^{(\nu)}\xi_i^{(\nu)} = 1$ and $J_{ii} = \sum_{\mu=1}^{p}\xi_i^{(\mu)}\xi_i^{(\mu)} = p$. Now, if $\sum_{\substack{j=1\\ j\neq i}}^{N} J_{ij}\xi_j^{(\nu)} < 0$, then both $\xi_i^{(\nu)} = +1$ and $\xi_i^{(\nu)} = -1$ will result in $h_i^{(\nu)}\xi_i^{(\nu)}$ taking a positive value if the first term on the right side of eq.(2.5) is large. As we shall see later in Chapter 4, this indicates the stability condition being satisfied. That is, $J_{ii} \neq 0$ leads to the presence of additional spurious states in the vicinity of the actual stable states corresponding to the $\boldsymbol{\xi}^{(\nu)}$'s. If $J_{ii} = 0$, $h_i^{(\nu)}\xi_i^{(\nu)} = [\sum_{\substack{i=1\\ i\neq j}}^{N} J_{ij}\xi_j^{(\nu)}]\xi_i^{(\nu)}$ which will be < 0 when $\xi_i{}^{(\nu)} = -1$ and > 0 when $\xi_i{}^{(\nu)} = +1$.

where C is a constant due to the ii terms. The parentheses refers to all the distinct (symmetric) ij pair terms for which ij terms $= ji$ terms. Our aim here is to show that the learning rule in eq.(2.3) can only lead to a decrease in the energy of the system. Now let H' be the new Hamiltonian on some neuron i such that $\xi_i^{'(\mu)} = sgn\left(h_i^{(\mu)}\right)$. Now if $\xi_i^{'(\mu)} = h_i^{(\mu)}$, then $H' = H$ and the Hamiltonian remains as it is. But if $\xi_i^{'(\mu)} = -h_i^{(\mu)}$, then,

$$\begin{aligned} H^{'(\mu)} - H^{(\mu)} &= -\sum_{\substack{i,j=1 \\ i\neq j}}^{N} J_{ij}\xi_i^{'(\mu)}\xi_j^{(\mu)} + \sum_{\substack{i,j=1 \\ i\neq j}}^{N} J_{ij}\xi_i^{(\mu)}\xi_j^{(\mu)}, \\ &= 2\xi_i^{(\mu)} \sum_{\substack{i,j=1 \\ i\neq j}}^{N} J_{ij}\xi_j^{(\mu)}, \\ &= 2\xi_i^{(\mu)} \sum_{j=1}^{N} J_{ij}\xi_j^{(\mu)} - 2J_{ii}. \end{aligned} \tag{2.8}$$

We can ignore the second term containing the self-connection terms. We can infer from eq.(2.6) the definition of and $\xi_i^{'(\mu)}$ that the Hamiltonian can only decrease. Thus, we see that the learning rule (in eq.(2.3)) ensures the minimization of the energy function.

Moreover, for patterns to be stable or retrievable, the signs of the LFP on each neuron must match with the corresponding element of the presented pattern. That is,

$$sgn\left(h_i^{(\nu)}\right) = sgn\left(\xi_i^{(\nu)}\right) \text{ for all } i\text{'s.} \tag{2.9}$$

Alternatively, pattern stability requires

$$s_i h_i^{(\nu)} > 0, \tag{2.10}$$

where $s_i = sgn\left(\xi_i^{(\nu)}\right)$ for all i.

We now use this condition to check the stability of the first pattern $\boldsymbol{\xi}^{(1)}$ when p patterns have been stored in the system. For this, we first evaluate $s_1 h_1$ following eq.(2.10). and separate it into two terms, the first containing the contribution of the

first pattern and the second due to all other patterns as,

$$\begin{aligned} s_1 h_1 &= \frac{1}{N} \sum_{j=2}^{N} \sum_{\mu=1}^{p} \xi_1^{(1)} \xi_1^{(\mu)} \xi_j^{(\mu)} \xi_j^{(1)}, \\ &= \frac{1}{N} \left(\sum_{j=2}^{N} \xi_1^{(1)} \xi_1^{(1)} + \sum_{j=2}^{N} \sum_{\mu=2}^{p} \xi_j^{(\mu)} \xi_j^{(1)} \right), \\ &= \frac{1}{N} \left(N - 1 + \sum_{j=2}^{N} \sum_{\mu=2}^{p} \xi_j^{(\mu)} \xi_j^{(1)} \right), \\ s_1 h_1 &= \frac{N-1}{N} + \sum_{j=2}^{N} \sum_{\mu=2}^{p} \xi_j^{(\mu)} \xi_j^{(1)}. \end{aligned} \tag{2.11}$$

The first term in the equation is straightforward, as $(\pm 1)^2 = 1$. As there are no self-connections and $\xi_j^{(\mu)}$ and $\xi_j^{(1)}$ are uncorrelated for $\mu \neq 1$, each of the four elements in the second term is independent of the other three. Besides, there is a term in each sum which is not correlated with any of the other terms. Hence, each element in the sum can be approximated as ± 1. The second term in eq. (2.11) can therefore be considered as a random walk with $(N-1)(p-1)$ terms.

The first term can be understood as the signal due to the pattern $\boldsymbol{\xi}^{(1)}$, and takes value 1 for $N >> 1$. The second term constitutes the noise due to the correlations between $\boldsymbol{\xi}^{(1)}$ and the remaining $p-1$ patterns. Its value is typically expressed as the root mean square σ given by:

$$\sigma = \frac{N-1}{N} + \sum_{j=2}^{N} \sum_{\mu=2}^{p} \xi_j^{(\mu)} \xi_j^{(1)}. \tag{2.12}$$

We see that eq.(2.11) satisfies the condition in (2.10) when the signal term is larger than the noise term, indicating that the pattern $\boldsymbol{\xi}^{(1)}$ is retrievable. We can generalize this analysis to any of the p memorized patterns.

$$s_i h_i = \frac{1}{N} \sqrt{(N-1)(p-1)} \approx \frac{p}{N}. \tag{2.13}$$

The above condition requires the signs on the neurons to match with perfect accuracy, that is, on all N neurons. However, a match of $\geq 97\%$ is deemed sufficient for pattern retrieval. A plot showing the average number of stable patterns for different memory loads is shown in Fig.2.1.

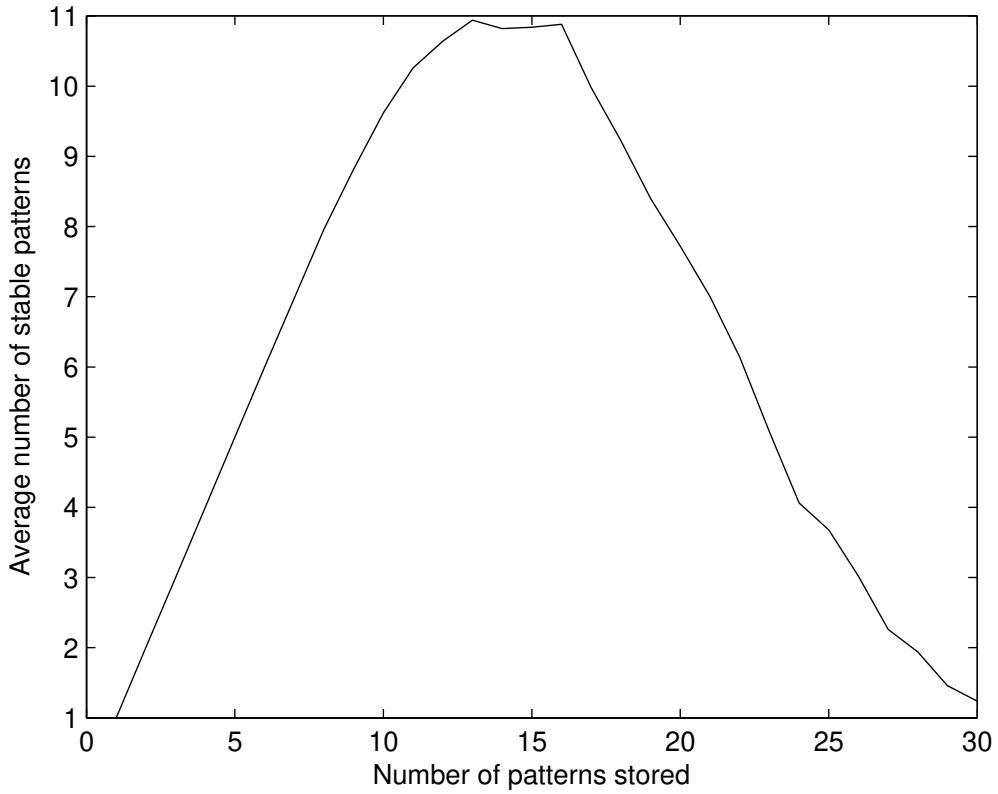


FIGURE 2.1: Stable patterns in the Hopfield model with 100 neurons plotted against the number of memorized patterns. The plot is shown for 50 trials, where each trial refers to a particular set of patterns. Patterns are retrieved completely upto $p = 11$, beyond which patterns start becoming unstable.

2.4 Memory capacity of the network and catastrophic blackout

We can now estimate the memory capacity of the network. The memory capacity of the network gives a measure of the efficacy of the network as an associative memory. The maximum number of patterns p such that all p inscribed patterns can be retrieved gives the memory capacity of the network of size N. In other words, it marks the limit within which the signals arising from the learnt patterns are clearly distinguishable from the noise, rendering the patterns stable. In order to calculate the value of p upto which there no degradation in the stored memory, we must work out the probability of a single neuron being unstable, i.e., $s_i h_i < 0$.

For large values of $(N-1)(p-1)$, when the random walk has a high number of steps, the noise term in eq.(2.13) can be approximated by a Gaussian. Then, the

probability of the neuron i being unstable is given by:

$$P(s_i h_i < 0) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{-1} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} \mathrm{d}x. \tag{2.14}$$

Rewriting the above equation using the error function, we get

$$P(s_i h_i < 0) = \frac{1}{2}\left(1 - erf\left(\frac{1}{\sigma\sqrt{2}}\right)\right). \tag{2.15}$$

The memory capacity α_c is now

$$\alpha_c = \frac{p}{N} = \sigma^2. \tag{2.16}$$

From theoretical and analytical studies, $\alpha_c = 0.144$ for $\sigma = 0.379$ [7, 8] (and references therein). The memory capacity of the network is shown in Fig.2.2 as the fraction of retrieval for different values of the load parameter (p/N).

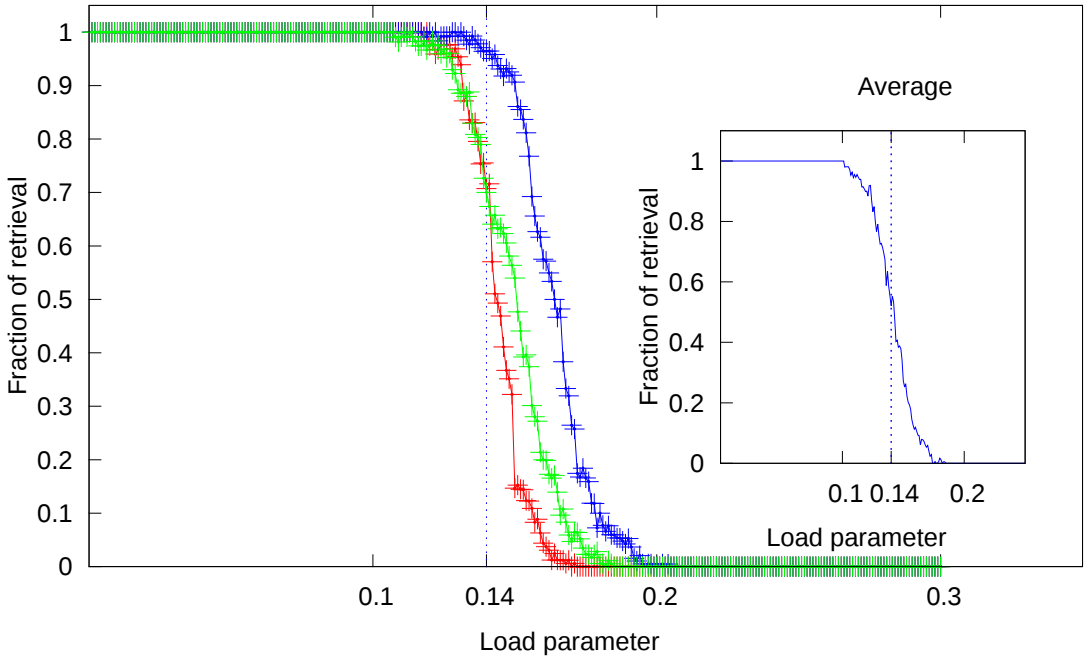


FIGURE 2.2: Plot showing the memory capacity of the network, with the fraction of retrieval plotted against the load parameter (p/N). The plotted data pertains to three different trials with $p = 300$ in a network of size $N = 1000$. The average of 18 sets including these three is plotted in the inset. The fraction of retrieved patterns deteriorates rapidly for loads beyond the theoretical limit of $p = 0.14N$, falling to 0 around $p = 0.17N$.

The above limit gives the critical value at which the network dynamics shift from the stored information being retrievable to all the stored memories being lost. This effect is referred to as catastrophic blackout. Beyond this point, the learnt patterns become unstable as shown in Fig.2.3

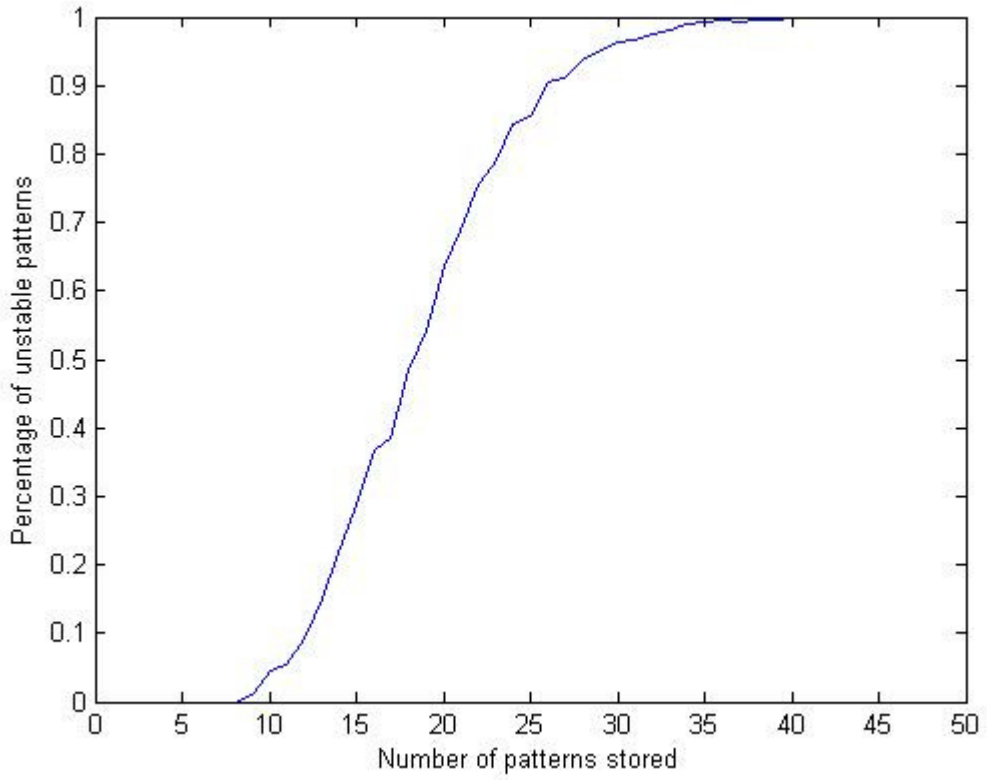


FIGURE 2.3: Unstable patterns in the Hopfield model with 100 neurons plotted against the number of memorized patterns. The plot is shown for 50 trials. Some of the patterns become unstable as p nears 10, but this fraction is negligible. With further increase in p, more and more patterns become unstable.

2.5 Relationship between basins of attraction, pattern stability and memory capacity

The stability of a pattern also indicates the presence of a finite or non-zero basins of attraction. Basin of attraction refers to the set of patterns around an inscribed pattern $\boldsymbol{\xi}^{(\nu)}$ (a minimum in the energy landscape) such that they settle down to that particular pattern $\boldsymbol{\xi}^{(\nu)}$ when presented to the network for retrieval. The patterns

within the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$ share some degree of similarity with $\boldsymbol{\xi}^{(\nu)}$ and are hence *associated* with it. It is intuitive that patterns are associated with or fall within the basin of attraction of the inscribed pattern to which they are the most similar. However, whether a test pattern lies within the basin of attraction of an inscribed pattern depends on the size(or extent or radius) of the basin. The maximum number of differences between a test pattern $\boldsymbol{\xi}^{(test)}$ and an inscribed pattern $\boldsymbol{\xi}^{(\nu)}$ (that is, the Hamming distance between them) such that $\boldsymbol{\xi}^{(test)}$ retrieves $\boldsymbol{\xi}^{(\nu)}$ when presented to the network gives the extent of the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$. The protocol for calculating basins of attraction along with an example is explained in Appendix B. An in-depth study of the basins of attraction will be presented in Chapter 3.

We will now look at the relationship between memory capacity of the network and the sizes of the basins of attraction. In order to get a general estimate the radius of a basin of attraction, following [8] we first choose a test pattern $\boldsymbol{\xi}^{(t)}$ which differs from the first pattern by b elements, i.e., the Hamming distance between $\boldsymbol{\xi}^{(t)}$ and $\boldsymbol{\xi}^{(t)}$ is b. For the sake of convenience, we make the test pattern such that its first b elements are the same as those of the first inscribed pattern $\boldsymbol{\xi}^{(1)}$ and the remaining $N-b$ elements are the inverses of the corresponding elements of $\boldsymbol{\xi}^{(1)}$. That is,

$$\xi_i^{(t)} = \begin{cases} \xi_i^{(1)}, & \text{i} \in \{1, 2, ...N-b\} \\ -\xi_i^{(1)}, & \text{i} \in \{N-b+1, N-b+2, ...N\}. \end{cases} \tag{2.17}$$

Now, to check if the test pattern is stable after p patterns have been stored, we can check the stabilization parameter of either the first or the last neuron (as the signs of the remaining neurons will follow, from eq.(2.17)). $\boldsymbol{\xi}^{(t)}$ will be stable if either $s_1 h_1 > 0$ or $s_N h_N < 0$. And so, using eq.(2.9) and eq.(2.4), we get,

$$s_1 h_1 = \xi_1^{(1)} h_1 = \xi_1^{(1)} \sum_{j=2}^{N-b} J_{1j} \xi_j^{(1)} + \xi_1^{(1)} \sum_{j=N-b+1}^{N} J_{1j}(-\xi_j^{(1)}) > 0, \tag{2.18}$$

or

$$s_N h_N = \xi_N^{(1)} h_N = (-\xi_N^{(1)}) \sum_{j=2}^{N-b} J_{Nj} \xi_j^{(1)} + (-\xi_N^{(1)}) \sum_{j=N-b+1}^{N} J_{Nj}(-\xi_j^{(1)}) < 0. \tag{2.19}$$

As the first and last neurons of $\boldsymbol{\xi}^{(t)}$ have the same and inverted values of the corresponding elements of $\boldsymbol{\xi}^{(1)}$, either of the above two conditions will ensure that the signs on the test pattern and the inscribed pattern match. They are also equivalent,

as eq.(2.19) multiplied by a factor of -1 yields the same criterion as in eq.(2.18). We will now focus on the first neuron and calculate its stabilization parameter:

$$\begin{aligned}
\xi_1^{(1)} h_1 &= \frac{1}{N}\sum_{j=2}^{N-b}\sum_{\mu=1}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)} - \frac{1}{N}\sum_{j=N-b+1}^{N}\sum_{\mu=1}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)}, \text{ which yields,} \\
&= \frac{1}{N}\sum_{j=2}^{N-b}\xi_1^{(1)}\xi_1^{(1)}\xi_j^{(1)}\xi_j^{(1)} + \frac{1}{N}\sum_{j=2}^{N-b}\sum_{\mu=2}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)} \\
&\quad - \frac{1}{N}\sum_{j=N-b+1}^{N}\xi_1^{(1)}\xi_1^{(1)}\xi_j^{(1)}\xi_j^{(1)} - \frac{1}{N}\sum_{j=N-b+1}^{N}\sum_{\mu=2}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)}, \\
&\quad \text{on separating the } \mu = 1 \text{ terms from the rest, and,} \\
&= \frac{1}{N}\left(N - b - 2 - (N - (N - b + 1))\right) \\
&\quad + \frac{1}{N}\sum_{j=2}^{N-b}\sum_{\mu=2}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)} - \frac{1}{N}\sum_{j=N-b+1}^{N}\sum_{\mu=2}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)}, \\
&\quad \text{since } \xi_1^{(1)}\xi_1^{(1)} = 1, \text{ and hence} \\
\xi_1^{(1)} h_1 &= \underbrace{\frac{N - 2b - 1}{N}}_{\text{signal}} + \\
&\quad \underbrace{\frac{1}{N}\sum_{j=2}^{N-b}\sum_{\mu=2}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)} - \frac{1}{N}\sum_{j=N-b+1}^{N}\sum_{\mu=2}^{p}\xi_1^{(1)}\xi_1^{(\mu)}\xi_j^{(\mu)}\xi_j^{(1)}}_{\text{noise}}.
\end{aligned} \tag{2.20}$$

We have separated the terms into a signal and a noise term. The noise term is a random walk whose constituent terms are uncorrelated, and can be evaluated as $(N-1)(p-1)$. When N and p are large, the 1 term can be neglected in both the signal and noise terms. So, we get an expression for the signal-to-noise ratio (SNR) as,

$$SNR = \frac{(N - 2b)/N}{\sqrt{p/N}} = \frac{N - 2b}{\sqrt{pN}}. \tag{2.21}$$

Now, to estimate the radius of the basin of attraction, we make the $SNR = \alpha_c$ (from eq.(2.16)) to get,

$$\begin{aligned} \frac{N-2b}{\sqrt{p/N}} &= \frac{1}{\alpha_c} \\ b &= -\frac{1}{2}\left(\frac{\sqrt{pN}}{\alpha_c} - N\right) \\ &= \frac{N}{2}\left(1 - \frac{\sqrt{pN}}{N\sqrt{\alpha_c}}\right) \\ b &= \frac{N}{2}\left(1 - \sqrt{\frac{p}{\alpha_c N}}\right). \end{aligned} \tag{2.22}$$

From the above analysis, we can see that the basin of attraction can be at most $N/2$ for very small values of p, and when p/N approaches α_c, the radius becomes 0, that is, the basin of attraction vanishes. We can hence see that the catastrophic blackout (discussed previously) also marks the limit where the basin of attraction surrounding each of the inscribed patterns disappears. This is due to the destructive interference (between the basins) resulting from the crosstalk between the stored patterns [10].

2.6 Conclusion

In this chapter, we have presented a brief overview of the Hopfield model. In the forthcoming chapters, we will build our model based on this network. We will also revisit the concept of pattern stability and elaborate further on some aspects of the network dynamics described in this chapter.

Chapter 3

Gram-Schmidt orthogonalization: a solution to Catastrophic Interference

In this chapter, we invoke Gram-Schmidt orthogonalization in the Hopfield model as a way of overcoming the detrimental effects of catastrophic interference between the patterns to improve the memory capacity of the network. We discover that it has profound biological implications on cognitive learning and memory.

3.1 Introduction

Learning and memory require a system to possess two properties - stability and plasticity. The system must be plastic or malleable in order to learn information, while at the same time remain stable and retain information in the presence of changes due to newer information being encountered by the system. These contradictory requirements consitute the stability-plasticity dilemma[11]. Neurobiologists must address this issue while studying the functioning of the nervous system, while in artificial intelligence research, the problem bears relevance in building and understanding memory systems. At the theoretical level, the stability-plasticity dilemma prompts the following question: how can the elements of a system store new information without affecting what has already been memorized?

In the nervous system, memory is stored in the synapses following the principles postulated by Donald Hebb[2]. Long-term potentiation of the synapses has been proposed as the primary mechanism underlying learning and memory[12]. The Bienenstock-Cooper-Munro (BCM) model [13] pointed out that this mechanism could suffer from an inherent instability- in a system whose synapses have a threshold for plasticity, the entry of new information causes a growth in the synapses which are above the threshold, while those below it decay. The growth or decay pertains to the patterns rather than to the synapses, resulting in certain patterns being favoured over others. As the same set of synapses is used to store more than one memory, there could be further growth or potentiation of synapses even in response to non-favoured patterns. The potentiation (or depression) due to patterns other than a specific pattern being stored contributes to the "ongoing plasticity"[22]. This could result in runaway cycles of potentiation or depression which would disrupt the information already in the memory. Such an unhindered increase in synaptic potentiation could lead to excitoxicity, cell death and epileptogenesis in biological systems[25].

A similar phenomenon is observed in connectionist neural networks. In artificial neural networks, as more and more patterns are stored, we reach a critical point beyond which there is a sudden abrupt, drastic and complete loss of memory, a blackout. This blackout is due to Catastrophic Interference (CI). Any information learnt by the system modifies the same set of synaptic weights, while the network still retains the information pertaining to the patterns memorized earlier; but this is possible only for a small set of patterns before CI sets in. CI is thus a manifestation of the stability-plasticity dilemma.

The set of patterns in the memory store overlap with each other. The presence of correlations between the patterns indicates that they share several common features and are similar to each other. (Refer to [15] for a mathematical interpretation.) These correlations result in further potentiation or depression of the same synapses. This is akin to the concept of stability described earlier. While newer memories in the human brain are labile and susceptible to change[26, 27], there is no abrupt or complete deterioration in the retrieval of earlier memories. Any degradation of prior memorized information happens gradually and "gracefully"[28, 29, 30, 31]. The catastrophic blackout is hence a limiting feature of connectionist networks which rely on modifications to the same finite set of synapses to store multiple memories. In contrast, the human brain is inherently capable of learning new information without

any drastic loss of earlier memories. (But, see [32] for an example of retroactive interference in humans wherein newly learnt information affects the retrieval of previously stored memories.) It would hence be useful to understand the mechanisms underlying human memory and explore their implementation in and relevance to artificial intelligence networks.

A number of ways of overcoming the detrimental effects of CI have been proposed. For instance, the usage of a cascade of states and including the strength of synapses[33], updating only a fraction of weight-restricted synapses for specific rates of information presentation[34] or dual systems, one for learning and another for retention[35]. Other strategies entail implementing neurogenesis i.e., the generation of new neurons [36] or following an anti-Hebbian learning prescription[37] to offset the negative effects of CI. Another biologically plausible approach involves having the synaptic efficacies traverse between a number of states which are bounded on either side[22]. The modifications to the efficacies depend on the current strength of the synapse, reflecting the so-called "soft bound plasticity"[38]. However, these proposed strategies are limited in scope of implementation or biological feasibility.

Given the commonalities between biological systems and artificial neural systems in terms of stability and plasticity, it would be useful and interesting to apply the strategies to overcome CI to a deeper analysis of the runaway cycle of potentiation. The BCM model proposes a threshold for synaptic modifications that is dependent on the current and previous states of the synapse[39]. This dependence represents the plasticity of the (plastic) synapse, the "metaplasticity"[40]. The "sliding threshold" regulates synaptic changes due to LTP and LTD and brings about homeostasis in the system by making the synapses stable and impervious to disturbances[40, 41]. A synapse grows following LTP or decays after LTD depending on whether or not the threshold is crossed. Crossing the threshold and memorizing a pattern require sufficient amount of activity on the synapses, thus ensuring that the changes due to the non-pertinent synapses in the system fall below the threshold without modifying all the synapses considerably.

The BCM model offers a means of limiting synaptic efficacies within bounds, making it a biologically tenable neural network model. It would be useful to establish a link between experimental results pertaining to cellular and synaptic processes and network function. While the presence of a shifting plasticity has been observed, there is no evidence yet of a BCM-like effect in human memory, nor has the model been studied with relation to CI in artificial neural networks. We implement Gram-Schmidt

orthogonalization in a Hopfield model following Hebbian learning (hereafter referred to as H-H-GS model) and show that the network then becomes capable of automatically containing the creation of runaway potentiation cycles, thereby effectively eliminating CI.

In our study, we use a basic network model which only incorporates the minimum essential features and underlying conditions of the system we wish to explore. While it does not appear biologically plausible at the outset, it can be modified to incorporate knowledge gained from experimental observations. The model captures the basic properties of our system of interest, as its components are capable of synapse-like encoding, and the network captures the essence of the phenomena we wish to understand. The network can be generalized provided it meets certain criteria (discussed later). We believe that the results from our current study may pertain to real observed phenomena. The model can also be generalized and brought closer to biology by systematically including more features, while also ensuring mathematical tractability at each step. The model should in principle then be capable of producing results similar to those observed experimentally.

3.2 Catastrophic interference

We adopt the Hebb-Hopfield model as the base for our study. The network was described in some detail in the previous chapter, and we saw that the network had a low memory capacity. When patterns are stored in the network, there is a sudden and complete loss of memory beyond a certain limit. This memory blackout is the result of the catastrophic interference discussed above. In order to understand the problem better, we perform an alternate signal-to-noise analysis of the LFP, the local field potential (discussed in 2.3 of the previous chapter). We have seen previously that the LFP or post-synaptic potential can be split into a signal and a noise term, resulting in the catastrophic blackout when the noise accumulates and overrides the signal. We now revisit the LFP and evaluate eq.(2.4) by substituting for J_{ij} from eq.(2.3) which can be rewritten as:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \left(\xi_i^{(\mu)} \xi_j^{(\mu)} - \delta_{ij} \xi_i^{(\mu)} \xi_i^{(\mu)} \right), \tag{3.1}$$

where the Kronecker delta function $\delta_{ij} = 1$ for $i = j$ and $= 0$ otherwise. The above learning rule was originally postulated by Cooper[14] to mimic Hebbian synaptic plasticity. We then have,

$$h_i^{(\nu)} = \sum_{j=1}^{N} \left[\frac{1}{N} \sum_{\mu=1}^{p} \left(\xi_i^{(\mu)} \xi_j^{(\mu)} - \delta_{ij} \xi_i^{(\mu)} \xi_j^{(\mu)} \right) \right] \xi_j^{(\nu)}. \tag{3.2}$$

We can express the above equation as:

$$\begin{aligned} h_i^{(\nu)} &= \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{(\mu)} \left[\left(\boldsymbol{\xi}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right) - \xi_i^{(\mu)} \xi_i^{(\nu)} \right] \\ &= \frac{1}{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{p} \xi_i^{(\mu)} \left[\left(\boldsymbol{\xi}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right) \right] + \left[\xi_i^{\nu} - \sum_{\mu=1}^{p} \xi_i^{(\mu)} \xi_i^{(\mu)} \xi_i^{(\nu)} \right] \\ h_i^{(\nu)} &= \left(1 - \frac{p}{N} \right) \xi_i^{(\nu)} + \frac{1}{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{p} \xi_i^{(\mu)} \left(\boldsymbol{\xi}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right). \end{aligned} \tag{3.3}$$

Here again, the first term represents the signal due to the inscribed pattern $\boldsymbol{\xi}^{(\nu)}$, while the second term denotes the noise. The correlations between the stored patterns, the *crosstalk* contribute to the noise in the system.

3.3 Overcoming the catastrophic blackout

The addition of new information leads to an increase in the noise in the system which eventually becomes so high that the signals due to the memorized patterns get submerged in it, while the basins of attraction shrink or vanish altogether. From the previous section, we see that the overlaps between the patterns are the source of noise which leads to the catastrophic interference. In order to overcome this, we try to reduce or remove the correlations between the inscribed patterns. A simple way of doing so is by incorporating the Gram-Schmidt orthogonalization procedure in the memorization process[15, 16].

3.3.1 Gram-Schmidt orthogonalization

For a set of linearly independent vectors $\{\boldsymbol{\xi}^{(\mu)}\} = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(p)}\}$, the Gram-Schmidt procedure yields a set $\{\boldsymbol{\eta}^{(\mu)}\} = \{\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \ldots, \boldsymbol{\eta}^{(p)}\}$ whose elements are mutually orthogonal. The orthogonalization process is done by removing the projections of all the other vectors on each individual $\boldsymbol{\xi}$. This can be expressed mathematically as,

$$\boldsymbol{\eta}^{(\nu)} = \boldsymbol{\xi}^{(\nu)} - \sum_{\mu=1}^{\nu-1} \boldsymbol{\eta}^{(\mu)} \frac{\boldsymbol{\eta}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)}}{\boldsymbol{\eta}^{(\mu)} \cdot \boldsymbol{\eta}^{(\mu)}}, \tag{3.4}$$

where ν is the pattern index which runs from 1 to p. The second term on the right hand side gives the projection of $\boldsymbol{\xi}^{(\nu)}$ on $\boldsymbol{\eta}^{(1)}$ to $\boldsymbol{\eta}^{(\nu-1)}$. We normalize $\boldsymbol{\eta}^{(\nu)}$ to yield $\hat{\boldsymbol{\eta}}^{(\nu)}$ whose components are denoted by $\underset{\sim}{\eta}^{(\nu)}$'s. $\hat{\boldsymbol{\eta}}^{(\nu)}$ is obtained by calculating $\hat{\boldsymbol{\eta}}^{(\nu)} = \frac{\boldsymbol{\eta}^{(\nu)}}{\|\boldsymbol{\eta}^{(\nu)}\|}$, where $\|\boldsymbol{\eta}^{(\nu)}\|$ gives the norm of the vector.

We first study the case of $p = 2$, where 2 patterns $\boldsymbol{\xi}^{(1)}$ and $\boldsymbol{\xi}^{(2)}$ are orthonormalized following the Gram-Schmidt procedure to yield $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\eta}^{(2)}$. We first verify whether $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\eta}^{(2)}$ are mutually orthogonal.

$$\begin{aligned} \boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(1)} &= \left(\boldsymbol{\xi}^{(2)} - \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(2)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} \right) \cdot \boldsymbol{\eta}^{(1)} \\ &= \boldsymbol{\xi}^{(2)} \cdot \boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(2)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} \\ &= \boldsymbol{\xi}^{(2)} \cdot \boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(2)} \end{aligned} \tag{3.5}$$

$$\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(1)} = 0. \tag{3.6}$$

Now, we add a third vector, $\boldsymbol{\xi}^{(3)}$ to the network following the Gram-Schmidt procedure, and obtain $\boldsymbol{\eta}^{(3)}$. We can now verify if this new third pattern is indeed orthogonal to the previous two. We have:

$$\begin{aligned} \boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\eta}^{(2)} &= \left(\boldsymbol{\xi}^{(3)} - \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} - \boldsymbol{\eta}^{(2)} \frac{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)}} \right) \cdot \boldsymbol{\eta}^{(2)} \\ &= \boldsymbol{\xi}^{(3)} \cdot \boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} - \boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)} \frac{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)}} \\ &= \boldsymbol{\xi}^{(3)} \cdot \boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)} - 0 \end{aligned} \tag{3.7}$$

$$\boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\eta}^{(2)} = 0, \tag{3.8}$$

and,

$$\begin{aligned}
\boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\eta}^{(1)} &= \left(\boldsymbol{\xi}^{(3)} - \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} - \boldsymbol{\eta}^{(2)} \frac{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)}} \right) \cdot \boldsymbol{\eta}^{(2)} \qquad (3.9) \\
&= \boldsymbol{\xi}^{(3)} \cdot \boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} - \boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)} \frac{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)}} \\
&= \boldsymbol{\xi}^{(3)} \cdot \boldsymbol{\eta}^{(1)} - \boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)} - 0 \\
\boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\eta}^{(1)} &= 0, \qquad (3.10)
\end{aligned}$$

as $\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(2)} = 0$. This can be generalized to all the $\boldsymbol{\eta}$'s. Having established their mutual orthogonality, we now study the projections of the $\boldsymbol{\xi}$'s on the $\boldsymbol{\eta}$'s to confirm that $\boldsymbol{\xi}^{(\nu)}$ does not project onto $\boldsymbol{\eta}^{(\mu)}$ for $\mu > \nu$.

$$\begin{aligned}
\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(1)} &= \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(1)} \\
\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(2)} &= \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(2)} - \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(2)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} \\
&= \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(2)} - \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(2)} \\
\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(2)} &= 0. \\
\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(3)} &= \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(3)} - \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(1)} \frac{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)}} - \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(2)} \frac{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)}} \\
&= \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(3)} - \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(1)} \frac{\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(3)}}{\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(1)}} \\
&= \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(3)} - \boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(3)} \\
\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\eta}^{(3)} &= 0. \qquad (3.11)
\end{aligned}$$

We see that $\boldsymbol{\xi}^{(1)}$ projects onto $\boldsymbol{\eta}^{(1)}$, but not $\boldsymbol{\eta}^{(2)}$ or $\boldsymbol{\eta}^{(3)}$. This result can be extrapolated amd generalized to show that $\boldsymbol{\xi}^{(\nu)}$ projects onto $\boldsymbol{\eta}^{(\mu)}$ only when $\nu \leq \mu$. (Note that the above holds true only for orthonormalized vectors.)

3.3.2 An example of the orthogonalization process

We now provide an example of the Gram-Schmidt orthogonalization procedure. Table3.1 shows the orthogonal and orthonormal bases for a set of 5 10−dimensional

random vectors $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(5)}$, the input patterns.

TABLE 3.1: Table showing the orthogonal($\boldsymbol{\eta}$'s) and orthonormal($\hat{\boldsymbol{\eta}}$'s) bases of the input patterns ($\boldsymbol{\xi}$'s) obtained using GS orthogonalization. The $\boldsymbol{\xi}$'s are a set of 5 randomly generated patterns whose elements are ± 1. The number of similarities and differences between the input pattern pairs are represented by S and D respectively.

Input patterns										
$\boldsymbol{\xi}^{(1)}$	1	-1	-1	1	-1	1	1	-1	1	-1
$\boldsymbol{\xi}^{(2)}$	1	-1	-1	1	-1	-1	-1	-1	1	1
$\boldsymbol{\xi}^{(3)}$	1	1	1	1	-1	-1	-1	1	-1	-1
$\boldsymbol{\xi}^{(4)}$	-1	-1	1	1	1	1	1	1	1	-1
$\boldsymbol{\xi}^{(5)}$	1	1	-1	-1	1	1	-1	-1	1	-1

$\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(2)} = 4$ $7S/3D$ | $\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(3)} = -2$ $4S/6D$ | $\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(4)} = 2$ $6S/4D$ | $\boldsymbol{\xi}^{(1)} \cdot \boldsymbol{\xi}^{(5)} = 2$ $6S/4D$

$\boldsymbol{\xi}^{(2)} \cdot \boldsymbol{\xi}^{(3)} = 0$ $5S/5D$ | $\boldsymbol{\xi}^{(2)} \cdot \boldsymbol{\xi}^{(4)} = -4$ $3S/7D$ | $\boldsymbol{\xi}^{(2)} \cdot \boldsymbol{\xi}^{(5)} = 0$ $5S/5D$

$\boldsymbol{\xi}^{(3)} \cdot \boldsymbol{\xi}^{(4)} = -2$ $4S/6D$ | $\boldsymbol{\xi}^{(3)} \cdot \boldsymbol{\xi}^{(5)} = 2$ $6S/4D$

$\boldsymbol{\xi}^{(4)} \cdot \boldsymbol{\xi}(5) = -2$ $4S/6D$

Orthogonalized patterns										
$\boldsymbol{\eta}^{(1)}$	1	−1	−1	1	−1	1	1	−1	1	−1
$\boldsymbol{\eta}^{(2)}$	0.6	−0.6	−0.6	0.6	−0.6	−1.4	−1.4	−0.6	0.6	1.4
$\boldsymbol{\eta}^{(3)}$	1.1429	0.8571	0.8571	1.1429	−1.1429	−0.6667	−0.6667	0.8571	−0.8571	−1.3330
$\boldsymbol{\eta}^{(4)}$	−0.7149	−1.0399	0.9601	1.2801	0.7199	−0.0801	−0.0801	0.9601	1.0399	−0.1601
$\boldsymbol{\eta}^{(5)}$	0.7235	0.8095	−0.2856	−0.3709	1.2757	0.5238	−1.4762	−0.2856	1.1905	−0.9524

$\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\eta}^{(1)} = 10$	$\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\eta}^{(2)} = 8.4$	$\boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\eta}^{(3)} = 9.5231$	$\boldsymbol{\eta}^{(4)} \cdot \boldsymbol{\eta}^{(4)} = 6.72$	$\boldsymbol{\eta}^{(5)} \cdot \boldsymbol{\eta}^{(5)} = 7.8847$

Orthonormalized patterns										
$\hat{\boldsymbol{\eta}}^{(1)}$	0.3162	−0.3162	−0.3162	0.3162	−0.3162	0.3162	0.3162	−0.3162	0.3162	−0.3162
$\hat{\boldsymbol{\eta}}^{(2)}$	0.2070	−0.2070	−0.2070	0.2070	−0.2070	−0.4831	−0.4831	−0.2070	0.2070	0.4831
$\hat{\boldsymbol{\eta}}^{(3)}$	0.3704	0.2777	0.2777	0.3704	−0.3704	−0.2160	−0.2160	0.2777	−0.2777	−0.4320
$\hat{\boldsymbol{\eta}}^{(4)}$	−0.2777	−0.4012	0.3704	0.4938	0.2777	−0.0309	−0.0309	0.3704	0.4012	−0.0618
$\hat{\boldsymbol{\eta}}^{(5)}$	0.2577	0.2883	−0.1017	−0.1321	0.4543	0.1865	−0.5257	−0.1017	0.4240	−0.3392

$\hat{\boldsymbol{\eta}}^{(1)} \cdot \hat{\boldsymbol{\eta}}^{(1)} = 1$	$\hat{\boldsymbol{\eta}}^{(2)} \cdot \hat{\boldsymbol{\eta}}^{(2)} = 0.9995$	$\hat{\boldsymbol{\eta}}^{(3)} \cdot \hat{\boldsymbol{\eta}}^{(3)} = 0.9996$	$\hat{\boldsymbol{\eta}}^{(4)} \cdot \hat{\boldsymbol{\eta}}^{(4)} = 1.001$	$\hat{\boldsymbol{\eta}}^{(5)} \cdot \hat{\boldsymbol{\eta}}^{(5)} = 1$

$\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(2)} = 4$ | $\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(3)} = -2$ | $\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(4)} = 2$ | $\boldsymbol{\eta}^{(1)} \cdot \boldsymbol{\xi}^{(5)} = 2$

$\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(3)} = 0.8$ | $\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(4)} = -4.8$ | $\boldsymbol{\eta}^{(2)} \cdot \boldsymbol{\xi}^{(5)} = -0.8$

$\boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\xi}^{(4)} = -1.1433$ | $\boldsymbol{\eta}^{(3)} \cdot \boldsymbol{\xi}^{(5)} = -1.5241$

$\boldsymbol{\eta}^{(4)} \cdot \boldsymbol{\xi}^{(5)} = -3.0402$

It has been proposed that the brain might use orthogonalization as a way of classifying information in an economical manner, by emphasizing the differences between information that are similar to each other, and the similarities in patterns that are very different from each other, and storing this information[15]. For instance, in the above table, $\boldsymbol{\xi}^{(4)}$ is dissimilar to $\boldsymbol{\xi}^{(2)}$ and $\boldsymbol{\xi}^{(3)}$ but shares more similarities with $\boldsymbol{\xi}^{(1)}$, while all 4 patterns have one element in common (4^{th} element). The 7^{th} element, $\xi_7^{(4)}$ is similar to $\xi_7^{(2)}$ and $\xi_7^{(3)}$ but dissimilar to $\xi_7^{(1)}$. These factors are reflected in the magnitudes of $\underset{\sim}{\eta}_7^{(4)}$ and $\underset{\sim}{\eta}_4^{(4)}$. In fact, $\underset{\sim}{\eta}_4^{(4)}$ has the maximum magnitude, emphasizing the shared commonality between all 4 patterns.

3.3.3 Memory capacity of the H-H-GS model

We now estimate the memory capacity of the H-H-GS model using the set of orthogonal vectors $\{\boldsymbol{\eta}^{(\mu)}\}$ instead of $\{\boldsymbol{\xi}^{(\mu)}\}$ to calculate J_{ij}'s and hence to inscribe patterns $\{\boldsymbol{\xi}^{(\mu)}\}$ in the network.

$$J_{ij} = \sum_{\mu=1}^{p} \left(\eta_i^{(\mu)} \eta_j^{(\mu)} - \delta_{ij} \eta_i^{(\mu)} \eta_i^{(\mu)} \right), \tag{3.12}$$

where δ_{ij} serves to replace the $i \neq j$ condition in the summation in eq.(2.3).

With these weights, we can calculate the local field potential $h_i^{(\nu)}$ when $\{\boldsymbol{\xi}^{(\nu)}\}$ is presented for retrieval (i.e., $\{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(p)}\}$ are imprinted in their orthogonalized form.

$$\begin{aligned} h_i^{(\nu)} &= \sum_{j=1}^{N} J_{ij} \xi_j^{(\nu)} \\ &= \sum_{j=1}^{N} \sum_{\mu=1}^{p} \left(\eta_i^{(\mu)} \eta_j^{(\mu)} \xi_j^{(\nu)} - \delta_{ij} \eta_i^{(\mu)} \eta_i^{(\mu)} \xi_i^{(\nu)} \right), \text{or,} \\ h_i^{(\nu)} &= \sum_{\mu=1}^{p} \eta_i^{(\mu)} \left\{ \left(\boldsymbol{\eta}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right) - \eta_i^{(\mu)} \xi_i^{(\nu)} \right\}. \end{aligned} \tag{3.13}$$

From eq.(3.4), we have

$$\eta_i^{(\nu)} = \xi_i^{(\nu)} - \sum_{\mu=1}^{\nu-1} \eta_i^{(\mu)} \frac{\boldsymbol{\eta}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)}}{\boldsymbol{\eta}^{(\mu)} \cdot \boldsymbol{\eta}^{(\mu)}}.$$

For orthonormal vectors, $\hat{\boldsymbol{\eta}}^{(\mu)} \cdot \hat{\boldsymbol{\eta}}^{(\mu)} = 1$, and so we get

$$\eta_i^{(\nu)} = \xi_i^{(\nu)} - \sum_{\mu=1}^{\nu-1} \underset{\sim}{\eta_i}^{(\mu)} \left(\hat{\boldsymbol{\eta}}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right). \tag{3.14}$$

Rearranging the terms in eq.(3.14), we get

$$\xi_i^{(\nu)} - \eta_i^{(\nu)} = \sum_{\mu=1}^{\nu-1} \underset{\sim}{\eta_i}^{(\mu)} \left(\hat{\boldsymbol{\eta}}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right). \tag{3.15}$$

As verified earlier, $\boldsymbol{\xi}^{(\nu)}$ does not project onto $\boldsymbol{\eta}^{(\mu)}$ for $\mu > \nu$, so we can rewrite eq.(3.13) using normalized vectors as

$$h_i^{(\nu)} = \sum_{\mu=1}^{\nu} \eta_i^{(\mu)} \left\{ \hat{\boldsymbol{\eta}}^{(\boldsymbol{\mu})} \cdot \boldsymbol{\xi}^{(\nu)} \right\} - \sum_{\mu=1}^{p} \eta_i^{(\mu)} \left(\eta_i^{(\mu)} \xi_i^{(\nu)} \right). \tag{3.16}$$

From (3.14),

$$\boldsymbol{\eta}^{(\mu)} = \boldsymbol{\xi}^{(\nu)} - \sum_{\mu=1}^{\nu-1} \left(\hat{\boldsymbol{\eta}}^{(\boldsymbol{\mu})} \cdot \boldsymbol{\xi}^{(\nu)} \right)$$

Normalizing $\boldsymbol{\eta}^{(\nu)}$ and multiplying it by $\hat{\boldsymbol{\eta}}^{(\boldsymbol{\mu})}$, we get

$$\hat{\boldsymbol{\eta}}^{(\nu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} = \boldsymbol{\xi}^{(\mu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} - \sum_{\mu=1}^{\nu-1} \left(\hat{\boldsymbol{\eta}}^{(\mu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} \right) \left(\hat{\boldsymbol{\eta}}^{(\mu)} \cdot \boldsymbol{\xi}^{(\nu)} \right). \tag{3.17}$$

As $\hat{\boldsymbol{\eta}}^{(\mu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} = 0$ for $\mu \leq \nu$, the second term on the right hand size of the equation vanishes to give

$$\hat{\boldsymbol{\eta}}^{(\nu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} = \boldsymbol{\xi}^{(\mu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)}. \tag{3.18}$$

This holds true for any $\boldsymbol{\xi}^{(\mu)}$ as long as $\boldsymbol{\eta}$'s are normalized.

Now, plugging in the values from equations (3.15) and (3.18) in eq.(3.16), we get

$$h_i^{(\nu)} = \xi_i^{(\nu)} - \eta_i^{(\nu)} + \eta_i^{(\nu)} \left(\hat{\boldsymbol{\eta}}^{(\nu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} \right) - \sum_{\mu=1}^{p} \left(\underset{\sim}{\eta}_i^{(\mu)} \right)^2 \xi_i^{(\nu)}. \tag{3.19}$$

Since $\hat{\boldsymbol{\eta}}^{(\nu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)} = 1$, we get

$$\begin{aligned} h_i^{(\nu)} &= \xi_i^{(\nu)} - \eta_i^{(\nu)} + \eta_i^{(\nu)} - \sum_{\mu=1}^{p} \left(\underset{\sim}{\eta}_i^{(\mu)} \right)^2 \xi_i^{(\nu)} \\ h_i^{(\nu)} &= \xi_i^{(\nu)} - \sum_{\mu=1}^{p} \left(\underset{\sim}{\eta}_i^{(\mu)} \right)^2 \xi_i^{(\nu)}. \end{aligned} \tag{3.20}$$

For orthonormal vectors, $\left(\underset{\sim}{\eta}_i^{(\mu)} \right)^2 \approx \frac{1}{N}$. Hence,

$$\boldsymbol{h}^{(\nu)} = \left(1 - \mathcal{O}\left(\frac{p}{N} \right) \right) \boldsymbol{\xi}^{(\nu)}. \tag{3.21}$$

We can thus see that though it is the $\boldsymbol{\eta}$'s that are stored in the network, the $\boldsymbol{\xi}$'s can actually be retrieved perfectly[17]. We see that the stability condition (2.9) is satisfied upto $p = N - 1$, as there can be at most $N - 1$ orthogonal vectors

of dimension N. Fig.3.1 shows stable patterns in the Hopfield model with Gram-Schmidt orthogonalization.

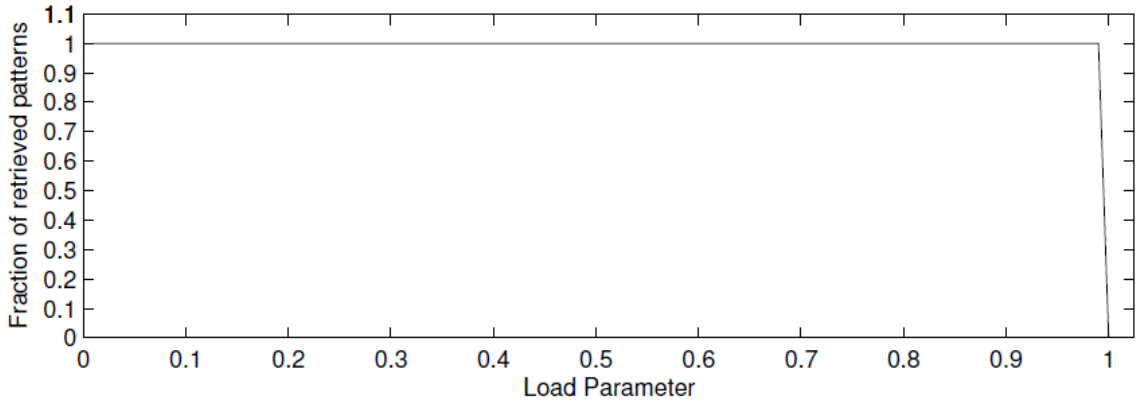


FIGURE 3.1: Plot showing stable patterns in the H-H-GS model from our computations with 100 neurons after 50 trials with 100 patterns each, plotted as fraction of retrieved patterns vs the load parameter (given by p/N). All the patterns are stable upto $p = 99$.

The memory capacity of a 1000−neuron H-H-GS network is shown in Fig.3.2, with retrieval being checked with the raw patterns ($\boldsymbol{\xi}$'s) while the orthonormalized set$\{\hat{\boldsymbol{\eta}}\}$ is used for storage. All the patterns are retrieved perfectly upto $p = 998$, with a sharp decrease in retrieval for $p = 999$. At $p = 1000$, the retrieval falls completely to zero.

3.4 Studying the post-synaptic potential (PSP)

3.4.1 Analyzing the PSP

We now return our focus to the PSP. We have seen that the second term of eq.(2.4) (hereafter denoted by $\mathcal{A}$) constitutes a random walk whose components are fractions and < 1. The values of p and N determine the bounds of the range of values $\mathcal{A}$ can take. While $\mathcal{A}$ can take any value within this range, the signs of the terms in the stability condition (eq.(2.9)) match provided $\mathcal{A} < (1 - p/N)$. The shaded regions in Fig.3.3 show the values of $\mathcal{A}$ for which the condition holds true.

As discussed earlier, the $\boldsymbol{\xi}^{(\nu)}$'s are correlated with each other as they are randomly generated and not necessarily mutually orthogonal. This implies that their dot products are non-zero, and can therefore take arbitrarily large positive or negative values

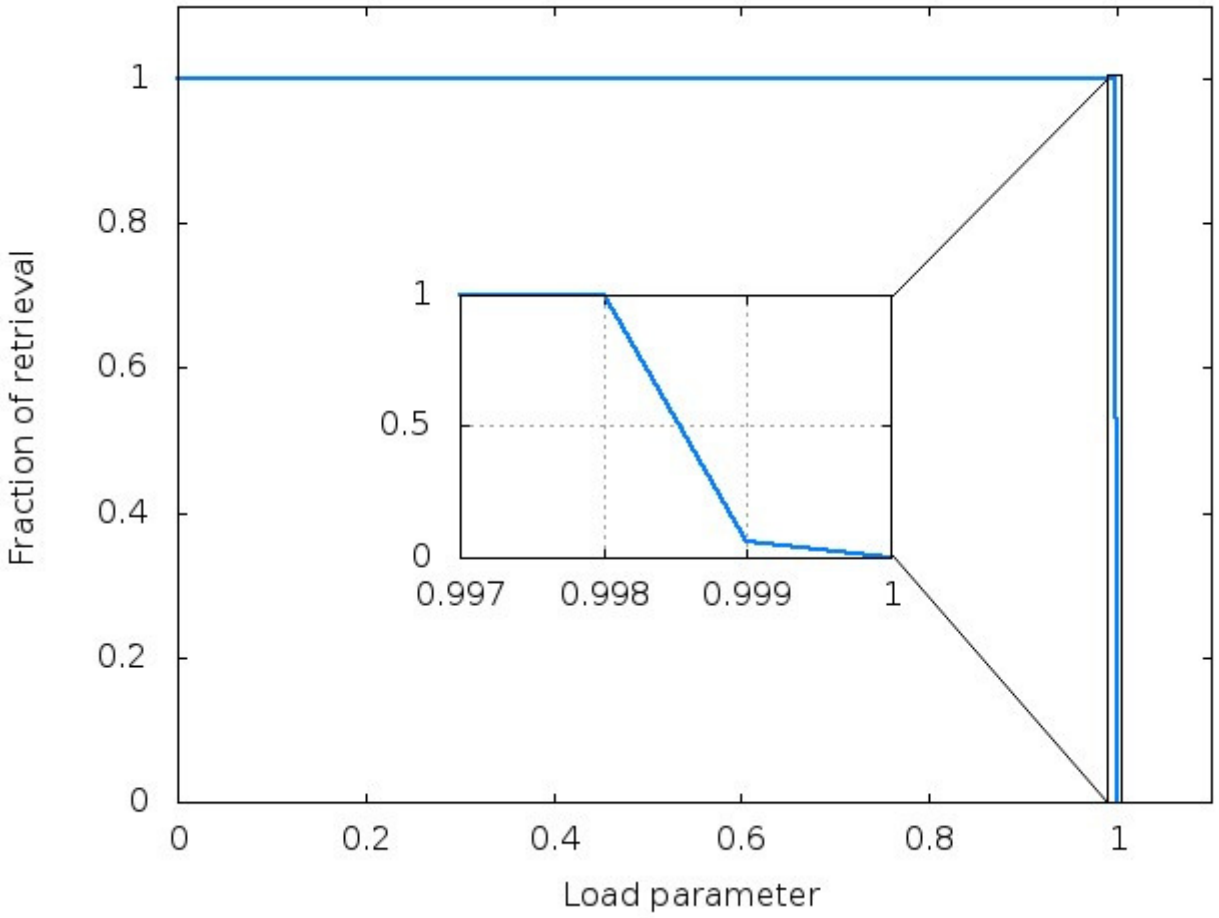


FIGURE 3.2: Plot from our compuations showing the retrieval of patterns in the H-H-GS model with 1000 neurons after 50 trials of 1000 patterns, plotted as fraction of retrieved patterns against load parameter. All the patterns are retrieved as long as $p < N$.

for high enough values of p. This leads to a higher likelihood of the occurence of CI. The reason for this is as follows:

We can see from eq.(3.3) that the PSP on neuron i can be expressed in terms of a signal and the noise $\mathcal{A}$. The signal term is obtained by distinguishing the contribution of the pattern presented to the network for retrieval, $\boldsymbol{\xi}^{(\nu)}$ from the overlaps of $\boldsymbol{\xi}^{(\nu)}$ with every other pattern. These non-zero overlaps obscure the signal and together constitute the noise term $\mathcal{A}$. It follows that the PSP will lie between $1 - p/N$ and $p/N - 1$ as long as $\mathcal{A}$ falls below $1 - p/N$ and above $p/N - 1$, indicating that CI is suppressed. However, as p increases, the correlations of the newer patterns with those stored previously add to the noise, as the Hopfield network lacks a default mechanism to contain these overlaps and hence limit the noise to the favourable range mentioned

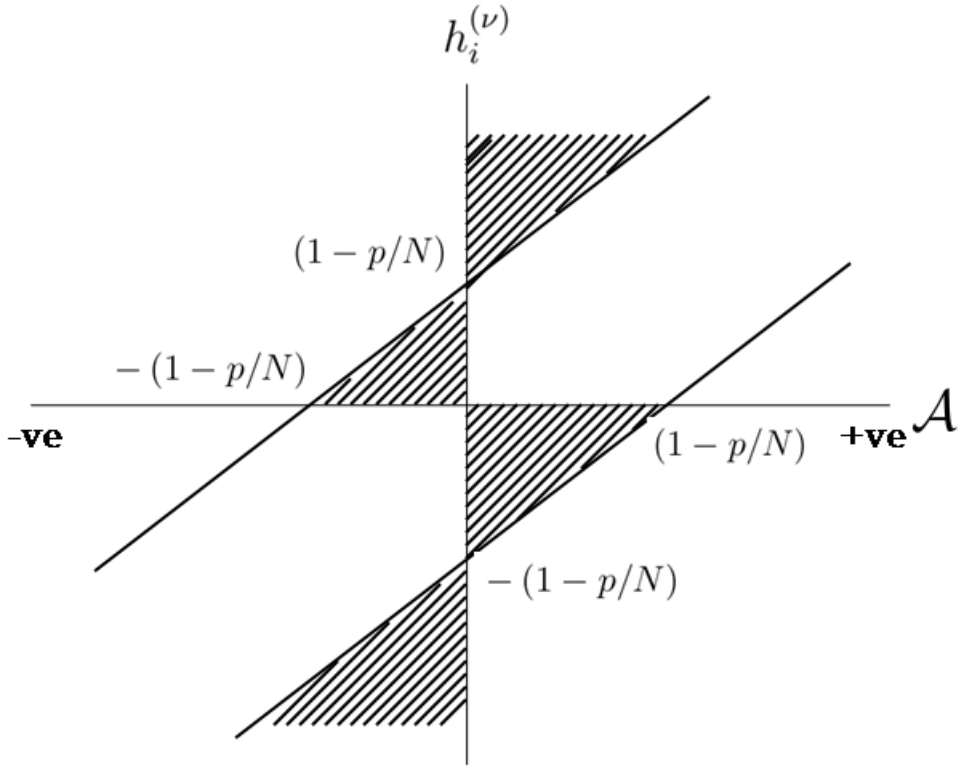


FIGURE 3.3: Schematic representation of the post-synaptic potential $h_i^{(\nu)}$ on a random neuron i on presenting pattern $\boldsymbol{\xi}^{(\nu)}$ to the network for retrieval, versus , the noise term in eq.(3.3). $\boldsymbol{\xi}^{(\nu)}$ is one of the learnt patterns chosen at random. The shaded areas represent the regions where $\mathcal{A}$ will be positive definite. The bounds on slide up and down with variations in p and N enabling, at least in principle, plasticity to control CI to some extent.

above. As more patterns are learnt by the network, there is an increase in the number of sites i where eq.(2.9) no longer holds true, eventually and inevitably resulting in CI. The values of p and N determine the limits within which $h_i^{(\nu)}$ satisfies the stability criterion, thereby making the network more or less prone to the detrimental effects of CI. For a network of fixed size N, high values of p would cause the limits to shrink, making the system more prone to CI. On the other hand, if the size of the network is increased such that $p/N \to 0$, the distance between the bounds gets extended, making CI less likely.

But, $\mathcal{A}$ can, in principle take very large values (both positive and negative) beyond the bounds. This is comparable to the runaway effect in the BCM model[13] discussed earlier. While the stability condition is still satisfied, the changes in the value of $\mathcal{A}$ are (apparently) arbitrary and unrestrained. Such a growth leads to CI and

affects the retrieval of information stored in the memory[42]. It results in the run-away phenomenon and even leads to false and incorrect associations with the feature represented by the neuron i. The unchecked growth of PSP on multiple neurons i eventually leads to the catastrophic blackout discussed earlier in Sec. 2.4.

3.4.2 Gram-Schmidt orthogonalization and PSP

Now consider a Hopfield network in which p patterns have been stored. On presenting a new $(p+1)^{th}$ pattern ($\boldsymbol{\xi}^{(p+1)}$) to the network, the orthogonalized pattern, $\boldsymbol{\eta}^{(p+1)}$ is obtained from the Gram-Schmidt orthogonalization scheme as:

$$\eta_i^{(p+1)} = \xi_i^{(p+1)} - \sum_{\mu=1}^{p} \eta_i^{(\mu)} \frac{\sum_{j=1}^{N} \eta_j^{(\mu)} \xi_j^{(p+1)}}{\sum_{j=1}^{N} \eta_j^{(\mu)} \eta_j^{(\mu)}}. \tag{3.22}$$

The system compares the new incoming pattern $\boldsymbol{\xi^{(p+1)}}$ with all the previously stored patterns. Taking eq.(3.2) into consideration, these differences get computed through the PSP as $\boldsymbol{\xi}^{(p+1)} - \boldsymbol{h}^{(p+1)}$ on each individual neuron i. This computation is equivalent to the process of orthogonalization[15], that is,

$$\boldsymbol{\eta}^{(p+1)} = \boldsymbol{\xi}^{(p+1)} - \boldsymbol{h}^{(p+1)}, \tag{3.23}$$

with $\boldsymbol{h}^{(p+1)} = \{h_i^{(p+1)}\}$ given by:

$$\boldsymbol{h}^{(p+1)} = \sum_{\mu=1}^{p} \boldsymbol{\eta}^{(\mu)} \left(\boldsymbol{\eta}^{(\mu)}.\boldsymbol{\xi}^{(p+1)}\right) - \mathcal{O}\left(\frac{p}{N}\right) \boldsymbol{\xi}^{(p+1)}. \tag{3.24}$$

A highlight of the H-H-GS model is that if $\boldsymbol{\xi}^{(p+1)}$ is one of the previously memorized patterns, say $\boldsymbol{\xi}^{(\nu)}$ $(1 \leq \nu < p)$, then $\boldsymbol{\xi}^{(\nu)}$ will project onto only the first $\boldsymbol{\eta}^{(\nu)}$ patterns and not $\boldsymbol{\eta}^{(\nu+1)} \ldots \boldsymbol{\eta}^{(p)}$ [3]. (We have verified this in Sec. 3.3.1.) The first $(\nu - 1)$ terms in eq.(3.24) yield $(\boldsymbol{\xi}^{(\nu)} - \boldsymbol{\eta}^{(\nu)})$ and so we get,

$$\boldsymbol{h}^{(\nu)} = \boldsymbol{\xi}^{(\nu)} - \boldsymbol{\eta}^{(\nu)} + \hat{\boldsymbol{\eta}}^{(\nu)} \left(\hat{\boldsymbol{\eta}}^{(\nu)} \cdot \boldsymbol{\xi}^{(\nu)}\right) - \mathcal{O}\left(\frac{p}{N}\right) \boldsymbol{\xi}^{(\nu)} = \left(1 - \mathcal{O}\left(\frac{p}{N}\right)\right) \boldsymbol{\xi}^{(\nu)}, \tag{3.25}$$

as $\hat{\boldsymbol{\eta}}^{(\nu)} \cdot \boldsymbol{\xi}^{(\nu)} = \hat{\boldsymbol{\eta}}^{(\nu)} \cdot \hat{\boldsymbol{\eta}}^{(\nu)}$ for orthonormal vectors. The network would hence identify $\boldsymbol{\xi}^{(p+1)}$ as $\boldsymbol{\xi}^{(\nu)}$ while $\boldsymbol{\eta}^{(p+1)}$ will be approximately 0. This indicates that $\boldsymbol{\xi}^{(p+1)}$ will not be orthogonalized and re-learnt by the system however many times it is presented.

However, any novel $\boldsymbol{\xi}^{(p+1)}$ would be identified as such and the corresponding $\boldsymbol{\eta}^{(p+1)}$ would be calculated, adding it to the memory store.

The main point to be emphasized here is that the process of orthogonalization restricts the noise $\mathcal{A}$ by removing the correlations of the new incoming patterns with those already in the memory. The PSP on each neuron i ($h_i^{(p+1)}$) takes the value $\left(1 - \mathcal{O}\left(\frac{p}{N}\right)\right) {\xi_i}^{(\nu)}$. Moreover, the PSPs are restricted to lie between the limits $\left(\left(\mathcal{O}\left(\frac{p}{N}\right) - 1\right), \left(1 - \mathcal{O}\left(\frac{p}{N}\right)\right)\right)$, as $\xi_i^{(\nu)} = \pm 1$. This implies that the synapses do not get modified on each presentation of a previously memorized pattern, and are thereby protected against possible runaway potentiation (or depression) cycles.

3.5 Discussion

Various strategies have been proposed to address the issue of loss of information stored in a system, due to CI in the case of artificial networks, or the stability-plasticity problem in biological systems. A system needs to be malleable in response to information entering it for storage; at the same time, it must also be impervious to modifications in order to retain the stored memories. In our current strategy, we have used a conventional Hopfield network. While we do not claim biological accuracy or incorporate finer physiological details in our model, we argue that our model is a simple tool but still provides us a handle on addressing the stability-plasticity dilemma. Our strategy uses the same set of units to learn and memorize information, with new information being compared to what is already present in the system, and stored in relation to that. The network gets endowed with the capability of recovering any of the memories in store, though only the similarities and differences between the patterns get recorded.

The H-H-GS model in our proposed strategy is capable of storing a much larger number of sequentially presented patterns compared to the conventional Hopfield model. Moreover, the scheme also provides a means of comparing and generalizing newer information with respect to the ones stored earlier. This is in contrast with other non-overlapping strategies to surmount CI (for instance, see [43] or [31]). Segregating the patterns to be memorized would eliminate CI. But at the same time, it would prevent the network from being able to identify shared features or generalize the input set, thereby affecting the capacity of the network to classify or categorize the stored information [35]. The chief merit of our model is that the information learnt using

the orthogonalization scheme will automatically be compared to and stored in the context of what has already been memorized, without any need to restrict synaptic efficacies or the learning rate.

Learning in humans involves newer information being added to and blending with previously stored memories without superimposing on them[44]. The Gram-Schmidt orthogonalization procedure inherently possesses this property, without having to implement it separately. This property can be harnessed to benefit any system– biological or artificial– implementing the scheme. The system would then compare new and old information, and identify their commonalities and differences. It would also store each information only once, even if it encounters the same information multiple times. The system would hence be self-organized, also acting as an "internal supervisor"[22] to identify the synapses to be modified to accommodate a new memory over earlier ones. Repeated presentations of a stimulus once learnt do not modify the local field potential on the neurons, and preventing the possibility of a BCM-like runaway effect by keeping in check any unrestricted growth (or decay) in the synaptic efficacies.

While orthogonalization has previously been suggested as a remedy to the issue of CI in artificial neural networks (for instance, see [45]), the term differs in meaning from our usage. The term 'orthogonalization' could typically refer to the use of sparse coding to eliminate the interference between correlated patterns by using different uncorrelated sets of nodes to store different items of information (see for instance [31, 46, 47] and references therein). In our scenario, orthogonalization refers to making vectors mutually perpendicular. Each vector represents an information, and new vectors are made perpendicular to the previous ones such that there is no overlap between the vectors. These vectors are all stored using the same set of nodes.

Randomly generated vectors whose components can take one of two values can be expected to be orthogonal or non-overlapping. However, this holds true only for infinitely large systems of vectors. For finite vectors, the orthogonalization is only approximate, and so their inner products will take a finite value rather than zero as in the case of truly (completely) orthogonal vectors. This orthogonalization is not deliberate, and the non-zero overlaps imply that the noise will eventually override the signal, beyond the limit of $p/N = 0.14$[16]. The general idea of orthogonal vectors is that of a sparsely coded set whose components do not overlap with each other [48], which can help overcome CI, irrespective of how the vectors are orthogonalized [45].

In the Gram-Schmidt procedure we have described here, the set of randomly generated input vectors is intentionally made mutually perpendicular. This eliminates the noise in the system, thus increasing the memory capacity of the network from $p/N = 0.14$ to $p/N \approx 1$. (Note that the set of vectors which is learnt and retrieved the network is still the original random set of correlated patterns- the input set is itself not altered or made mutually orthogonal as in [49])

Our study has been carried out on an artificial neural network, but experiments are needed to see how it pertains to living networks. While this strategy may have some bearing on actual biological systems, it must be verified experimentally, given that the degree of biological accuracy can influence the behaviour of theoretical systems differently (refer to [34]). While there is evidence of sliding thresholds for plasticity[39], orthogonalization of input information necessitates certain network architecture and physiological conditions. For instance, a network must have feedforward excitation and feedback inhibition[16], or dendritic multiplication and the nature of inputs to different dendrites of the same neuron[48]. These criteria pertain to general structural or functional properties of biological networks, making experimental verification of theoretical predictions feasible.

Chapter 4

Stability and associativity of memories in Attractor Neural Networks

In this chapter, we analyze the H-H-GS model in more detail, focusing on pattern stability which is crucial for associativity. For this, we first define precisely the terms retrieval, recognition and recall. We then study the effects of GS orthogonalization on pattern stability and the associative property of the network. We also discuss what the results mean in terms of cognition.

4.1 Introduction

Learning and memory are inherently associative by nature. An information learnt and stored in the memory can be recovered not just on coming across the same information, but also by encountering something even partially similar. The brain thus possesses the property of associativity– it sorts and groups information within the memory store as well as external stimuli it encounters. Theoretical and mathematical modelling of such associative networks using ideas from physics and mathematics could shed light on functional aspects of learning and memory[1, 17, 50, 51, 52, 53].

Associativity can be modeled using Attractor Neural Networks (ANNs) which can learn information presented as sequences of ± 1s, or *patterns*, in such a way that the memorized patterns become fixed points and *attractors* in the network dynamics[7].

Each attractor is surrounded by a *basin of attraction*, the region containing the set of patterns associated with the attractor[7, 8]. When the network is presented with any of the patterns within a basin of attraction, it identifies and recovers the attractor pertaining to the basin. Even presenting an incorrect or inexact version of the pattern would yield the memorized pattern, indicating that the network is capable of 'error correction'. The network is also capable of classification or categorization of information, as categorization in essence amounts to partitioning the pattern space and separating the patterns into basins of attraction[8].

The proper functioning of a network as an effective memory system depends critically on the stability of the patterns lodged in the memory. That is, memorized patterns must form (stable) fixed points in the network dynamics. In addition to being fixed points, the learnt patterns must also be attractors. Only then can the system be an efficient associative memory. The reason for this distinction will be explained shortly. The chief objective of this chapter is to examine the meaning of pattern stability and study how it is influenced by the network dynamics. We also define clearly and distinguish between the terms retrieval, recognition and recall which are often used interchangeably in the literature. The reason for this lies in how each of these terms relates to pattern stability, which we will explore in detail in a later section.

The deterioration in the memory capacity and the ability of the Hopfield model to function as an effective associative memory beyond a certain low limit has been established widely. The cause for this degradation is the so-called catastrophic interference, the consequence of the memorized patterns overlapping with each other. The H-H-GS model[15, 16, 53] described in the previous chapter provides a way of getting around the problems of CI[17]. The model was able to improve the memory capacity of the network and model some cognitive aspects of learning and memory. But whether the effects of orthogonalization on the processes of retrieval, recognition and recall are uniform or vary in extent is not clear at the outset. It is therefore imperative to provide proper definitions for these terms, how they relate to each other, how orthogonalization affects these relations and what the ramifications are for pattern stability. We address these issues in this chapter through an in-depth study of the influence of orthogonalization on memory stability and associativity.

4.2 Retrieval, recognition and recall

It is essential that we define the terms retrieval, recognition and recall at the outset, prior to elaborating on our study. In physics and neuroscience literature, the terms are treated as being equivalent synonymous with each other. While they are differentiated in cognitive-psychology literature, their interpretation there is different, and it is hard to establish a one-to-one correspondence in the meanings of the terms across domains. We hence need to define the terms clearly and also in mathematical terms.

If the presentation of any of the previously memorized information to the network yields that particular information instantaneously and accurately, we consider that information as *retrieved* by the network. Within our framework, *exactness* and *instantaneity* are crucial for the process of retrieval. However, the reproduction of the presented learnt pattern may not be perfectly accurate in the first instance, but may take multiple steps, with a similar pattern reproduced at each step before culminating accurately in the learnt pattern. We term this process *recognition* of the presented pattern as part of the stored information. We must point out that the process of recognition comprises two parts, which we will elaborate on, but first we define recall.

If the network is able connect a memorized pattern within a few steps when it comes across the same pattern or a very similar one, then the network can *recall* information. This is akin to *pattern completion* or *error correction*. The process of recovering a memorized pattern when the network encounters an incomplete or incorrect or erroneous version of that pattern is referred to as pattern completion or error correction[54, 55, 56, 57]. Note that *recollection* of an inscribed pattern can occur even when the presented information is different from that learnt pattern.

The information reproduced during recognition is identified as being already in the memory store, that is, *familiar*. One part of recognition is thus *familiarity*, as used in cognitive-psychology literature [58, 59, 60, 61, 62, 63]. The other part of recognition is *recollection*. In our context as in cognitive psychology literature, the combined effect of the processes of *familiarity/retrieval* and *recollection* results in *memory recognition*.

We will later come upon a scenario where a memorized pattern is neither retrieved nor recognized when presented to the network, but instead gets associated with a new pattern which is not any of the stored patterns but is very similar to that presented

pattern. We now consider the issue of pattern stability before further discussion of the scenario.

We would like to point out here that *associative recall* is present in our model system so long as the learnt patterns each have a set of novel but similar patterns which eventually lead to the stored patterns, and so are associated with them.

We will now present mathematical definitions of the terms retrieval, recognition and recall and connect them to the discussion above. We know from eq.(2.9) that for a pattern inscribed in the network to be treated as *recovered*, it must satisfy the following condition:

$$sgn\left(h_i^{(\nu)}\right) = sgn\left(\xi_i^{(\nu)}\right) \text{ (for all } i\text{'s)}, \tag{4.1}$$

where $h_i^{(\nu)}$ is the LFP/PSP from eq.(2.4)

$$h_i^{(\nu)} = \sum_{\substack{i=1 \\ i \neq j}}^{N} J_{ij}\xi_j^{(\nu)}. \tag{4.2}$$

We will henceforth refer to each occurence (instance) of the above equations as an *iteration*.

In the following analysis, we use $\boldsymbol{\xi}^{(\nu)}$ and $\boldsymbol{\xi}^{(t)}$ to represent an inscribed pattern and a test pattern respectively. $\boldsymbol{\xi}^{(\nu)}$ is one of the learnt patterns chosen at random, while $\boldsymbol{\xi}^{(t)}$ is a pattern similar to $\boldsymbol{\xi}^{(\nu)}$ or $\boldsymbol{\xi}^{(\nu)}$ itself. Now, if $\boldsymbol{\xi}^{(t)}$ is presented to the network and checked for retrieval using eqs.(4.2) and (4.1), then

- if $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}^{(\nu)}$ and $\boldsymbol{\xi}^{(\nu)}$ is recovered spontaneously, that is, a test pattern which is exactly the same as an inscribed pattern recovers that pattern in the first instance itself, then $\boldsymbol{\xi}^{(\nu)}$ is said to be *retrieved*. Also,
- if $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}^{(\nu)}$ converges to $\boldsymbol{\xi}^{(\nu)}$ but in more than one iteration, then $\boldsymbol{\xi}^{(\nu)}$ is considered as *recognized*. In other words, $\boldsymbol{\xi}^{(t)}$ reaches $\boldsymbol{\xi}^{(\nu)}$ within a few steps and does not deviate from it with subsequent iterations. But,
- if $\boldsymbol{\xi}^{(t)}$ is similar to but not the same as $\boldsymbol{\xi}^{(\nu)}$ ($\boldsymbol{\xi}^{(t)}$ differs from $\boldsymbol{\xi}^{(\nu)}$ on some components) but still converges to $\boldsymbol{\xi}^{(\nu)}$ within a small number of steps, then $\boldsymbol{\xi}^{(\nu)}$ is considered to be *recalled*.

Note that the condition for retrieval does not inherently check for or guarantee convergence. Prior to settling down to an attractor, $\boldsymbol{\xi}^{(t)}$ may recover an inscribed pattern $\boldsymbol{\xi}^{(\nu)}$, but further iterations would bring it away from $\boldsymbol{\xi}^{(\nu)}$. $\boldsymbol{\xi}^{(t)}$ would eventually settle down to an attractor which may or may not be one of the memorized patterns. This is illustrated in Table4.1.

TABLE 4.1: Examples of the evolution of a presented pattern till convergence : how a pattern reaches an attractor. The values represent the overlap of the presented pattern $X = \xi^{(\nu)}$ and the retrieved pattern X'. Convergence is immediate, as in the case of X_1, or after a very small number of iterations, as in the case of X_2. In both these cases, the inscribed pattern presented for convergence is stable and is an attractor. In the case of an unstable pattern like X_3, even though X_3 is retrieved in the first iteration, it moves away and eventually converges at X_3', which is 96% similar to X_3 and is the new attractor.

Iteration	1	2	3	4	5	6	7	8	9	10
X_1	100	100	100	100	100	100	100	100	100	100
X_2	99	100	100	100	100	100	100	100	100	100
X_3	100	99	97	96	96	96	96	96	96	96

TABLE 4.2: Table illustrating the difference between retrieval and recognition. For different values of p ($p = 12, 14, 16$) in a network of size $N = 100$, the rows marked Ret indicate whether a pattern is retrieved (✓) or not (×), while the rows marked Rec indicate whether or not recognition has happened. A pattern which is retrieved is also trivially understood as also being recognized. For patterns which are not retrieved, the overlap between the inscribed pattern and the recovered pattern after the first iteration are indicated within square brackets [ ]. Upto $p = 10$, all the patterns are retrieved. When a pattern is not retrieved, it is presented back to the network and the process is repeated to check whether the pattern converges within 10 iterations. The ( ) give the number of iterations to convergence, and (-) indicates lack of convergence after 10 iterations. The table shows the results for the first 10 patterns $\boldsymbol{\xi}^{(1)}$ to $\boldsymbol{\xi}^{(\mathbf{10})}$ for each value of p.

		$\boldsymbol{\xi}^{(1)}$	$\boldsymbol{\xi}^{(2)}$	$\boldsymbol{\xi}^{(3)}$	$\boldsymbol{\xi}^{(4)}$	$\boldsymbol{\xi}^{(5)}$	$\boldsymbol{\xi}^{(\mathbf{6})}$	$\boldsymbol{\xi}^{(\mathbf{7})}$	$\boldsymbol{\xi}^{(\mathbf{8})}$	$\boldsymbol{\xi}^{(\mathbf{9})}$	$\boldsymbol{\xi}^{(\mathbf{10})}$
$p = 12$	Ret	✓	✓	✓	✓	✓	✓	✓	×[99]	✓	✓
	Rec	✓	✓	✓	✓	✓	✓	✓	✓(2)	✓	✓
$p = 14$	Ret	✓	✓	✓	✓	✓	✓	×[99]	×[99]	✓	✓
	Rec	✓	✓	✓	✓	✓	✓	✓(3)	✓(3)	✓	✓
$p = 16$	Ret	✓	✓	✓	✓	✓	✓	✓	×[99]	✓	✓
	Rec	✓	✓	✓	✓	✓	✓	✓	×(-)	✓	✓

Table4.2 provides some examples of the difference between retrieval and recognition. The reason for the emphasis on convergence and for distinguishing between retrieval and recognition is that convergence (to self) is a defining feature of an attractor and crucial for pattern stability, as we shall see later. When we talk about convergence, we mean an exact match between the presented and recovered patterns. While some errors are generally permitted and a $\geq 97\%$ accuracy deemed sufficient to term a pattern recovered, we impose the stricter condition of perfect fidelity, which we will justify shortly. Fig.4.1 shows how convergence quality relates to different degrees of accuracy. Their cognitive relevance will be discussed in a later section.

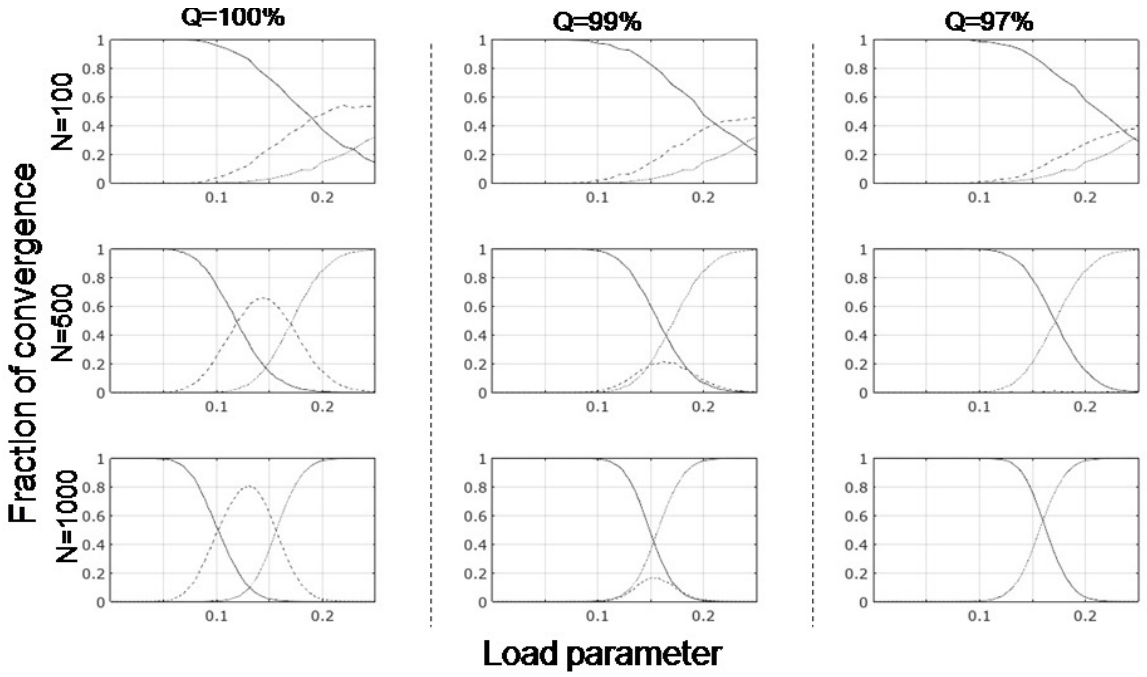


FIGURE 4.1: Quality of convergence as a function of load parameter: Plot showing where the inscribed patterns in the Hopfield network converge for various values of N ($N = 100, 500, 1000$), for different qualities of convergence ($Q = 100\%, 99\%, 97\%$). A pattern can converge either to itself (solid line) or to a different pattern (dashed line), depending on the error tolerance for that value of Q. The thin dotted line indicates lack of convergence even after 10 iterations. For $Q = 100\%$, the error tolerance is 0%, and only a perfect match is treated as converging to 'itself'. For $Q = 99\%$ and 97%, the error tolerances are 1% and 3% respectively, which means that convergence to a nearby pattern is acceptable in such calculations . The plots show data from 50 different sets of patterns.

4.3 Criteria for pattern stability

Having defined the various terms, we now focus on establishing the conditions for an inscribed pattern to be considered stable, as pattern stability is a fundamental requirement for the network to function effectively as an associative memory.

Since the stability condition in eq.(2.4) does not assure or necessarily imply convergence of a test pattern to itself, it cannot be the sole criterion for memory stability. As seen from Table4.1, a test pattern may reproduce an inscribed pattern in some iteration before settling down to a different pattern on further iterations. Errors, or mismatches between the presented and recovered patterns in an iteration could potentially lead to more mismatches on later iterations, eventually creating an 'avalanche' of errors[19]. Hence, recognition rather than retrieval must be a criterion for the stability of a memory.

We explain how a pattern may be rendered unstable by expressing eq.(4.2) using a signal term and a noise term arising from the correlations between the patterns. This noise is referred to as *slow noise*. We now introduce a stabilization parameter $\mathfrak{s}_i^{(\nu)}$ [7, 9] given by,

$$\mathfrak{s}_i^{(\nu)} = h_i^{(\nu)} \xi_i^{(\nu)}. \tag{4.3}$$

From the signal-to-noise analysis in Sec.2.5 of Chapter2, we have

$$h_i^{(\nu)} = \xi_i^{(\nu)} + \frac{1}{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\nu)} \right), \tag{4.4}$$

which can be expressed in terms of the stabilization parameter as,

$$\mathfrak{s}_i^{(\nu)} = 1 + \frac{1}{N} \sum_{\substack{\mu=1 \\ \mu \neq \nu}}^{p} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\xi_i^{(\mu)} \xi_j^{(\mu)} \xi_j^{(\nu)} \right) \xi_i^{(\nu)}. \tag{4.5}$$

From the above equation (and from Sec.2.4), we can see that it is the noise term whose value determines whether or not the signal will be detected, thereby indicating whether or not the pattern will be stable. The noise term can take negative values, and the likelihood of it doing so goes up as more patterns are stored in the network. From eqs. (4.1) and (4.3), we see that pattern stability requires the stabilization parameter to take a positive value. We can also understand this from eq.(4.5), for when $\mathfrak{s}_i^{(\nu)} > 0$, the signal is clearly distinguishable from the noise and $\boldsymbol{\xi}^{(\nu)}$ will be

recognized (or retrieved, depending on the pattern) or recalled if $\boldsymbol{\xi}^{(\nu)}$ is different from $\boldsymbol{\xi}^{(\nu)}$.

Taking our cue from [8] (also refer to [19] and [9] for other discussions pertaining to the conditions for pattern stability), we will now list out the conditions for pattern stability: for a pattern $\boldsymbol{\xi}^{(\nu)}$ to be stable,

1. it must converge to itself with perfect accuracy within a few iterations when presented to the network, that is, it must be recognized.
2. it must be an attractor in the network dynamics, with its own basin of attraction. In other words, the network should be able to recall $\boldsymbol{\xi}^{(\nu)}$ when presented with any of the (similar) patterns within its basin of attraction.
3. it must have a positive value for the stabilization parameter on each of the neurons - $\mathfrak{s}_i^{(\nu)}$ must be > 0 for all i's.

The first two conditions are interlinked, as the inscribed pattern will be an attractor and have a basin of attraction only if it converges to itself without any errors. If it converges to a different pattern, then that pattern would be an attractor with its own basin of attraction, and the inscribed pattern would be one among the patterns within the basin. This new pattern would also necessarily converge to itself. In general, any pattern which is an attractor and has a basin of attraction will converge to itself when presented for recovery. Thus, condition1 can be treated as a necessary condition for the stability of a pattern (any pattern, not necessarily one in the memory store), while conditiom2 would be a sufficient one. Also, note that condition 3 will hold true as long as condition1 is satisfied.

We will come back to the pattern stability criteria discussed above and examine their pertinence to the network with orthogonalization(the H-H-GS model described in the previous chapter), where the patterns to be memorized are first orthogonalized and normalized and stored. The network is able to recover the memorized patterns even though it is their orthonormalized versions that are stored. The H-H-GS model will present a case which challenges the above criteria - there will be a situation where patterns will be recognized (or retrieved, at times) but will not have basins of attraction. Such patterns which are stable fixed points but not attractors are referred to as stable non-attractors. In the H-H-GS model, we will find that all the inscribed patterns are stable attractors, but with increasing memory loads, the memorized

patterns become stable non-attractors. We will hence claim that the learnt patterns are stable states though they only satisfy the necessary condition of recognition, but not the sufficient condition of non-zero basins of attraction. This will bring us to question what a basin of attraction with radius zero means.

As the basins of attraction are integral to the concept of memory stability, we first perform a detailed investigation of the network dynamics of the H-H model. We then invoke orthogonalization in the network and examine how the network dynamics, including basins of attraction are modified following orthogonalization.

4.4 A study of the behaviour and dynamics of the H-H model

In light of the definitions presented earlier in the chapter, and the criteria for pattern stability listed above, we now look at the behaviour and dynamics of the Hopfield network.

4.4.1 A detailed analysis of the basins of attraction

We discussed the concept of basins of attraction briefly in Chapter2 (Sec.2.5), and will look at them more closely now. We will first discuss the size and shape of the basins of attraction and how to compute them (emulating [8]). We can visualize a $N+1$-dimensional space with N space dimensions and an energy dimension. Each pattern $\boldsymbol{\xi}^{(\nu)}$ is a point in this *configuration space* and possesses some energy $E\left(\boldsymbol{\xi}^{(\nu)}\right)$. We can picture an *energy landscape* made up of hills and valleys, with the bottoms of the valleys containing the energy minima, the points with lowest energy. Each inscribed pattern must minimize the total energy of the system, the Hamiltonian (eq.(2.6)) and hence be an energy minimum lying at the bottom of a valley. The patterns in the valley surrounding the minimum converge to it and form its basin of attraction containing the set of patterns associated with it.

While the configuration space needs to be large and capable of containing $2^{(N)}$ patterns, it must also be finite (for mathematical reasons not discussed here). With increasing number of memorized patterns, the energy landscape gets reorganized to accommodate each new pattern. This process brings about topographical changes

with basins shrinking and becoming shallower, and energy minima getting displaced. Consequently, the minima corresponding to some of the inscribed patterns may be replaced by other new patterns which were not learnt by the network, and those inscribed patterns now lie in the basins of attraction of these new minima.

The size of the basin is typically expressed as the average of the extent of the basin in different directions from the minimum[8]. In the H-H model, it is expressed in terms of the Hamming distance, a measure of the number of sites on which two patterns differ. But, the average may be an inaccurate and even a deceptive measure, as the term basin of attraction encompasses the complete structure around a minimum, and is generally irregular in shape. as represented schematically in Fig.4.2. Thus, instead of a single average value, a better choice of representation of a basin would be to use a set of Hamming distances for different directions starting from the minimum. Each of these Hamming distances denotes the maximum distance or number of differences for which a test pattern still converges to that minimum. We will justify our choice later.

We can see from Table4.3 that the basin of attraction is not uniform or anisotropic. It may extend to different extents in different directions, as reflected by the great variations in the set of Hamming distances denoting the basin. While the basin of a stable pattern may have one or a few '0' values, it will also have other non-zero values. However, all the Hamming distances in the basin of an unstable pattern will be zeroes, as an unstable pattern lacks convergence to itself. Although an unstable pattern may converge elsewhere, to a 'new' attractor with a non-zero basins, an unstable pattern itself will not have basin of attraction.

A brief summary of the calculation of a basin is as follows: we start with a pattern $\boldsymbol{\xi}^{(\nu)}$ whose basin we wish to evaluate. We then choose a random sequence, referred to as a *sample*, and flip the elements of $\boldsymbol{\xi}^{(\nu)}$ according to the sequence and present it to the network for recognition. At each step, we test whether the presented pattern converges to $\boldsymbol{\xi}^{(\nu)}$. The process is continued as long as there is convergence, and the maximum such value gives the Hamming distance for that sample. We would like to point out here that this methodical flipping of spins acts as another source of noise, referred to as *fast noise*. We repeat the procedure for a specified number of samples and get the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$. We then repeat the complete procedure for multiple sets with the same number of memorized patterns, with each set constituting a *trial*. Refer to AppendixB for the detailed protocol illustrated with an example.

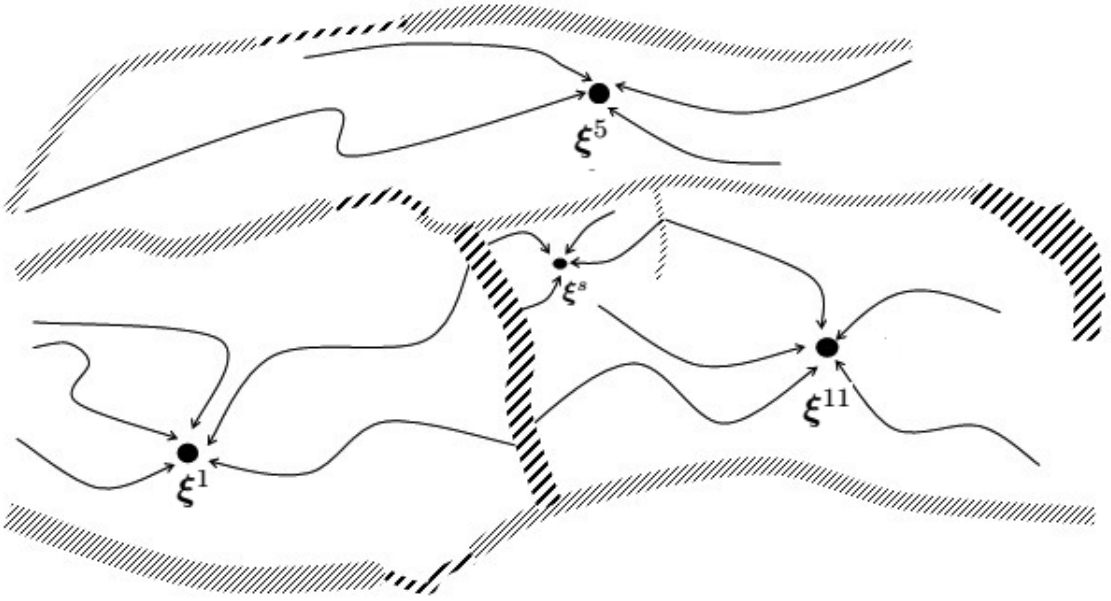


FIGURE 4.2: Schematic diagram showing the basins of attraction for three arbitrary memories, $\boldsymbol{\xi}^1$, $\boldsymbol{\xi}^5$, and $\boldsymbol{\xi}^{11}$ inscribed in the configuration space - these form minima in an energy landscape created by eq. ((2.6)). A configuration or pattern of ± 1's forming a vector similar to an inscribed pattern and falling in the latter's basin would converge to the inscribed vector or pattern. There can be other shallow minima within a basin of attraction of an inscribed patterns, but their basins of attraction are usually very small, and patterns falling within these basins eventually converge to the inscribed pattern by increasing the temperature. An example of such a spurious minimum $\boldsymbol{\xi}^s$, is also shown. Note that the minima are separated by hilly regions which are nonuniform in their height and width as shown by the hatched boundaries of the basins.

4.4.2 Network dynamics and exploration of the energy landscape

When patterns are learnt by the network following the Hebbian learning prescription in eq.(2.3), they minimize the total energy of the system given by eq.(2.6) and become attractors. This is due to the inherent nature of the learning rule, as we have seen earlier in Sec.2.3. However, with an increase the number of stored memories, the network dynamics get altered and the energy landscape curtailed . We can see this from energy calculations (after [20]). A pattern $\boldsymbol{\xi}^{(\mu)}$ is characterized by an energy $E(\boldsymbol{\xi}^{(\mu)})$ given by,

$$E(\boldsymbol{\xi}^{(\mu)}) = -\frac{1}{2}\boldsymbol{\xi}^{(\mu)} J \boldsymbol{\xi}^{(\mu)\,T}, \tag{4.6}$$

TABLE 4.3: **Table showing the basins of attraction of certain attractors corresponding to various inscribed patterns ($\boldsymbol{\xi}^{(\nu)}$'s) for different values of p for $N = 100$. $\boldsymbol{\xi}^{(\nu')}$ represents the (new) attractor to which an unstable $\boldsymbol{\xi}^{(\nu)}$ converges. The basin of attraction of a stable pattern can include a Hamming distance of value zero: 0 can be one amongst the various numbers representing the basin. The basin of attraction of $\boldsymbol{\xi}^{(4)}$ when $p = 12$ provides an example of this. In addition, we can see the differences in basins of attraction even between stable patterns like in $\boldsymbol{\xi}^{(1)}$ and $\boldsymbol{\xi}^{(4)}$ for the same value of p. Also note the anisotropy in the spread of the Hamming distances in the case of $\boldsymbol{\xi}^{(4)}$. The basins of attraction of the stable patterns shrink in most directions as p increases, as seen for $p = 14$ and $p = 16$. The new attractors $\boldsymbol{\xi}^{(\nu')}$ have non-zero basins of attraction.**

	$\boldsymbol{\xi}^{(\nu)}$	Hamming distances forming the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$									
$p = 10$	$\boldsymbol{\xi}^{(1)}$	31	29	37	35	40	24	30	34	41	31
	$\boldsymbol{\xi}^{(7)}$	44	39	48	39	39	32	40	36	35	37
$p = 12$	$\boldsymbol{\xi}^{(1)}$	32	36	32	44	34	34	35	32	34	37
	$\boldsymbol{\xi}^{(4)}$	0	4	6	7	3	6	6	8	1	19
	$\boldsymbol{\xi}^{(7)}$	0	0	0	0	0	0	0	0	0	0
	$\boldsymbol{\xi}^{(7')}$	31	32	39	35	40	26	30	47	37	37
$p = 14$	$\boldsymbol{\xi}^{(1)}$	26	31	36	41	37	22	30	31	31	39
	$\boldsymbol{\xi}^{(7)}$	0	0	0	0	0	0	0	0	0	0
	$\boldsymbol{\xi}^{(7')}$	5	16	8	22	3	7	6	10	6	6
$p = 16$	$\boldsymbol{\xi}^{(1)}$	31	21	8	27	28	35	27	33	17	28
	$\boldsymbol{\xi}^{(7)}$	0	0	0	0	0	0	0	0	0	0
	$\boldsymbol{\xi}^{(7')}$	13	6	10	7	7	23	8	17	18	7

where J is a matrix containing the synaptic efficacies and T gives the transpose of the vector $\boldsymbol{\xi}^{(\mu)}$ which is typically expressed as a row vector. The transpose of a vector converts it into a column vector if it is a row vector and vice-versa. Now, after memorizing p patterns, the synaptic weight matrix becomes,

$$J = \sum_{\mu=1}^{p} \boldsymbol{\xi}^{(\mu)\,T} \boldsymbol{\xi}^{(\mu)} - pI. \tag{4.7}$$

Here I denotes an $N \times N$ identity matrix. The second term in the equation containing I serves to eliminate self-connections from the efficacy matrix.

Now, from eqs.(4.2),(4.3) and (4.6), we get,

$$E(\boldsymbol{\xi}^{(\mu)}) = -\frac{1}{2}\mathfrak{s}^{(\mu)}. \tag{4.8}$$

$\mathfrak{s}^{(\mu)}$ is a vector comprising the energy on all the neurons. We can see from eq.(4.5) that the stabilization parameter will be positive and its value large as long as p is much smaller compared to N. From eq.(4.8) it follows that the memorized patterns are energy minima at these memory loads. We see from Table4.4 that the average energy of an attractor, whether a memorized pattern or not, lies close to a mean value which is generally $N/2$. The spread of the energies of the attractor around this mean value remains impervious to changes in both p and N, though the value of p influences the sizes of the basins of attraction.

TABLE 4.4: **Table showing the mean, or average energy of patterns, $\boldsymbol{E_{avg}^{(\mu)}}$, and that of the attractors, $\boldsymbol{E_{avg}^{(\mu')}}$, for a particular value of p for various N. The energies of the patterns are distributed around a mean, which is typically around $N/2$. Even as p increases, the energies of the patterns are still scattered around $N/2$.**

p	Average Energy	$N = 100$	$N = 200$	$N = 500$	$N = 700$	$N = 1000$
$0.10N$	$\boldsymbol{E_{avg}^{(\mu)}}$	-48.58	-100.18	-249.49	-349.63	-500.66
	$\boldsymbol{E_{avg}^{(\mu')}}$	-48.58	-100.18	-249.49	-350.15	-501.14
$0.14N$	$\boldsymbol{E_{avg}^{(\mu)}}$	-49.06	-98.62	-249.83	-349.55	-503.45
	$\boldsymbol{E_{avg}^{(\mu')}}$	-49.06	-98.66	-251.70	-350.43	-504.46
$0.16N$	$\boldsymbol{E_{avg}^{(\mu)}}$	-49.36	-102.24	-248.38	-350.95	-501.30
	$\boldsymbol{E_{avg}^{(\mu')}}$	-49.62	-102.52	-250.34	-353.21	-503.94
$0.20N$	$\boldsymbol{E_{avg}^{(\mu)}}$	-48.52	-102.76	-247.82	-352.95	-499.95
	$\boldsymbol{E_{avg}^{(\mu')}}$	-49.00	-103.98	-249.83	-360.21	-504.42

Lodging a new information in the memory modifies the J_{ij}'s. The energy landscape also undergoes changes to fit in the new information along with its basin of attraction. These changes affect the basins of attraction of the earlier memories causing them to stretch in some directions and contract in others. Additionally, the energy minima may get displaced, and the depths of the basins could reduce. As a result, the

anisotropy or non-uniformity in the shapes of the basins increases, making the energy landscape more convoluted. Storing new memories also contributes to a growth in the slow noise. The basins also shrink further as a consequence of the increased noise due to interference between the various basins of attraction[10].

As the neurons in the Hopfield network are all interconnected, each pattern within the system is correlated with all the other patterns, and not just with those in close proximity. Eq.(4.4) also shows this. Storing more information hence contributes to the cross-talks and hence to the noise. The growth in the noise may affect the recognition of some of the imprinted patterns. The noise term grows so much that the stabilization parameter (eq.(4.3)) turns negative on some sites. In other words, the noise *destabilizes* some of the neurons in stable patterns. As a result, the signs on some of the elements of the inscribed patterns differ from the corresponding elements of the recovered patterns, indicating that the patterns have now become unstable. Consequently, their basins of attraction shrink to zero or disappear.

In brief, the memorized patterns are all stable at first, and they all minimize the Hamiltonian in eq.(2.6). They are also all attractors with large basins of attraction. As the energy landscape restructures itself to allow for the addition of more information, some of the attractors may get displaced. An attractor may shift away from an inscribed pattern, making that pattern unstable. While these new attractors may at first lie near the inscribed patterns, increasing memory loads shift them further away. Fig.4.3 illustrates this displacement with rising p .

The energy landscape contains other attractors besides the learnt patterns. Storing a pattern in the memory also creates a minimum or an attractor corresponding to its inverse. Table4.5 shows the basins of attraction of some of the inverse states. We can see that the inverse states are attractors and that their basins of attraction are similar in size and shape to their counterparts among the inscribed patterns. We can hence consider the configuration space as being divided in two– one half pertains to the imprinted patterns, and the other to the inverses, with each half also containing the basins of attraction of the respective attractors.

Apart from these, there is another class of attractors known as *spurious states* or *false memories*. They are attractors that arise from and pertain to combinations of an odd number the memorized patterns. Hence, they are also referred to as *combination states* or *mixture states*. If a combination state is made from combining an even number of imprinted patterns, some of its elements may take value 0, instead

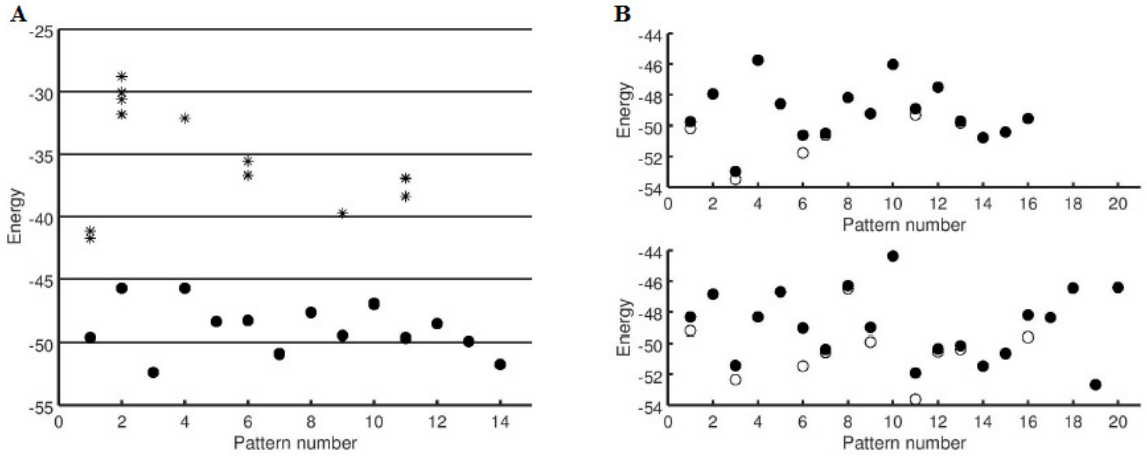


FIGURE 4.3: Plot showing the energies for the inscribed patterns (•) and the attractors(∘) for **(A)** $p = 14$ and **(B)** at $p = 16$ and 20 for $N = 100$. At $p = 14$, for stable patterns, the attractor corresponds to the inscribed pattern, indicated by the • and ∘ coinciding. For an unstable inscribed patterns (for instance, $\boldsymbol{\xi}^{(7)}$), the attractor is at a slightly lower energy than the inscribed pattern but very close to it. (The difference between the attractors in the plot is not very distinct at this resolution). The $*$'s represent various symmetric mixture states with $3, 5$ and 7 components, and are plotted against the inscribed pattern with which they have maximum overlap. These mixture states have much higher energy than any of the inscribed patterns. From Fig.**(B)**, we can see that even at such high values of p, there are still stable patterns, but with increase in p, the number of unstable patterns goes up. As p increases from 16 to 20, the attractors pertaining to unstable patterns move further away from the corresponding inscribed patterns.

of the permitted values of ± 1[19]. Moreover, combination states with an odd number of component patterns form stable fixed points, while mixtures with even number of components are saddle points in the network dynamics[7]. Mixture states formed from such simple and straightforward combinations typically overlap with their component states to a (fairly) similar extent and are termed *symmetric*. The energy landscape also contains minima corresponding to *asymmetric* mixture states, resulting from more complex combinations involving integer or fractional prefactors. An example of a symmetric mixture state could be $\boldsymbol{\xi}^{(mix)} = sgn\left(\boldsymbol{\xi}^{(1)} + \boldsymbol{\xi}^{(12)} + \boldsymbol{\xi}^{(14)}\right)$, and that of an asymmetric mixture, $\boldsymbol{\xi}^{(mix)} = sgn\left(2\boldsymbol{\xi}^{(10)} + (3/8)\boldsymbol{\xi}^{(2)} + (2/5)\boldsymbol{\xi}^{(4)}\right)$ (for $p = 14$).

These spurious memories are attractors that are distinct from the inscribed patterns. They also overlap with all the other patterns, but more so with their component patterns. However, compared to the inscribed patterns, they typically possess higher energy (as seen from Fig.4.3(**A**)) and have smaller and shallower basins of attraction[8], as we can see from Table4.6. The spurious memories are hence termed *local* minima, while the inscribed patterns are *global* minima.

TABLE 4.5: Table showing the basins of attraction of some inscribed patterns ($\boldsymbol{\xi}^{(\nu)}$'s) and their inverses($\boldsymbol{\xi}^{(\nu)}_{inv}$'s) for different values of p for $N = 100$. For stable patterns, the basins of attraction of the inverse or mirror states are similar to those of the inscribed patterns. The inverses of the unstable states (with no basins) may have basins of attraction.

	$\boldsymbol{\xi}^{(\nu)}$	Hamming distances in the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$									
$p = 10$	$\boldsymbol{\xi}^{(1)}$	31	29	37	35	40	24	30	34	41	31
	$\boldsymbol{\xi}^{(1)}_{inv}$	39	32	40	33	40	34	42	38	43	35
$p = 12$	$\boldsymbol{\xi}^{(1)}$	32	36	32	44	34	34	35	32	34	37
	$\boldsymbol{\xi}^{(1)}_{inv}$	32	35	30	35	42	40	27	32	39	36
	$\boldsymbol{\xi}^{(7)}$	0	0	0	0	0	0	0	0	0	0
	$\boldsymbol{\xi}^{(7)}_{inv}$	13	6	0	0	12	36	0	19	0	14
$p = 14$	$\boldsymbol{\xi}^{(1)}$	26	31	36	41	37	22	30	31	31	39
	$\boldsymbol{\xi}^{(1)}_{inv}$	12	26	14	15	34	41	32	18	19	20
	$\boldsymbol{\xi}^{(7)}$	0	0	0	0	0	0	0	0	0	0
	$\boldsymbol{\xi}^{(7)}_{inv}$	0	10	8	5	13	14	0	10	18	0

When p is low, a mixture state and its basin of attraction may lie anywhere within the larger basin corresponding to an inscribed pattern, typically the component imprinted pattern with which it has maximum overlap, or its inverse. A pattern within the basin of a false minimum, including the minimum itself, can be brought out of the basin by introducing a little extra noise (such as by flipping spins to raise the temperature); the eventual convergence occurs at the component imprinted state. The spurious states can hence be considered to be pseudoattractors. By contrast, a tremendous amount of energy would be required to bring a pattern out of the basin of a deep global minimum.

As p increases, there is a corresponding combinatorial increase in the number of spurious states[8]. Some of them may even form in the region of configuration space between the basins of other attractors. The growing number of spurious patterns crowd the configuration space, reducing the area available to the basins of attraction

TABLE 4.6: Table showing the basins of attraction of some mixture states for $p = 14$ with $N = 100$. A mixture state $\boldsymbol{\xi}^{(mix_{m,n,q,\cdots})}$ is formed from the combination of some of the inscribed patterns, viz. its 'Components' ($\boldsymbol{\xi}^{(m)}$,$\boldsymbol{\xi}^{(n)}$,$\boldsymbol{\xi}^{(q),\cdots}$) and converges to an attractor $\boldsymbol{\xi}^{(\mu)}$. This $\boldsymbol{\xi}^{(\mu)}$ can be one of the inscribed patterns, typically the component with which $\boldsymbol{\xi}^{(mix_{m,n,q,\cdots})}$ has maximum overlap, or $\boldsymbol{\xi}^{(mix_{m,n,q,\cdots})}$ itself . When $\boldsymbol{\xi}$ß$^{(mix_{m,n,q,\cdots})}$ converges to $\boldsymbol{\xi}^{(\mu)}$, it lacks a basin of attraction of its own, but when it converges to itself, the basin of attraction is very small as seen below.

Components of $\boldsymbol{\xi}^{(mix_{m,n,q,\cdots})}$	Attractor	Basin of attraction of $\boldsymbol{\xi}^{(mix_{m,n,q,\cdots})}$									
($\boldsymbol{\xi}^{(m)}$,$\boldsymbol{\xi}^{(n)}$,$\boldsymbol{\xi}^{(q),\cdots}$)	$\boldsymbol{\xi}^{(\mu)}$	Hamming distances									
$\boldsymbol{\xi}^{(1)}$,$\boldsymbol{\xi}^{(3)}$,$\boldsymbol{\xi}^{(4)}$	$\boldsymbol{\xi}^{(1)}$	0	0	0	0	0	0	0	0	0	0
$\boldsymbol{\xi}^{(2)}$,$\boldsymbol{\xi}^{(5)}$,$\boldsymbol{\xi}^{(8)}$	$\boldsymbol{\xi}^{(3)}$	0	0	0	0	0	0	0	0	0	0
$\boldsymbol{\xi}^{(3)}$,$\boldsymbol{\xi}^{(6)}$,$\boldsymbol{\xi}^{(4)}$,$\boldsymbol{\xi}^{(8)}$,$\boldsymbol{\xi}^{(11)}$	$\boldsymbol{\xi}^{(mix_{3,6,4,8,11})}$	0	0	0	6	2	0	1	0	0	3

of the imprinted patterns, making the basins smaller or even making the minima change positions. The network eventually enters a *spin glass* state[7, 8] where the synaptic efficacies become, to a large extent, random and there is no longer a direct correspondence between the imprinted patterns and the synaptic weights. Once the network has entered this state, none of the imprinted patterns can even be recognized anymore.

4.4.3 Memory capacity of the H-H network

We now look at the memory capacity of the network in the light of our above discussions on retrieval, recognition and recall, the energy and the basins of attraction. We know that the requirement for a pattern to be added to the memory is for it to minimize the Hamiltonian. However, we have seen that this condition is no longer satisfied by some of the inscribed patterns as the memory store expands, while there may be other novel unlearnt patterns which minimize the total energy. These novel patterns cannot be included in the list of memorized patterns. This brings us to a need to reexamine the definition of memory capacity.

We have pointed out subtle differences between retrieval, recognition and recall in our definitions of the terms. So it becomes clear that the memory capacity should include the imprinted patterns only if the network can recognize them. Only those

patterns can be guaranteed to be energy minima and attractors with their own basins of attraction, thereby also ensuring recall. The other imprinted patterns which cannot be recognized are neither energy minima nor attractors.

Our study thus highlights the importance of complete accuracy, recognition and recall for pattern stability and the storage capacity of the memory. We can hence expect our estimated memory capacity to be less than $p = 0.14N$[64], which we had also estimated/calculated earlier (in 2.4). The reason for this is our stricter condition of a 100% match between the presented and recovered patterns.

We would like to point out here that following our definitions in Sec.4.2 and the discussion in Sec.2.4 of Chapter2, α_c gives the memory capacity of the Hopfield network for recognition. We can see from the analysis in Sec.2.5 that it also marks the capacity for associative recall. To sum up, the memory capacity of the Hopfield model is the same for both recognition and recall, and is given by α_c.

4.5 An in-depth analysis of the H-H-GS model

It has previously been shown that when the Gram-Schmidt orthogonalization procedure is incorporated into the H-H model, the system gains some properties that make it an attractive representation of cognition[15, 16, 53]. In the previous chapter, we saw how the H-H-GS model provided the system a way out of the negative consequences of CI, thereby improving the memory capacity vastly, from $p = 0.14N$ in the H-H model to $p = N - 1$ in the H-H-GS model. Here, we must reiterate that though the network uses the orthonormalized vectors, the $\hat{\boldsymbol{\eta}}^{(\nu)}$'s to calculate the weights and inscribe the information coming into the system, the network dynamics is still capable of recognizing and associatively recalling the raw patterns, the $\boldsymbol{\xi}^{(\nu)}$'s. (Note that the input information need not be mutually orthogonal. Refer to [49] for a study on storing orthogonal patterns in the Hopfield network.)

Our analysis shows that the network can recognize (or retrieve, depending on the pattern) the $\hat{\boldsymbol{\eta}}^{(\nu)}$'s as well as $\boldsymbol{\xi}^{(\nu)}$'s for upto $p = N - 1$. However, the case of recall is not as straightforward. We will first list out some of our observations before going into a detailed discussion of associative recall in the H-H-GS model.

4.5.1 Basins of attraction in the H-H-GS network

It is clear from our discussion earlier in the chapter that the stability of the memories is critically linked to the basins of attraction. We have seen in the previous chapter that the GS scheme removes the noise in the system. So, we now examine how this absence of noise affects the radii of the basins of attraction.

TABLE 4.7: Table showing the basins of attraction of some of the raw patterns, $\boldsymbol{\xi}^{(\nu)}$'s, for different values of p for $N = 100$ after invoking orthogonalization. The basins are fairly large and uniform.

	$\boldsymbol{\xi}^{(\nu)}$	Hamming distances forming the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$									
$p = 10$	$\boldsymbol{\xi}^{(1)}$	38	38	35	33	36	42	44	43	34	42
	$\boldsymbol{\xi}^{(5)}$	40	36	39	42	43	43	37	44	41	41
	$\boldsymbol{\xi}^{(10)}$	37	43	34	41	36	37	41	37	42	40
$p = 20$	$\boldsymbol{\xi}^{(1)}$	32	32	29	34	29	34	31	29	36	37
	$\boldsymbol{\xi}^{(10)}$	35	32	32	31	40	30	33	35	33	39
	$\boldsymbol{\xi}^{(20)}$	35	36	34	32	39	33	28	34	26	31
$p = 30$	$\boldsymbol{\xi}^{(1)}$	28	28	28	28	28	26	32	27	29	28
	$\boldsymbol{\xi}^{(15)}$	25	28	29	26	29	31	21	24	20	27
	$\boldsymbol{\xi}^{(30)}$	26	30	29	25	22	32	30	28	28	27
$p = 40$	$\boldsymbol{\xi}^{(1)}$	23	24	17	19	20	18	19	19	17	16
	$\boldsymbol{\xi}^{(20)}$	20	23	20	23	17	16	20	22	20	21
	$\boldsymbol{\xi}^{(40)}$	22	17	19	19	18	17	19	19	18	18
$p = 50$	$\boldsymbol{\xi}^{(1)}$	8	6	10	9	8	7	11	10	8	8
	$\boldsymbol{\xi}^{(25)}$	8	8	9	11	11	7	10	9	10	12
	$\boldsymbol{\xi}^{(50)}$	11	6	12	9	7	8	8	7	12	11
$p = 60$	$\boldsymbol{\xi}^{(1)}$	4	3	5	6	3	3	4	5	4	7
	$\boldsymbol{\xi}^{(30)}$	6	5	4	4	4	3	5	3	3	6
	$\boldsymbol{\xi}^{(60)}$	5	5	4	5	4	3	4	4	3	3

Table4.7 shows the basins of attraction of some of the learnt patterns ($\boldsymbol{\xi}^{(\nu)}$'s). We can see that the radii of the basins are quite big and rather uniform at low memory loads. As p increases, the radii become smaller, but the basins are still fairly isotropic. This scenario is in sharp contrast to the H-H model in which some patterns may have

non-uniform basins or may not even have a basin of attraction for p/N as small as 0.1.

4.5.2 Network dynamics and energy landscape subsequent to orthogonalization

While our study is focused on the raw patterns, the $\boldsymbol{\xi}^{(\nu)}$'s, we must mention that in the H-H-GS model, both $\hat{\boldsymbol{\eta}}^{(\nu)}$'s and $\boldsymbol{\xi}^{(\nu)}$'s are global minima in the energy landscape and are all recognized by the network as long as $p \leq N-1$. Moreover, the elimination of the noise in the system also prevents the minima from being displaced. Hence, in contrast to the Hopfield network, all the learnt patterns are energy minima.

We can now work out the energy $E(\boldsymbol{\xi}^{(\mu)})$ of the pattern $\boldsymbol{\xi}^{(\mu)}$ as

$$E(\boldsymbol{\xi}^{(\mu)}) = -\frac{N}{2} + \frac{N}{2}\left(\mathcal{O}\left(\frac{p}{N}\right)\right), \tag{4.9}$$

from eq.(4.6) and taking into account that the synaptic weights matrix is obtained from the $\{\hat{\boldsymbol{\eta}}\}$. We can see that the energy of the pattern is determined by and proportional to the memory load p. This is in contrast to the H-H model where the energies of the attractors remain distributed around the same mean value irrespective of the value of p. As the number of memorized patterns increases, so do their energies, as we can see from Fig.4.4**(A)**. The distribution of the radii of the basins of attraction of the $\boldsymbol{\xi}$'s for increasing memory loads is shown in Fig.4.4**(B)**. The size of the basin is calculated for each pattern, for various p. The complete set of Hamming distances representing the basins of attraction are then binned using the interval $[0, N/2]$. This interval represents the range of values that the basin radius can take, as we have seen from the analysis in Sec.2.5 (of Chapter2) that the minimum and maximum values the basin size can take are 0 and $N/2$ respectively. The histograms show the probability of the basin sizes taking each of the terms in this interval. As p goes up, there is a corresponding increase in the energies and a reduction in the basin sizes.

We can also verify theoretically (following [20]) that $\boldsymbol{\xi}$'s form minima and are hence attractors even though it is the $\boldsymbol{\eta}$'s that are inscribed in the network. The energy $E(\boldsymbol{\xi}^{(\mu)})$ corresponding to a pattern $\boldsymbol{\xi}^{(\mu)}$ is given by eq.(4.6). Now, on inscribing a

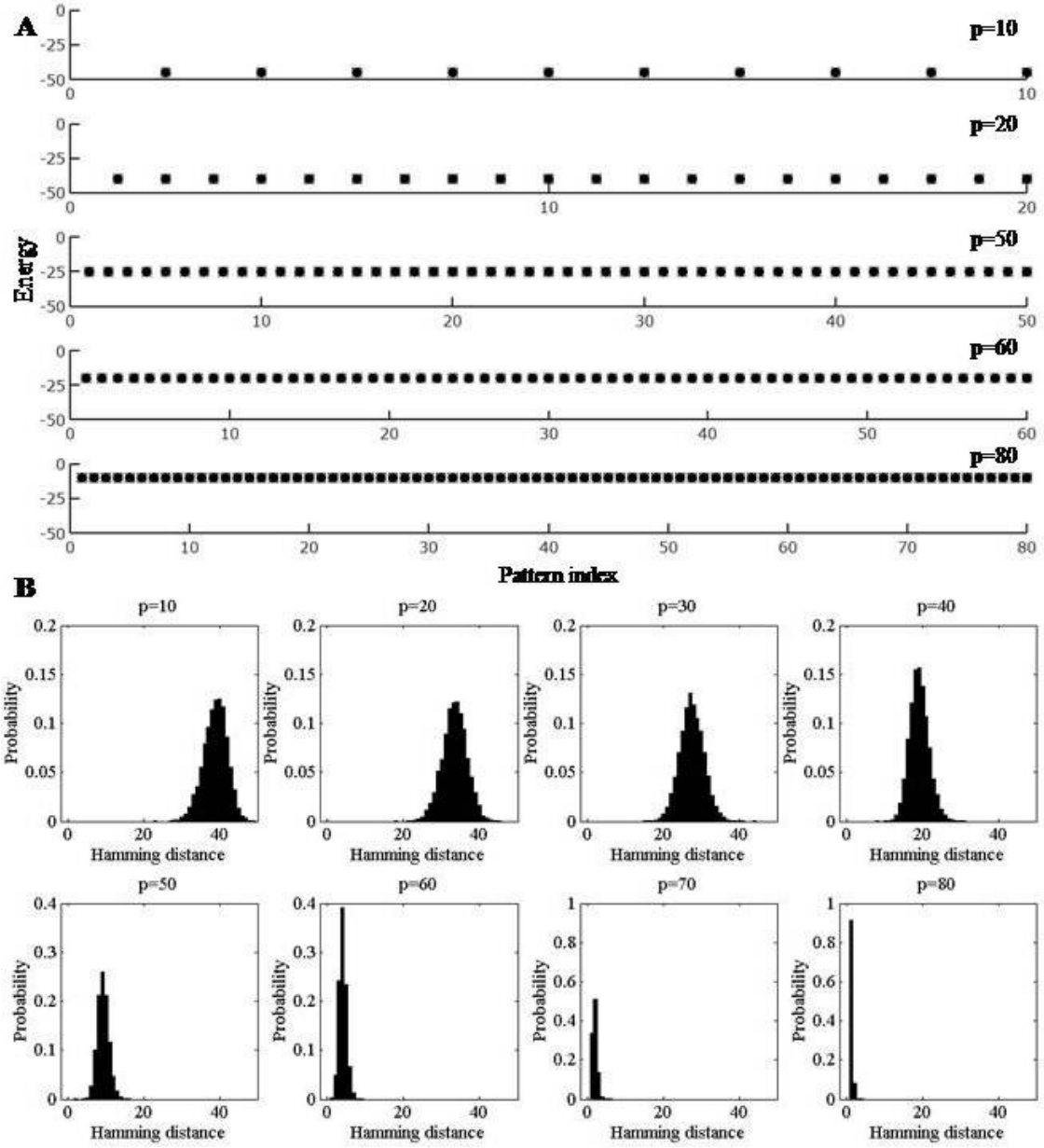


FIGURE 4.4: Fig.**(A)** shows the energy of the $\boldsymbol{\xi}$'s (■) and the attractors or energy minima(□) for various values of p following the orthogonalization scheme for $N = 100$. The x-axis gives the pattern index, or pattern number for each value of p. As the $\boldsymbol{\xi}$'s all form energy minima, the ■ and □ coincide, indicating that there are no instances of minima shifting with increasing p. As p increases, the energy of the $\boldsymbol{\xi}$'s go up, but they remain as energy minima, i.e., with rising p, the valleys in the energy landscape shrink in size and the bottoms are at higher energy. Fig.**(B)** illustrates the evolution of basins of attraction after orthogonalization. The histograms show the distribution of the Hamming distances in the basins of attraction for different values of p after orthogonalization for 50 trials with $N = 100$. Each Hamming distance pertains to one particular sample of a pattern from a particular trial. The x-axis gives the range of values that Hamming distances can take, while the y-axis shows the probability for each of these values. The probability (P) is calculated as $P = f(x)/c$, where $f(x)$ gives the number of Hamming distances with value x, and c is the total number of Hamming distances. The value of c is given by $p * s * T$, where p gives the number of patterns, s is the number of samples and T gives the number of trials. Note that the basins are initially large and relatively isotropic. Zeroes begin to appear around $p = 60$ and by $p = 80$, the basins of attraction are dominated by 0's.

single vector $\boldsymbol{\xi}^{(1)}$ in the network, the synaptic weights are given by

$$\begin{aligned} J^{(1)} &= \frac{1}{N}\sum_{\substack{i,j=1 \\ i\neq j}}^{N} \xi_i^{(1)}\xi_j^{(1)} \\ &= \frac{1}{N}\left(\sum_{i,j=1}^{N} \xi_i^{(1)}\xi_j^{(1)} - \xi_i^{(1)}\xi_i^{(1)}\right). \end{aligned} \tag{4.10}$$

This can be rewritten as

$$J^{(1)} = \boldsymbol{\xi}^{(1)\,T}\boldsymbol{\xi}^{(1)} - I, \tag{4.11}$$

where I represents the $N \times N$ identity matrix. This identity matrix I ensures a zero diagonal in the weights matrix because $\boldsymbol{\xi^{(1)}} \cdot \boldsymbol{\xi^{(1)}} = 1$ for any pattern ν whose components are ± 1. It performs the function of removing self-connections from the weights matrix. The outer product of $\boldsymbol{\xi}^{(1)}$ with its transpose, $\boldsymbol{\xi^{(1)\,T}\xi^{(1)}}$ yields an $N \times N$ matrix . Calculating the energy(4.6) for a random test pattern $\boldsymbol{\xi}^{(\nu)}$,

$$\begin{aligned} E(\boldsymbol{\xi}^{(\nu)}) &= -\frac{1}{2}\boldsymbol{\xi}^{(\nu)} J \left(\boldsymbol{\xi}^{(\nu)}\right)^T \\ &= -\frac{1}{2}\left(\boldsymbol{\xi}^{(\nu)}\boldsymbol{\xi}^{(1)T}\boldsymbol{\xi}^{(1)}\boldsymbol{\xi}^{(\nu)T} - \boldsymbol{\xi}^{(\nu)}\boldsymbol{\xi}^{(\nu)\,T}\right), \end{aligned} \tag{4.12}$$

the minimum of the energy function of the networks with synaptic weights given by $J^{(1)}$ is located at $\boldsymbol{\xi}^{(1)}$. Then $\boldsymbol{\xi^{(\nu)}} = \boldsymbol{\xi}^{(1)}$ is a minimum of the energy function as

$$E(\boldsymbol{\xi}^{(\nu)}) = -\frac{1}{2}\|\boldsymbol{\xi}^{(\nu)} J \boldsymbol{\xi}^{(\nu)\,T}\|^2 + \frac{N}{2}, \tag{4.13}$$

since

$\boldsymbol{\xi}^{(\nu)}\boldsymbol{\xi}^{(1)\,T}\boldsymbol{\xi}^{(1)}\boldsymbol{\xi}^{(\nu)T} = \boldsymbol{\xi}^{(\nu)}\boldsymbol{\xi}^{(1)\,T}\left(\boldsymbol{\xi}^{(\nu)}\boldsymbol{\xi}^{(1)\,T}\right)^T = \|\boldsymbol{\xi}^{(\nu)}(\boldsymbol{\xi}^{(1)\,T}\|^2$ and $\boldsymbol{\xi}^{(1)}\cdot\boldsymbol{\xi}^{(\nu)} = N$ for $\xi_i^{(\nu)} = \pm 1$.

We then have,

$$E(\boldsymbol{\xi}^{(\nu)}) = -\frac{N^2}{2} + \frac{N}{2}, \tag{4.14}$$

whose value will be < 0 indicating that the energy has decreased, thereby making $\boldsymbol{\xi}^{(1)}$ a stable state of the network.

After p patterns have been inscribed in the network, the weight matrix is given by

$$J = \sum_{\mu=1}^{p} \boldsymbol{\xi}^{(\mu)T}\boldsymbol{\xi}^{(\mu)} - pI. \tag{4.15}$$

We can then calculate $\boldsymbol{h}^{(1)}$ as

$$\begin{aligned}\boldsymbol{h}^{(1)} &= \boldsymbol{\xi}^{(1)} J \\ &= \boldsymbol{\xi}^{(1)} \left(\sum_{\mu=1}^{p} \boldsymbol{\xi}^{(1)T} \boldsymbol{\xi}^{(1)} - pI \right) \\ &= \boldsymbol{\xi}^{(1)} \boldsymbol{\xi}^{(1)T} \boldsymbol{\xi}^{(1)} + \boldsymbol{\xi}^{(1)} \boldsymbol{\xi}^{(2)^T} \boldsymbol{\xi}^{(2)} + \ldots + \boldsymbol{\xi}^{(1)} \boldsymbol{\xi}^{(p)T} \boldsymbol{\xi}^{(p)} - p\boldsymbol{\xi}^{(1)} I \\ &= N\boldsymbol{\xi}^{(1)} + \sum_{\mu=2}^{p} \boldsymbol{\xi}^{(1)} \boldsymbol{\xi}^{(\mu)^T} \boldsymbol{\xi}^{(\mu)} - p\boldsymbol{\xi}^{(1)},\end{aligned} \tag{4.16}$$

as $\boldsymbol{\xi}^{(1)} \boldsymbol{\xi}^{(1)^T} = N$. Thus,

$$\boldsymbol{h}^{(1)} = (N - p)\, \boldsymbol{\xi}^{(1)} + \sum_{\mu=2}^{p} \alpha_{1\mu} \boldsymbol{\xi}^{(1)}, \tag{4.17}$$

where $\alpha_{1\mu}$ gives the correlation between the inscribed pattern $\boldsymbol{\xi}^{(1)}$ and each of the remaining $p-1$ patterns. As long as $p \ll N$, the second term in eqn.(4.17) would be negligible and the pattern $\boldsymbol{\xi}^{(1)}$ would be stable: it would have a positive stabilization parameter(4.5). We can show similarly that each of the remaining p inscribed patterns is stable and a minimum in the energy landscape.

In order to make our study comprehensive, we also test whether the $\boldsymbol{\xi}$'s can still form spurious states, given that it is their orthogonalized forms that is used in the inscription process. Our results show that the mirror states, or the inverses of the $\boldsymbol{\xi}$'s are minima, as are combination states which are present at low values of p, and converge to one of their constituent states, as shown in Fig.4.5. As p increases, to 50, say, the mixture states are no longer stable. This is logical, given that the radii of the basins of even the $\boldsymbol{\xi}$'s are miniscule at high memory loads.

4.5.3 Improvement in the memory capacity

Our results show that, following orthogonalization, all the $\boldsymbol{\xi}$'s are minima in the energy landscape and are recognized by the network as long as $p \leq N - 1$.

For these values of p, the stabilization parameter $\mathfrak{s}_i^{(\nu)} > 0$ and will have positive values on all neurons i. This is due to the fact that the orthogonalization process removes all the noise arising from the crosstalk between the patterns. However, the magnitude of the $\mathfrak{s}_i^{(\nu)}$'s decrease as $p \to N - 1$, though they are still > 0(/positive).

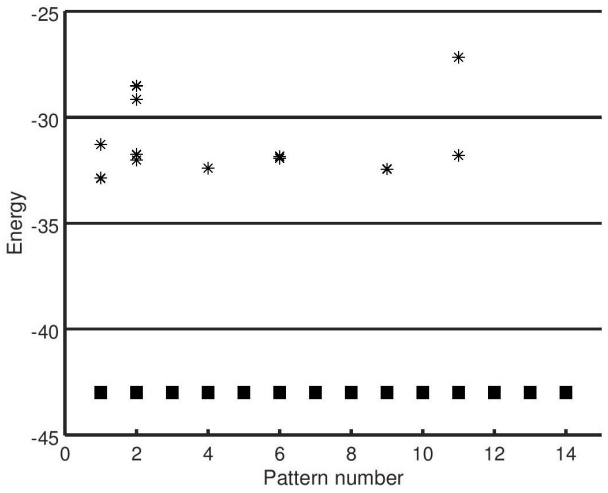


FIGURE 4.5: Plot showing the energy (from eq.(4.9)) of the raw patterns, $\boldsymbol{\xi}$'s (■) and the corresponding attractors(□) for $p = 14$ following the orthogonalization scheme. ■ and □ overlap each other completely for all the patterns. Some mixture states (formed from combinations of $\boldsymbol{\xi}$'s, denoted by ($*$'s)) are also shown for $p = 14$, plotted against the pattern $\boldsymbol{\xi}$ with which they have maximum overlap. The mixture states shown here lack basins of attraction of their own and typically converge to one of their component patterns. Refer to Fig.4.3**(A)** for comparison.

At the same time, there is an increase in the energies of the patterns, while the basins of attraction shrink. In due course, as p reaches a limit ($0.63N$, shown in Fig.4.6), the basins of attraction of some of the memorized patterns vanish. Within this limit, the patterns all have non-zero basins of attraction, while the number of patterns lacking a basin of attraction goes up once the limit is crossed. This limit delineates the upper bound for associative recall by the network. However, the network remains capable of recognition well beyond this bound.

The basins of attraction of the memories in an efficient associative memory should ideally be big and uniform in size and shape. The memorized patterns in the H-H model possess large and uniform basins of attraction, but only when p is very small (as seen in Table4.8). Modifying the learning prescription (as in ref. [65]) is one way of achieving the preferred traits in basin properties. However, the basins of attraction of the memories in the H-H-GS model are large and uniform or isotropic without the need for any such modifications to the learning rule, nor were the input patterns pre-processed in any way. This upgrade in the efficacy of the network as an associative memory is the result of the process of orthogonalization itself.

The changes in the energies of the attractors and basin radii in the H-H and H-H-GS models can be seen from Fig.4.7.

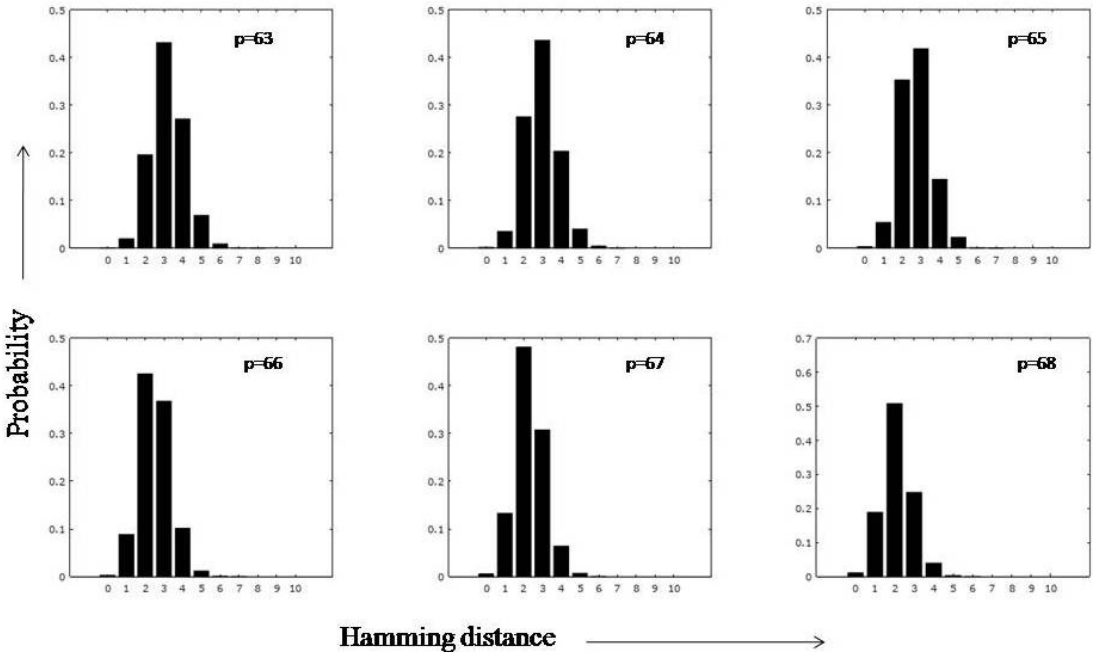


FIGURE 4.6: Histograms showing the capacity for associative recall in the H-H-GS network. The plots show the data for $N = 100$. Beyond $p = 63$, zero basins of attraction begin to appear and grow in number as p increases

TABLE 4.8: Table showing the basin of attraction of a stable pattern ($\xi^{(1)}$) for different values of p ($p = 2, 4, 6$) for $N = 100$. The basin is fairly uniform/isotropic for low values of p.

Samples	1	2	3	4	5	6	7	8	9	10
$p = 2$	43	49	42	49	48	43	48	47	42	46
$p = 4$	43	36	40	45	37	43	45	42	46	38
$p = 6$	35	41	37	38	29	35	40	39	38	42

4.6 Ramifications of correlations in the H-H and H-H-GS models

While the H-H-GS model removes the overlaps affecting pattern stability, and learns the orthogonalized patterns, it would be useful to study the role of overlaps between the raw patterns in the network dynamics of the model. We look at the extreme case where there is a very high degree of similarity between the patterns to be memorized. We first check to see if such patterns can be stored in both the H-H and H-H-GS models, and whether the high amount of overlaps affects the processes of recognition

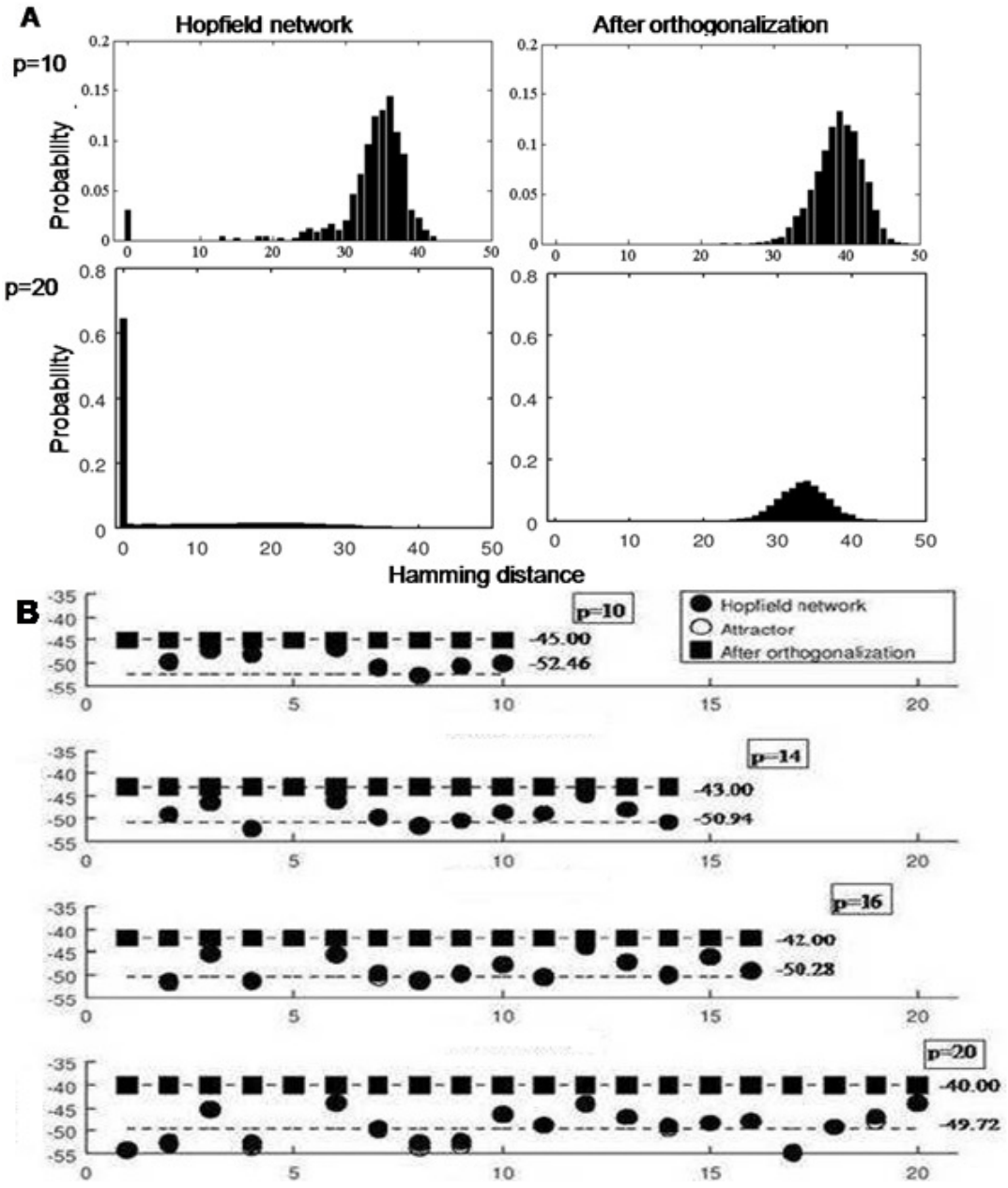


FIGURE 4.7: The histograms in Fig. **(A)** show the distribution of the Hamming distances in the basins of attraction for $p = 10$ and $p = 20$ in the Hopfield network and in the network after orthogonalization for 50 trials. When $p = 10$, some of the basins of attraction in the Hopfield network become zeroes, by $p = 20$, the basins of attraction are predominantly 0. By contrast, after orthogonalization, the sizes of the basins are larger and the distribution of the Hamming distances is also relatively uniform. Fig.**(B)** plots the energy of the $\boldsymbol{\xi}$'s and the corresponding attractors $\boldsymbol{\xi}'$'s for various values of p ($p = 10, 14, 16, 20$) for both the Hopfield network and orthogonalization scheme. In the Hopfield network, as p increases, some of the $\boldsymbol{\xi}$'s are no longer minima. However, the energies of the attractors remains distributed around a mean (close to $0.5N$). The average energy of the inscribed patterns is shown next to each value of p. After orthogonalization, all the $\boldsymbol{\xi}$'s have uniform energy for a particular value of p, and while the energy increase with the number of patterns, the $\boldsymbol{\xi}$'s remain minima.

and/or recall. We show that the H-H-GS model is capable of distinguishing between patterns that are very similar to each other. While very similar patterns in the Hopfield model tend to fall within the same basin of attraction, after orthogonalization, they become two separate and distinct attractors, though they may or may not have basins of attraction. In terms of cognition, an attractor pertaining to a $\boldsymbol{\xi}^{(\nu)}$ is a separate category and may have a set of items associated with it. After orthogonalization, the network still remains capable of identifying and separating individual categories, irrespective of the amount of correlations between the $\boldsymbol{\xi}$'s.

If we try to store two very similar(98% match) patterns in the Hopfield model, then either one or both the patterns become unstable - one of the inscribed patterns will be in the basin of the other, or they may both fall within the basin of a different attractor altogether. This new attractor would be very similar (99%, say) to both of those inscribed patterns. Table4.9 shows an example of each of these cases. In terms of cognition, both the similar patterns would belong to the same category.

Even at very small values of p, high levels of similarity between the patterns contribute significantly to the slow noise in the Hopfield network.

But in the H-H-GS model, high amounts of overlaps do not affect pattern stability, though it causes changes in the basin radii and shapes. Table4.10 shows the basins of attraction in a H-H-GS network with $p = 5$, out of which two have a high degree (98%) of overlap. We see that though both the similar patterns are attractors and can be recognized by the network, their basins are no longer uniform - the extents of the basins get constrained in some directions.

TABLE 4.9: Table showing examples of where patterns may converge in multiple trials in the Hopfield network and in the network after Gram-Schmidt orthogonalization when two very similar patterns are inscribed in a network of size $N = 100$. We first inscribe 4 patterns ($\boldsymbol{\xi}^{(1)}$ to $\boldsymbol{\xi}^{(4)}$) in the network, then choose $\boldsymbol{\xi}^{(\nu)}$ randomly from one of these 4 to make a fifth pattern $\boldsymbol{\xi}^{(5)}$ which is similar (98% similarity) to it, and differs on the sites marked as 'Differences'. We now store this fifth pattern and check whether all the 5 patterns are attractors. The inscribed patterns other than the chosen $\boldsymbol{\xi}^{(\nu)}$ remain attractors even as p changes from 4 to 5. We are interested in the situation with the two similar patterns, and whether $\boldsymbol{\xi}^{(\nu)}$ and $\boldsymbol{\xi}^{(5)}$ are attractors (✓) or not (×), when $p = 5$. When $\boldsymbol{\xi}^{(5)}$ is presented to the network, three possibilities arise: (i) $\boldsymbol{\xi}^{(5)}$ falls within the basin of attraction of $\boldsymbol{\xi}^{(\nu)}$ and converges there (Case 1), (ii) $\boldsymbol{\xi}^{(5)}$ is an attractor, and $\boldsymbol{\xi}^{(\nu)}$ which was previously stable now falls within the basin of attraction of $\boldsymbol{\xi}^{(5)}$ (Case 2), or (iii) both $\boldsymbol{\xi}^{(5)}$ and $\boldsymbol{\xi}^{(\nu)}$ converge to a third pattern which is not any of the inscribed patterns and has 99% overlap with both $\boldsymbol{\xi}^{(5)}$ and $\boldsymbol{\xi}^{(\nu)}$ (Case 3). We have used examples from different trials to illustrate each of these cases.

	$\boldsymbol{\xi^{(\nu)} \equiv \xi^{(5)}}$ (98% similarity) (ν=1–4, chosen randomly)			Status of $\boldsymbol{\xi}^{(\nu)}$ and $\boldsymbol{\xi}^{(5)}$ when $p = 5$					
				Hopfield network			After orthogonalization		
	Case	$\boldsymbol{\xi}^{(\nu)}$	Differnces	$\boldsymbol{\xi}^{(\nu)}$	$\boldsymbol{\xi}^{(5)}$	Remarks	$\boldsymbol{\xi}^{(\nu)}$	$\boldsymbol{\xi}^{(5)}$	Remarks
Randomly generated patterns $\boldsymbol{\xi}^{(\mu)}$, $\mu = 1-4$, ($\boldsymbol{\xi}^{(1)}$ - $\boldsymbol{\xi}^{(4)}$)	1	$\boldsymbol{\xi}^{(1)}$	56,71	✓	×	$\boldsymbol{\xi}^{(1)}$ is an attractor, $\boldsymbol{\xi}^{(5)}$ falls within its basin of attraction.	✓	✓	$\boldsymbol{\xi}^{(1)}$ and $\boldsymbol{\xi}^{(5)}$ are both attractors
	2	$\boldsymbol{\xi}^{(4)}$	15,62	×	✓	$\boldsymbol{\xi}^{(5)}$ is an attractor, $\boldsymbol{\xi}^{(4)}$ falls within its basin of attraction.	✓	✓	$\boldsymbol{\xi}^{(4)}$ and $\boldsymbol{\xi}^{(5)}$ are both attractors
	3	$\boldsymbol{\xi}^{(2)}$	52,91	×	×	$\boldsymbol{\xi}^{(2)}$ and $\boldsymbol{\xi}^{(5)}$ both fall within the basin of a third attractor distict from both but 99% similar to each.	✓	✓	$\boldsymbol{\xi}^{(2)}$ and $\boldsymbol{\xi}^{(5)}$ are both attractors

Our computations show that the above results hold true even at higher memory loads ($p = 10, 20, 30 \ldots, 90$). In the network with orthogonalization, a high level of correlations between even two of the memorized patterns disturbs the sizes and shapes of the basins, making them anisotropic when their radii are non-zero. However, recognition remains intact even in the presence of high correlations. This is in sharp contrast to the behaviour of the H-H network, where high levels of similarity between the patterns are detrimental to both recognition and recall. This is because the two processes are interconnected in the Hopfield model where recognition necessarily implies recall as the attractors always have non-zero basins. The inverse is also true, as attractors are by nature stable fixed points in the network dynamics, and so recall automatically indicates recognition.

This analysis brings us to an important observation. We have seen that as p goes up, more and more patterns imprinted in the Hopfield network become unstable and cease to be attractors. But, after orthogonalization, all the imprinted patterns remain stable in that their recognition remains intact, though their recollection is not always assured. As $p/N \to 1$,recall of an imprinted pattern through an erroneous version of the pattern acquires a novel status. This is understandable, as the basin radii are reduced for rising p, while the minima remain in place. Beyond the limit $p \approx 0.63N$ established earlier, some of the basins shrink so much that they resemble inverted $\delta-$functions, that is, the basins consist of just the inscribed patterns themselves.

There are no patterns associated with such imprinted patterns, not even those closest to the imprinted patterns. A basin is, in essence, a category, with the set of patterns within the basin being the items associated with the category label or the attractor. In this context, we can visualize the $\delta-$function-basins as a representation of the extreme situation where each of the inscribed patterns is a category by itself. That is, each imprinted pattern is a category label and also the only item in the category. In order to reach and identify the item/category, we must specify the information exactly and in its entirety. Such situations are encountered often in real life, and we can see here how they can be modelled using the H-H-GS model.

4.7 Discussion

The memory capacity of the network is enhanced following orthogonalization. The network can recognize all the $\boldsymbol{\xi}$'s, as they are all global minima for upto $p = N - 1$.

TABLE 4.10: Basins of attraction for p=4 and p=5, with $\boldsymbol{\xi}^{(5)}$ (98%) similar to $\boldsymbol{\xi}^{(2)}$. Following the addition of a pattern very similar to one of the previously inscribed patterns, the basin of attraction of $\boldsymbol{\xi}^{(5)}$ in the Hopfield network(HN) vanishes or becomes 0. After orthogonalization (GS) , the basins of attraction are all non-zero, however, the addition of a similar pattern makes the basins anisotropic.

	$\boldsymbol{\xi}^{(\nu)}$		Hamming distances in the basins of attraction									
$p=4$	$\boldsymbol{\xi}^{(1)}$	HN	46	34	39	41	47	40	46	43	38	46
		GS	42	40	46	41	38	38	32	45	40	43
	$\boldsymbol{\xi}^{(2)}$	HN	44	41	40	41	47	45	39	45	28	44
		GS	42	41	43	34	37	44	40	39	38	44
	$\boldsymbol{\xi}^{(3)}$	HN	39	44	46	37	34	45	39	28	45	48
		GS	47	47	43	41	35	31	45	45	43	37
	$\boldsymbol{\xi}^{(4)}$	HN	46	35	41	27	42	40	43	43	33	44
		GS	39	41	43	44	41	38	40	46	39	37
$p=5$	$\boldsymbol{\xi}^{(1)}$	HN	38	20	41	39	15	38	43	42	25	42
		GS	15	34	38	27	5	45	27	38	34	14
	$\boldsymbol{\xi}^{(2)}$	HN	48	47	42	44	50	48	47	46	47	50
		GS	8	29	13	25	39	6	16	34	5	38
	$\boldsymbol{\xi}^{(3)}$	HN	31	37	29	43	15	17	43	25	16	31
		GS	41	47	45	44	43	42	45	44	46	33
	$\boldsymbol{\xi}^{(4)}$	HN	40	40	38	34	34	40	26	44	38	33
		GS	4	41	38	42	18	4	39	2	43	30
	$\boldsymbol{\xi}^{(5)}$	HN	0	0	0	0	0	0	0	0	0	0
		GS	38	30	7	16	16	30	41	44	44	42

This is an improvement from the H-H network whose memory capacity is $\approx p = 0.1N$. This limit is the same for both recognition and recall, as they are interlinked. However, after orthogonalization, the capacity for associative recall increases to about $p = 0.63N$, the point till which the $\boldsymbol{\xi}$'s all have non-zero basins. Recognition does not always guarantee recall in the H-H-GS model.

Moreover, it is clear from our discussion earlier in the chapter that after orthogonalization, the memorized patterns satisfy the criterion 1 for stability whether or not they satisfy 2. This led us to separate the criteria for pattern stability into a necessary condition (recognition) and a sufficient one (recall).

We must also point out that the processes of *pattern separation* and *pattern completion* are separable in our model. Pattern separation is the process of separating and identifying individual patterns no matter how much they overlap with each other, and is basically the orthogonalization procedure itself. Pattern completion is the process of accurate associative recall of an imprinted pattern when the network is presented with an incorrect or partial version of the same. There is some biological evidence of that these two processes are dissociated and pertain to different parts of the hippocampus [55, 56].

Our study details the active changes to the energy landscape as a result of growing amount of memorized information. Nearly all aspects of these changes have some bearing on cognition. A basin of attraction is essentially a valley, or a category containing items similar to the category label pattern, the attractor which lies at the bottom. As new information enters the network and gets added to the memory store, the categories naturally get rearranged and honed. In the context of cognition, we can say that the introduction of more and newer information causes a change in perception of some of the previously known information. This phenomenon is widely known and, corresponds to the displacement of minima in the H-H model where some of the minima pertaining earlier to the imprinted patterns may shift to other novel patterns.

However, there does not appear to be a cognitive analogue of the catastrophic blackout in the Hopfield network where there is neither recognition nor recall if the number of learnt patterns is too high. This is where the orthogonalization scheme plays a role and makes the model brain more robust with respect to retrieval/recognition/recall as well as in classification and categorization.

Evidence from biology indicate the presence of attractor dynamics in the cerebellum and hippocampus, and CA3 in particular [66, 67]. These areas could possibly perform the process of orthogonalization. For a model of recognition in the perirhinal cortex, see [68, 69].

4.8 Some issues related to the H-H and H-H-GS networks

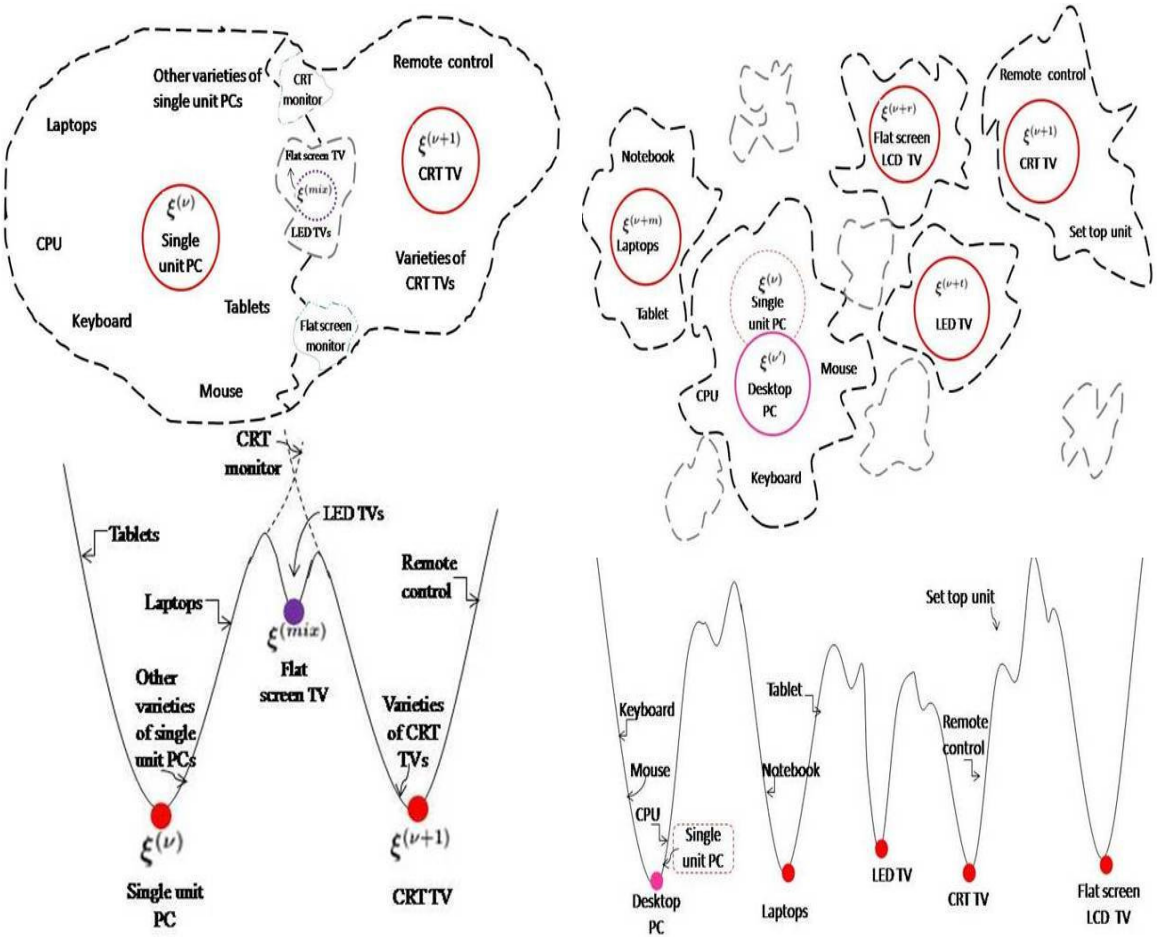
Some more issues related to the Hopfield network and the model with orthogonalization are discussed below:

4.8.1 Cognitive relevance of the network dynamics of the Hopfield network

When patterns are inscribed in the network, the energy landscape is partitioned into basins of attraction of the attractors. This is equivalent to the process of categorization in which the attractors can be considered to be the 'category labels' or 'representatives' of the categories. The basins of attraction give the set of patterns in each category, i.e. those associated with the category label [7, 8].

We present below in Fig.4.8 an example of categorization and the minima shifting described earlier. Consider for instance a network whose dynamics now includes a category called 'Computers', with the category label or default representation corresponding to a single unit PC, such as a Mac, which we hapenn to come across in the beginning. Other items within the category could include 'laptops' and 'tablets'. When we come across more versions of computers which means more patterns are inscribed and we happen to come across relatively older versions of computers, the representation of a computer could change to a traditional desktop PC with a separate CPU, while single unit PC's would still be associated with computers. This situation would still be associated with computers. This situation would be akin to the shiftig of minima in the energy landscape. As more patterns are added, a large number of which are related to laptops, say, then 'laptops' would now be a separate category with a set of patterns associated with it. A subset of patterns that was previously associated with generic computers would now be associated with

laptops. This corresponds to the reduction of basin radii with the accumulation of more patterns. There would also be sub-categories pertaining to spurious minima. For instance, there could be a smaller category in between 'TV' and 'Computer'. A flat-screen TV may fall into this category. Such categories have fewer items within them.



(a) Examples of categories: The inscribed patterns corresponding to single unit PC ($\boldsymbol{\xi^{(\nu)}}$) and CRT TV ($\boldsymbol{\xi^{(\nu+1)}}$) form attractors (red circles), with the black dashed lines marking their basins of attraction. The pattern corresponding to Flat screen TV may be a spurious minimum ($\boldsymbol{\xi^{(mix)}}$), with a small basin of attraction shown as grey dashed lines. Other patterns such as CRT monitor and Flat screen monitor are similar to both $\boldsymbol{\xi^{(\nu)}}$ and $\boldsymbol{\xi^{(\nu+1)}}$, and eventually converge to one of them. The red circles in the energy landscape pertain to the attractors, while the purple circle is the minimum corresponding to the mixture state.

(b) Modifications in the categories, the scenario after many more patterns have been inscribed in the network: The basins of attraction become smaller and more anisotropic. Some of the inscribed patterns may become unstable: $\boldsymbol{\xi^{(\nu)}}$ corresponding to single unit PC is no longer a minimum but converges to a new attractor $\boldsymbol{\xi^{(\nu')}}$. The number of spurious states also increases with increasing p.

FIGURE 4.8: Examples of categorization

When the number of patterns is small, the categories are broad and there would be more 'items' (i.e., patterns) in each category. As more patterns are inscribed, the

classification gets increasingly fine-tuned and more categories are created, resulting in a decrease in the number of items in each category (shrinking of basins of attraction). Further increase in the number of inscribed patterns leads to some of the inscribed patterns becoming unstable with the network no longer being able to recognize or recall these unstable patterns in that when presented for association the convergence does not happen on them, instead it happens on a nearby pattern. As a result of shifting of minima, the unstable patterns themselves fall within the basins of attraction of the newly found attractors, referred to as new attractors. These new attractors are initially very close to the nearest unstable inscribed pattern but move farther away as p increases. The presence of these new attractors indicates that the category labels have changed while the inscribed patterns can still be within the same category. Cognitively, this is like a change in perception.

4.8.2 An issue with definition of forgetting

We have attributed precise criteria to stability of patterns. We now take another look at those inscribed patterns that were stable to start with (in accordance with the strict condition in eq.(2.9)) but become unstable with increasing p. Should they be discarded as irrelevant? When presented for association, these patterns are not retrieved or recognized in the manner we have defined, instead the iteration procedure converges to a novel attractor. We have given this a cognitive interpretation that with increasing assimilation of information (i.e. increasing p) the perception associated with the inscribed information gets modified and the new attractor represents the new perception. So, as we have asked above, is the original inscribed pattern forgotten?

In the present scheme of definition of eq.(2.9) the answer would be 'yes'. But intuitively this would not appear to be correct. The inscribed pattern is in the basin of the new attractor, so the two are intimately connected. The inscribed pattern uniquely leads to the attractor. It might appear to be a limitation of the models that only the attractors are treated as stable memories, while the huge number of patterns in their individual basins that help in recalling or connecting with the patterns at the bottom of the basins are dubbed as unstable. It would be inappropriate to treat them as forgotten. They were in the memory and ought to be still there if they lead to recollection albeit in a different form. Qualitatively we can take a larger view and say that all the patterns in the basin of an attractor are familiar to the system and the attractor represents not just a single memory but a memory of a group of

items. This calls for a modification of the mathematical definition of memory such that it can address the problem of treating something as forgotten that was earlier in memory.

In cognitive psychology domain Tulving [70] defines forgetting as the failure in recalling information which could previously be recalled. This failure in recall can be attributed to the absence of adequate cues necessary to retrieve the information that is still present in the system. In our case the iterations that take the inscribed pattern to the new attractor can be viewed as supplying on every iteration those extra cues. However, the difference remains that the extra cues, in our case, do not take us back to something that was once a stable memory or an attractor. We need a mathematical framework that enables us to recall something that could be recalled previously.

After orthogonalization, there is no forgetting since the raw patterns can always be accessed or recognized by the network: the network always converges to the pattern when presented with any of the raw patterns. We reiterate here that there is no shifting of minima in the dynamics of the H-H-GS model, which ensures the recognition of raw patterns. At lower memory loads, the set of patterns within the basin of attraction of each pattern act as a set of cues as they all lead to the attractor, namely the raw pattern, and converge there. Presenting any of these associated patterns is sufficient to recall a raw pattern $\boldsymbol{\xi}^{(\nu)}$. At higher loads, each of the raw patterns becomes an individual category. As the categorization is now very strict, each pattern is now the only item within and associated with a category and therefore the patterns $\boldsymbol{\xi}$'s are themselves the only and unique cues for the retrieval of $\boldsymbol{\xi}$'s.

4.8.3 Two pertinent issues about the new minima or attractors

First, since they do not belong to any of the inscribed patterns, strictly speaking in our scheme of formulation, such minima ought to be termed as spurious. But that would be, for quite obvious reasons, inappropriate as is evident from the above discussion. So, the definition of spurious states needs to be modified in the conventional associative neural network models.

The second issue is about making cognitive sense out of the distance between an unstable inscribed pattern and the corresponding new attractor. Starting from one

or two mismatches for very low values of p (say $0.1N$) the new attractor progressively moves further away from the inscribed pattern as p increases. While we have interpreted a few mismatches for small p as indicating changing perception, it is hard to make a simple interpretation when the mismatch is on say 20% of the sites. It is also hard to draw a line to demarcate the highest percentage of mismatches that would be cognitively relevant.

4.8.4 Information contained in a pattern

The input patterns, the $\boldsymbol{\xi}$'s are N-dimensional vectors whose components represent the various features of an information. In the Hopfield network, these components can take values $+1$ or -1, which we can interpret as the presence $(+1)$ or absence (-1) of a certain feature in a particular pattern $\boldsymbol{\xi}^{(\nu)}$. After orthogonalization, the inscribed patterns, viz. $\boldsymbol{\eta}$'s, can take $\pm$ fractional values. We can consider this range of values to represent a spectrum or continuum ranging from very weak or very strong on either side of zero, that is different degrees to which particular features may be present or absent in a pattern. A value of slightly greater than zero would indicate a weak presence of that feature, while one closer to $1/\sqrt{N}$ would indicate an emphatically strong presence, similarly for negative values. Thus we like to infer that orthogonalization enables a lot more information to be embedded in the network within Hebbian mechanism.

4.8.5 Sharing of the configuration space

In the present state of formulation of Hopfield network, any of the stable inscribed patterns can become unstable as p increases. It is not necessarily the patterns that are stored later that become unstable, but even the earlier patterns may lose their stability. Although with further increase in memory loading, the new attractors corresponding to the unstable inscribed patterns drift further away from the latter, they can have large basins of attraction, which may contain not just the inscribed pattern and some of the patterns previously associated with it, but may also include patterns which were previously not associated with any of the earlier inscribed patterns. In this sense, the network is able to explore more of the configuration space, though this is not of much cognitive significance as the majority of the inscribed patterns are deemed to be unstable at such high loading. The anisotropy in the basins can

also be considered as an exploration of the configuration space, but it is not quite desirable, as uniform basins are supposed to improve the associative performance of the network. We hence do not attach much cognitive significance to it at this stage of formulation of network.

Orthogonalization is merely a transformation of the set of $\boldsymbol{\xi}$'s into an orthogonalized set, and the $\boldsymbol{\eta}$'s span the same space as $\boldsymbol{\xi}$'s. It would thus be logical to conclude that the two sets of patterns and their respective basins of attraction lie in the same space.

However, the signs of $\boldsymbol{\xi}$'s and $\boldsymbol{\eta}$'s match only at very low values of p. The two sets of patterns have similar basins of attraction (same range) for the same value of p. If the signs matched, we could treat the minima pertaining the two sets as coinciding with each other. If they are distinct sets within the same space, then an $\boldsymbol{\xi}^{(\nu)}$ which differs from the corresponding $\boldsymbol{\eta}^{(\nu)}$ by 3 elements, say, would lie fairly close to that $\boldsymbol{\eta}^{(\nu)}$. If they both form attractors within the same space, then where would the patterns falling in the shared region converge to? Should we instead treat the two spaces as separate, with the $\boldsymbol{\xi}$-space containing only patterns whose components are either $+1$ or -1, and a separate $\boldsymbol{\eta}$-space whose components are fractions?

Furthermore, the process of orthogonalization renders all the inscribed patterns 'equivalent' in that they all have the same energy and similar basins of attraction. The energies of the patterns go up as more patterns are inscribed in the network. This is in contrast to the scenario pertaining to the Hopfield network, where the energies of the inscribed patterns are scattered around a mean value (typically around $N/2$) regardless of the value of p, as seen from Fig.4.7.

As we have seen, the correlations between the patterns in the Hopfield network lead to non-uniform basins of attraction and displacement of attractors. We can also see that the presence of correlations between the $\boldsymbol{\xi}$'s in varying degrees influence the size and shape of the basins of attraction even after orthogonalization but do not affect the location of energy minima.

4.9 Conclusion

We have presented exact mathematical definitions of the terms retrieval, recognition and recall. We have also shown that the criteria for memory stability can be expressed

in terms of a necessary and a sufficient condition, namely recognition and recall. After studying the network dynamics and examining the energy landscape of the Hopfield model, we analyzed the H-H-GS model before comparing the two. When orthogonalization is introduced in the network learning, it presents new results for the stability of the memories. While the network improves greatly in terms of its memory capacity and effectiveness as an associative memory, the network is not completely free of the destructive effects of catastrophic interference. However, the CI does not affect recognition, while its negative consequences for associative recall are delayed considerably.

Various schemes[71, 72, 73, 74] have been proposed to increase the memory capacity of the Hopfield network. However, the orthogonalization scheme proposed here has a significant advantage over them - it does need any modifications to either the network structure or the learning prescription.

The biological equivalents of the H-H-GS model and the network dynamics analyzed here need to be investigated and validated experimentally. There is some literature (, including *in vivo* data-based modeling) on the effects of the introduction of orthogonalization in the network dynamics of the H-H model. We must point out here that stability is not an actual default state of a memory. Whenever a memory is recovered, it must be added back to the memory store once more for reconsolidation [75, 76]. This process makes the memory labile and malleable to changes prior to (re-)memorization. The concept of stability in our context is consistent with the idea of lability of memories, though our focus is on the robustness or stability of the memories in the face of newer information being added to the network. Our study must hence be generalized taking into consideration the fact that memories are labile in nature. This can be done by focusing on understanding the interaction between stability and pliability.

It would be of interest to extend the current study and its results to other orthogonalization schemes[3, 77], while also exploring and comparing the behaviour of the system in each case. Another potential avenue to be explored is the situation where the system uses sparse coding.

Chapter 5

Relevance of Löwdin orthogonalization schemes to cognition

So far we have studied the Gram-Schmidt orthogonalization scheme, which is sequential in nature. In this chapter, we will invoke two democratic orthogonalization schemes due to Löwdin, namely Symmetric and Canonical procedures, in the Hebb-Hopfield network. The democratic nature of the schemes refers to the set of vectors being orthogonalized as a whole, whereas the Gram-Schmidt process orthogonalizes individual vectors as they are added, while the previously orthogonalized vectors are retained. We will present some preliminary results and present our hypothesis that these two schemes may have some relevance to the learning and memorization capabilities of the brain, and to the physiological mechanisms underlying certain kinds of memories.

5.1 Introduction

We know that the Gram-Schmidt procedure is sequential in nature - the vectors are orthogonalized and added to the orthonormal basis in the same order as the input set. That is, a new vector presented to the system is made mutually perpendicular to all the vectors currently in the basis. In other words, each new vector is orthogonalized and added to the basis set without disturbing the previously orthogonalized vectors.

It has been discussed that in cognitive terms, the Gram-Schmidt procedure could be applied to semantic memory which includes a sequential process such as language learning[15]. In the previous chapters, we have analyzed in detail the H-H-GS model and discussed its relevance to learning and memory. We now wish to explore if other kinds of memory can be modelled in a similar fashion. For instance, episodic memory, where a memory is stored along with a set of other information including the environment or context in addition to that memory, rather than just one piece of information at a time. Such memories could be regarded as 'snapshots', and are not sequential in nature.

The two orthogonalization schemes due to Löwdin, namely symmetric and canonical orthogonalization [4, 5, 6, 78] are 'democratic' in nature in that the order of presentation does not affect the orthonormal basis. Each time a new vector is added, the complete set of vectors consisting of the new vector and all the previous vectors is used to calculate the new orthonormal basis. Given this, the two schemes could possibly be applied in situations such as the episodic memories mentioned above. In such scenarios, the brain processes and stores multiple information simultaneously - a particular information along with some contextual information, for instance, information about an event along with the environment in which the event occurred. Coming across a similar environment later would trigger the memory of the event in the brain. In other words, the two Löwdin schemes should be able to shed light on the cognitive mechanisms underlying the memorization and recall of memories where the contexts serve as cues to remembering the memories.

The interesting properties of the two schemes have been highlighted earlier by Srivastava [3]. In the case of symmetric orthogonalization, the sums on the squares of the projections of the orthonormalized vectors is the same for each of the vectors in the basis. In the orthonormal basis obtained from canonical orthogonalization, there is a hierarchy among the sums of the squares of the projections, with the sum taking highest value for one of the basis vectors, and the remaining sums following a hierarchical order. This highest value sum also happens to be the maximum value for any possible orthonormal basis.

Prior to examining what the above-mentioned properties of these two schemes could mean in terms of cognition of learning and memory, we first try and interpret the possible cognitive meanings of the orthonormal basis set and its constituent vectors.

As discussed in 4.8.4, the information entering the model brain of size N is denoted by a $N-$dimensional vector. The components of the input vectors are ± 1 generated randomly and represent a set of features. $+1$ and -1 represent 'Yes' and 'No' respectively, with respect to the corresponding feature. Together, the components characterize the information in terms of its features[15, 16, 17]. On encountering such information, the model brain first orthogonalizes them[15], and then memorizes the orthogonalized information in a Hebbian manner[2]. The components of these orthogonalized vectors are fractions, and may be understood as the firing rates of the neurons in response to the incoming information which is then learnt by the system. In our hypothesis, each of the orthonormal vectors would be representative of an information stored in the system. The vectors in the basis are all mutually orthogonal, making each of the representative vectors unique. The positive or negative values in the vector components show whether the corresponding features in the information get highlighted/emphasized or de-emphasized to the degree given by the magnitude. The signs of the components of the different representative vectors adapt themselves in order to maintain the mutual orthogonality of the vectors.

We can now attempt to conjecture what the Symmetric and Canonical orthogonalization schemes could mean in terms of cognition. Symmetric orthogonalization could be applied in situations where we take in information about our surroundings, noting everything in general but not focusing on anything in particular, while Canonical orthogonalization would be used for instant classification of incoming information and for sorting them into groups in a hierarchical manner.

With these considerations in mind, we now turn our attention to the possible studies that might shed more light on how the brain might implement orthogonalization. We will now comment on a few prospective research avenues.

The modular nature of the brain has been a subject of debate for more than a hundred years– is the brain made up of different task-specific modules, or does the brain use a common mechanism to address a variety of tasks?[79] A study by Tsao et al [80] demonstrating a face-specific region in the cortex provides evidence in favour of the modular brain theory. Faces possess the same characteristic sets of features, but a large number of individual faces can still be identified distinctly. This prompts the following question: how are the faces encoded by the cortical neurons such that each face can be distinctly separated from the rest?[79]. We believe that the brain might use Löwdin orthogonalization schemes to memorize, sort and recognize faces.

We plan to study the issues mentioned above through simulations, but here, we discuss whether Löwdin orthogonalization schemes help with content addressability. To do so, we now study associativity in the network following Symmetric and Canonical orthogonalization schemes, as we have already elaborated on the effects of Gram-Schmidt orthogonalization on associativity in the previous chapter.

This result bears some resemblance to episodic memory, where an episode is remembered along with its context. Encountering a similar environment could lead to remembering the memory of the episode and its environment. The content-addressable or associative nature of our model system is comparable to episodic memory, as we will show in our simulations. But first, we briefly recapitulate the reason for introducing orthogonalization in the Hopfield model, and the significant difference in the associative property of the modified model compared to that of the conventional Hopfield model.

Prior to elaborating on the Löwdin orthogonalization and their relevance to cognition, we remark on a curious (yet telling) coincidence – the factors that hinted at the parallels between the memory catastrophe problem in learning and memory and the 'non-orthogonality catastrophe'[81] in chemistry which prompted Löwdin to develop his orthogonalization schemes[4] inspired us to apply orthogonalization to the problem in cognition. As we have seen earlier, the memory catastrophe in the brain is the result of growing correlations between the patterns as more information is memorized, similarly, small overlaps between the orbitals of ions in close proximity to each other contibute to the rise in the number of overlap integrals, eventually leading to the non-orthogonality catastrophe[82].

5.2 A précis of the orthogonalization schemes

Consider an $N-$ dimensional space where a set $\boldsymbol{V}$ whose component vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \ldots, \boldsymbol{v}_N$ are linearly independent. Let $\boldsymbol{A}$ represent the transformation taking $\boldsymbol{V}$ to an orthonormal basis $\boldsymbol{\mathcal{Z}}$. Then,

$$\boldsymbol{\mathcal{Z}} = \boldsymbol{VA}, \tag{5.1}$$

with

$$< \boldsymbol{\mathcal{Z}} | \boldsymbol{\mathcal{Z}} >= \boldsymbol{I}, \tag{5.2}$$

where $\boldsymbol{I}$ is the identity matrix. We can see from [3] that

$$\begin{aligned} < \boldsymbol{\mathcal{Z}} | \boldsymbol{\mathcal{Z}} > &=< \boldsymbol{VA} | \boldsymbol{VA} >, \\ &= \boldsymbol{A}^{\dagger} < \boldsymbol{\mathcal{Z}} | \boldsymbol{\mathcal{Z}} > \boldsymbol{A}, \\ &= \boldsymbol{A}^{\dagger} \boldsymbol{A} \boldsymbol{A}; \end{aligned}$$

yields the general solution to the orthogonalization problem on substituting for $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{M}^{-\frac{1}{2}} \boldsymbol{B}, \tag{5.3}$$

with $\boldsymbol{M}$ being the Hermitian metric matrix of the basis $\boldsymbol{V}$, and $\boldsymbol{B}$ being the unitary matrix. Two particular solutions of the general solution(5.3) generate the Symmetric and Canonical bases. $\boldsymbol{B} = \boldsymbol{I}$ in the case of Symmetric Orthogonalization $\boldsymbol{\mathcal{Z}} = \boldsymbol{\Phi} = \boldsymbol{V} \boldsymbol{M}^{-\frac{1}{2}}$; while $\boldsymbol{B} = \boldsymbol{U}$ where $\boldsymbol{U}$ diagonalizes $\boldsymbol{M}$ and

$$\boldsymbol{U}^{\dagger} \boldsymbol{M} \boldsymbol{U} = \boldsymbol{d}, \tag{5.4}$$

yields the Canonical Orthogonalization $\boldsymbol{\mathcal{Z}} = \boldsymbol{\Lambda} = \boldsymbol{V} \boldsymbol{U} \boldsymbol{d}^{-\frac{1}{2}}$.

Some interesting characteristics of the two schemes[3] can be gleaned from the Scheinler-Wigner matrix[21]:

$$\begin{pmatrix} |(\boldsymbol{v}_1, \boldsymbol{z}_1)|^2 & |(\boldsymbol{v}_1, \boldsymbol{z}_2)|^2 & \dots & |(\boldsymbol{v}_1, \boldsymbol{z}_N)|^2 \\ |(\boldsymbol{v}_2, \boldsymbol{z}_1)|^2 & |(\boldsymbol{v}_2, \boldsymbol{z}_2)|^2 & \dots & |(\boldsymbol{v}_2, \boldsymbol{z}_N)|^2 \\ \vdots & \vdots & \ddots & \vdots \\ |(\boldsymbol{v}_N, \boldsymbol{z}_1)|^2 & |(\boldsymbol{v}_N, \boldsymbol{z}_2)|^2 & \dots & |(\boldsymbol{v}_N, \boldsymbol{z}_N)|^2 \end{pmatrix}.$$

The elements of the matrix are the squares of the projections of the vectors $\boldsymbol{v}_k$'s on the vectors $\boldsymbol{z}_\kappa$'s of the orthonormal basis. While the row sums give the squares of the lengths of the given vectors, the column sums contain fascinating information. A column sum is given by c_κ[3]

$$\sum_k | \, (\boldsymbol{v}_k, \boldsymbol{z}_\kappa)|^2 = (\boldsymbol{AMMA}^{\dagger})_{\kappa\kappa} = (\boldsymbol{BMB}^{\dagger})_{\kappa\kappa} = c_\kappa; k = 1, \dots, N. \tag{5.5}$$

We must point out here that,

$$\sum_{\kappa=1}^{N} c_\kappa = \sum_{k=1}^{N} |\boldsymbol{v}_k|^2, \text{ takes a constant value for a given set } \boldsymbol{V}, \tag{5.6a}$$

while

$$\sum_{\kappa} c_\kappa^2 = m \text{ , the SW parameter. [3, 21]} \tag{5.6b}$$

It has been shown that in the case of Symmetric orthogonalization,

$$m = m_{min} \text{ for the orthonormal basis } \boldsymbol{\mathcal{Z}} = \boldsymbol{\Phi}, \tag{5.7a}$$

while

$$m = m_{max} \text{ for the orthonormal basis } \boldsymbol{\mathcal{Z}} = \boldsymbol{\Lambda}. \tag{5.7b}$$

in the case of Canonical orthogonalization, when the vectors $\boldsymbol{v}_k$ are normalized[3].

In other words, the orthonormal basis vectors have an average distribution in the case of Symmetric orthogonalization, that is, $c_1 = c_2 = \ldots = c_N = (c_1 + c_2 + \ldots + c_N)/N$, whereas in Canonical orthogonalization, the c_κ's are distributed in a maximally skewed manner. To paraphrase, for Symmetric orthogonalization, $\boldsymbol{\mathcal{Z}} = \boldsymbol{\Phi} \equiv \{\phi_\kappa\}$, and

$$\sum_{k} |(\boldsymbol{v}_k, \boldsymbol{\phi}_\kappa)|^2 = \sum_{\kappa} |(\boldsymbol{v}_k, \boldsymbol{\phi}_\kappa)|^2 = |\boldsymbol{v}_k|^2, \tag{5.8a}$$

indicating that the projection squares of all the normalized vectors $\boldsymbol{v}_k$'s on the individual $\boldsymbol{\phi}_\kappa$'s add up to the same value. Whereas, in the case of Canonical orthogonalization, $\boldsymbol{\mathcal{Z}} = \boldsymbol{\Lambda} \equiv \{\boldsymbol{\lambda}_\kappa\}$, and

$$\sum_{k} |(\boldsymbol{v}_k, \boldsymbol{\lambda}_l)|^2 = \text{the maximum for, say, } \kappa = l; \tag{5.8b}$$

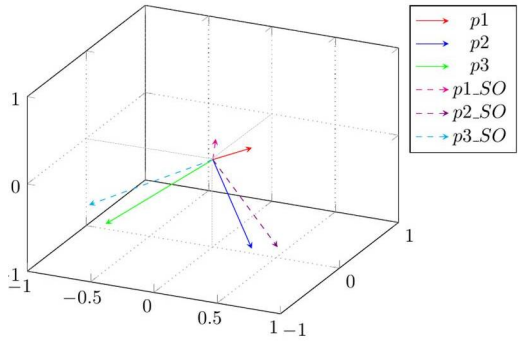
$$\sum_{k} |(\boldsymbol{v}_k, \boldsymbol{\lambda}_m)|^2 = \text{the next to maximum for, say } \kappa = m,$$

and so on. The maximum value of the collective projections of all the $\boldsymbol{v}_k$'s is encapsulated in $\boldsymbol{\lambda}_l$, indicating a hierarchy among the sums of projection squares of $\boldsymbol{v}_k$'s on

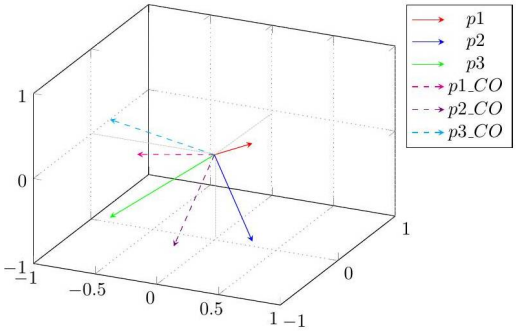
the $\boldsymbol{\lambda}_\kappa$'s.

In the case of Symmetric orthogonalization, the orientation of the orthogonal basis set is such that the sums of squares of the projections of the input vectors $\boldsymbol{v}_k$'s on each of the orthogonal vectors $\boldsymbol{\phi}_\kappa$'s takes the same value for each individual $\boldsymbol{\phi}_\kappa$. This is a significant trait of the orthogonal basis which retains any symmetry properties possessed by the input set. However, the Canonically orthogonalized basis set vectors $\boldsymbol{\lambda}_\kappa$'s are oriented in a descending order – the highest amount of information about the shared features present in the input set is captured in one of the orthogonal vectors $\boldsymbol{\lambda}_l$, followed by other orthogonal vectors $\boldsymbol{\lambda}_m$ and $\boldsymbol{\lambda}_n$ which capture the information but to successively lesser extents. The component vectors of the orthogonal basis can thus typically be sorted by the amount of information captured, thereby making the Canonical orthogonalization a powerful classification procedure.

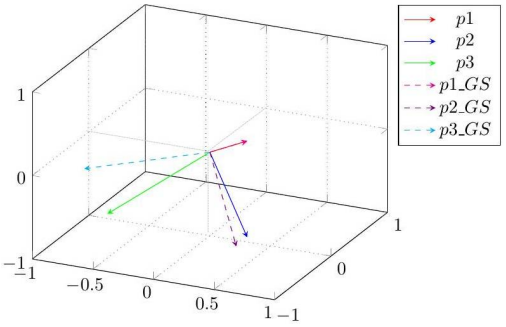
The properties of the orthogonalization discussed above are illustrated using a simple example in Fig. 5.1. The Gram-Schmidt procedure is also illustrated for comparison and correctness. The data pertaining to the figure is given in Table5.1.



(a) Symmetric orthogonalization



(b) Canonical orthogonalization



(c) Gram-Schmidt orthogonalization

FIGURE 5.1: 3D Plots depicting various orthogonalization schemes. The 3-dimensional input vectors $p1, p2$ and $p3$ are orthogonalized using Symmetric 5.1a, Canonical 5.1b and Gram-Schmidt 5.1c schemes, with the corresponding orthonormalized vectors plotted using dashed lines. The vectors are tabulated in Table5.1.

TABLE 5.1: Table showing an example of the various orthogonalization schemes usning 3D vectors along with their corresponding SW matrices. The row sums and column sums of the SW matrices are also shown. The vectors are plotted in Fig. 5.1.

Normalized input patterns			
$p1$:	0.57735	−0.57735	0.57735
$p2$:	0.57735	−0.57735	−0.57735
$p3$:	−0.57735	−0.57735	−0.57735
Symmetric orthogonalization			
$p1\_SO$:	0.33333	−0.66667	0.66667
$p2\_SO$:	0.66667	−0.33333	−0.66667
$p3\_SO$:	−0.66667	−0.66667	−0.33333

SW Matrix

$$\begin{pmatrix} 0.92593 & 0.03704 & 0.03704 \\ 0.03704 & 0.92593 & 0.03704 \\ 0.03704 & 0.03704 & 0.92593 \end{pmatrix} \begin{matrix} 1.00001 \\ 1.00001 \\ 1.00001 \end{matrix}$$
$$\begin{matrix} 1.00001 & 1.00001 & 1.00001 \end{matrix}$$

Canonical orthogonalization			
$p1\_CO$:	−0.8165	0.40825	−0.40825
$p2\_CO$:	0	−0.70711	−0.70711
$p3\_CO$:	−0.57735	−0.57735	0.57735

SW matrix

$$\begin{pmatrix} 0.88889 & 0.22222 & 0.22222 \\ 0 & 0.66667 & 0.66667 \\ 0.11111 & 0.11111 & 0.11111 \end{pmatrix} \begin{matrix} 1.33333 \\ 1.33334 \\ 0.33333 \end{matrix}$$
$$\begin{matrix} 1 & 1 & 1 \end{matrix}$$

Gram-Schmidt orthogonalization			
$p1\_GS$:	0.57735	−0.57735	0.57735
$p2\_GS$:	0.40825	−0.40825	−0.81650
$p3\_GS$:	−0.70711	−0.70711	0

SW matrix

$$\begin{pmatrix} 1 & 0.11111 & 0.11111 \\ 0 & 0.88889 & 0.22222 \\ 0 & 0 & 0.66667 \end{pmatrix} \begin{matrix} 1.22222 \\ 1.11111 \\ 0.66667 \end{matrix}$$
$$\begin{matrix} 1 & 1 & 1 \end{matrix}$$

5.3 Numerical Illustration

Prior to discussing what these two Löwdin orthogonalization could mean in terms of cognition, we first implement Symmetric and Canonical orthogonalizations in a small system to demonstrate numerically the features of the two schemes discussed in the previous section.

Our input set of 10-dimensional vectors $\{\boldsymbol{v}_k\}$ comprises randomly generated ± 1 elements, which are each divided by a factor of $\sqrt{10}$ to normalize the vectors. Initially, two such random vectors are generated and orthogonalized using the two Löwdin schemes. More vectors are then added to the input set, with the entire set being orthogonalized following the addition of each new vector. Table5.2 gives the data pertaining to Symmetric and Canonical orthogonalization schemes.

TABLE 5.2: Five randomly generated 10-dimensional vectors are normalized and orthogonalized following Symmetric (Table I) and Canonical (Table II) schemes. First, two vectors are orthogonalized, then new vectors are added, one at a time. The SW matrices are presented sequentially to examine how they change as new vectors are added. For the convenience of presentation, these numbers are presented as transpose of the SW matrix in the text.

Unnormalized input vectors

$\boldsymbol{v}_1$:	1	-1	1	1	1	-1	1	1	-1	-1
$\boldsymbol{v}_2$:	1	-1	1	-1	-1	-1	-1	1	-1	-1
$\boldsymbol{v}_3$:	-1	1	-1	-1	1	-1	-1	1	1	-1
$\boldsymbol{v}_4$:	1	-1	1	-1	-1	-1	1	-1	-1	-1
$\boldsymbol{v}_5$:	-1	-1	1	1	1	1	1	1	1	-1

Table - I : Symmetric Orthogonalization

of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ (normalized) : $(p = 2)$

0.2673	−0.2673	0.2673	0.4082	0.4082	−0.2673	0.4082	0.2673	−0.2673	−0.2673
0.2673	−0.2673	0.2673	−0.4082	−0.4082	−0.2673	−0.4082	0.2673	−0.2673	−0.2673

of $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ (normalized) : $(p = 3)$

0.2351	−0.2351	0.2351	0.3804	0.4531	−0.3078	0.3804	0.3078	−0.2351	−0.3078
0.2770	−0.2770	0.2770	−0.4000	−0.4223	−0.2547	−0.4000	0.2547	−0.2770	−0.2547
−0.2966	0.2966	−0.2966	−0.2743	0.3693	−0.3470	−0.2743	0.3470	0.2966	−0.3470

of $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$ (normalized) : $(p = 4)$

0.2259	−0.2259	0.2259	0.4207	0.4715	−0.2768	0.3392	0.3582	−0.2259	−0.2768
0.2578	−0.2578	0.2578	−0.2639	−0.3711	−0.1506	−0.5441	0.4307	−0.2578	−0.1506
−0.2835	0.2835	−0.2835	−0.3670	0.3344	−0.4178	−0.1763	0.2272	0.2835	−0.4178
0.1499	−0.1499	0.1499	−0.4220	−0.2313	−0.3405	0.4301	−0.5115	−0.1499	−0.3405

of $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5$ (normalized) : $(p = 5)$

0.3470	−0.1135	0.1135	0.4084	0.4617	−0.4002	0.2779	0.2972	−0.3470	−0.1667
0.2019	−0.3096	0.3096	−0.2581	−0.3664	−0.0936	−0.5157	0.4590	−0.2019	−0.2013
−0.2813	0.2856	−0.2856	−0.3669	0.3345	−0.4202	−0.1773	0.2262	0.2813	−0.4158
0.0959	−0.2001	0.2001	−0.4161	−0.2265	−0.2855	0.4577	−0.4841	−0.0959	−0.3897
−0.3797	−0.3624	0.3624	0.1505	0.1462	0.3840	0.2547	0.2539	0.3797	−0.3580

Schweinler-Wigner matrices for $p = 2, 3, 4, 5$

$$p = 2 : \begin{pmatrix} 0.9583 & 0.0417 \\ 0.0417 & 0.9583 \end{pmatrix} \begin{matrix} : 1.0000 \\ : 1.0000 \end{matrix}$$

$$p = 3 : \begin{pmatrix} 0.9472 & 0.0422 & 0.0106 \\ 0.0422 & 0.9577 & 0.0001 \\ 0.0106 & 0.0001 & 0.9893 \end{pmatrix} \begin{matrix} : 1.0000 \\ : 1.0000 \\ : 1.0000 \end{matrix}$$

$$p = 4 : \begin{pmatrix} 0.9283 & 0.0341 & 0.0078 & 0.0298 \\ 0.0341 & 0.8655 & 0.0019 & 0.0985 \\ 0.0078 & 0.0019 & 0.9454 & 0.0448 \\ 0.0298 & 0.0985 & 0.0448 & 0.8269 \end{pmatrix} \begin{matrix} : 1.0000 \\ : 1.0000 \\ : 0.9999 \\ : 1.0000 \end{matrix}$$

$$p = 5 : \begin{pmatrix} 0.8602 & 0.0406 & 0.0079 & 0.0358 & 0.0555 \\ 0.0406 & 0.8510 & 0.0019 & 0.0937 & 0.0128 \\ 0.0079 & 0.0019 & 0.9454 & 0.0447 & 0.0000 \\ 0.0358 & 0.0937 & 0.0447 & 0.8131 & 0.0127 \\ 0.0555 & 0.0128 & 0.0000 & 0.0127 & 0.9190 \end{pmatrix} \begin{matrix} : 1.0000 \\ : 1.0000 \\ : 0.9999 \\ : 1.0000 \\ : 1.0000 \end{matrix}$$

Table - II : Canonical Orthogonalization

of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ (normalized) : $(p = 2)$

0.3780	−0.3780	0.3780	0	0	−0.3780	0	0.3780	−0.3780	−0.3780
0	0	0	−0.5774	−0.5774	0	−0.5774	0	0	0

of $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ (normalized) : $(p = 3)$

−0.4353	0.4353	−0.4353	−0.1027	0.0635	0.2690	−0.1027	−0.2690	0.4353	0.2690
−0.1414	0.1414	−0.1414	−0.4243	0.1414	−0.4243	−0.4243	0.4243	0.1414	−0.4243
−0.1027	0.1027	−0.1027	0.4353	0.7042	−0.1663	0.4353	0.1663	0.1027	−0.1663

of $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$ (normalized) : $(p = 4)$

0.4311	−0.4311	0.4311	−0.0717	−0.2142	−0.2886	0.1969	0.0200	−0.4311	−0.2886
−0.1002	0.1002	−0.1002	−0.3716	0.1691	−0.4405	−0.4143	0.4832	0.1002	−0.4405
0.0402	−0.0402	0.0402	0.5954	0.6310	−0.0758	0.2902	0.3810	−0.0402	−0.0758
−0.1512	0.1512	−0.1512	−0.2482	0.2297	−0.3268	0.5766	−0.4981	0.1512	−0.3268

of $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4, \boldsymbol{v}_5$ (normalized) : $(p = 5)$

0.0843	−0.2014	0.2014	0.1859	−0.2886	0.3903	−0.5709	0.4837	−0.0843	0.2731
0.3869	0.2584	−0.2584	0.4342	0.3750	−0.3277	−0.0485	0.1650	−0.3869	0.3177
−0.1153	0.0764	−0.0764	−0.3088	0.2319	−0.4253	−0.3718	0.5272	0.1153	−0.4643
0.1778	0.2645	−0.2645	−0.4882	−0.4508	−0.2153	−0.3843	−0.3311	−0.1778	0.2271
−0.4380	0.4225	−0.4225	0.0838	0.2259	0.2960	−0.1865	−0.0101	0.4380	0.2805

Schweinler-Wigner matrices for $p = 2, 3, 4, 5$

$$p = 2 : \begin{pmatrix} 0.7000 & 0.7000 \\ 0.3000 & 0.3000 \end{pmatrix} \begin{matrix} \\ : 1.4000 \\ : 0.6000 \end{matrix}$$

$$p = 3 : \begin{pmatrix} 0.7236 & 0.5789 & 0.1447 \\ 0.0000 & 0.2000 & 0.8000 \\ 0.2764 & 0.2211 & 0.0553 \end{pmatrix} \begin{matrix} \\ : 1.4472 \\ : 1.0000 \\ : 0.5528 \end{matrix}$$

$$p = 4 : \begin{pmatrix} 0.4984 & 0.5810 & 0.2150 & 0.7640 \\ 0.0120 & 0.2497 & 0.7398 & 0.0046 \\ 0.4884 & 0.0678 & 0.0014 & 0.1010 \\ 0.0012 & 0.1015 & 0.0438 & 0.1304 \end{pmatrix} \begin{matrix} \\ : 2.0585 \\ : 1.0061 \\ : 0.6585 \\ : 0.2769 \end{matrix}$$

$$p = 5 : \begin{pmatrix} 0.0079 & 0.1135 & 0.0428 & 0.1090 & 0.0026 \\ 0.1423 & 0.0108 & 0.0009 & 0.0571 & 0.1021 \\ 0.0342 & 0.2196 & 0.7359 & 0.0100 & 0.0038 \\ 0.3384 & 0.0651 & 0.0064 & 0.0491 & 0.8889 \\ 0.4772 & 0.5910 & 0.2141 & 0.7749 & 0.0025 \end{pmatrix} \begin{matrix} \\ : 0.2759 \\ : 0.3131 \\ : 1.0035 \\ : 1.3479 \\ : 2.0596 \end{matrix}$$

Examining the data for a pair of vectors, that is, for $p = 2$ highlights how the two schemes compare the vectors of the input set during the orthogonalization process. The elements of the Symmetric basis vector each take one of two values, one representing similarities and the other, differences. The magnitude of the elements depends on the number of similarities or differences - higher the count, greater the magnitude. That is, if the pair of vectors share more similarities than differences, the number denoting the similarities in the orthogonal basis vector takes a higher value, while the differences are represented by a smaller number. The opposite occurs when the two vectors are more different from each other.

The similarities and dissimilarities between the two patterns are reflected sharply in the case of Canonical orthogonalization – the similarities or differences, whichever is smaller in number, is represented by a number of greater magnitude, and in the second vector. This comparison is highlighted further in the first vector, with the corresponding elements taking value zero. Also, the differences, or similarities, which are more in number are reflected in the first vector, but with smaller magnitude, with the corresponding elements in the second vector being zeroes. In short, the first vector emphasizes the majority feature, while the second vector emphasizes the minority. That is, the model brain first pays attention to the majority feature, reflecting it in the first vector, and then in the second vector, it stresses solely the minority feature.

Following the introduction of a third vector and orthogonalization, the bases get altered completely instead of a cumulative change as in the case of Gram-Schmidt orthogonalization. However, as in the case of $p = 2$, the patterns of similarities and differences present in the input set can still be seen in the orthogonal bases of either scheme.

Beyond $p = 3$, as the number of input vector goes up, it gets harder to attribute the values in the orthogonal bases to the patterns of similarities and differences present in the input set. However, the parameters c_κ's provide a general picture of the reorganization of the orthonormal bases, showing how the bases re-orient themselves following the introduction of newer vectors.

The eq.(5.8a) is satisfied perfectly in the case of Symmetric orthogonalization, indicating the perfect symmetry in the SW matrices. Another interesting aspect of the SW matrices is that they are diagonally dominated. This implies a close alignment of the basis set vectors with the input vectors – $\boldsymbol{\phi}_1$ with $\boldsymbol{v}_1$, $\boldsymbol{\phi}_1$ with $\boldsymbol{v}_1, \ldots$, though

there is a marginal reduction in the degree of alignment as more input vectors are added.

In the case of Canonical orthogonalization, one of the basis vectors $\boldsymbol{\lambda}_k$ records maximally the projections of the input vectors$\boldsymbol{v}_k$'s. In our example, it is $\boldsymbol{\lambda}_k = \boldsymbol{\lambda}_1$ which captures the majority of the projections of $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_4$ for $p = 2, 3, 4$. However, following the introduction of the fifth vector $\boldsymbol{v}_5$, $\boldsymbol{\lambda}_k$ changes to $\boldsymbol{\lambda}_5$ which is now the vector sampling the input set to the greatest extent. The hierarchical ordering of the basis vectors can be seen from the sums of projection squares in Table II.

5.4 Löwdin orthogonalization schemes - implementation and cognitive relevance

In Chapter 2, we saw how the Hebb-Hopfield network suffers from a catastrophic blackout when the number of patterns p in a network of size N crosses the small limit of $p = 0.14N$. We also saw that the reason for this is the growing amount of correlations between the patterns manifesting as a noise which eventually submerges the signals due to the learnt information. One way of evading this catastrophe is the elimination of noise by using orthogonalization. In the Chapters 3 and 4, we discussed in detail the network properties of the H-H-GS model in which Gram-Schmidt orthogonalization is invoked in the H-H model. While information in the H-H-GS network is stored as the orthogonalized versions $\boldsymbol{\eta}$'s of the input patterns $\boldsymbol{\xi}$'s, the original patterns or the raw input $\boldsymbol{\xi}$'s can still be recovered with total accuracy. Not just that, the network is also capable of associatively recalling the input memories when presented with patterns similar to them.

Encouraged by these results, we now study the H-H network with Löwdin orthogonalizations and what the results could mean in cognitive terms. That is, we orthogonalize the input vectors using Symmetric and Canonical orthogonalizations to get the basis vectors which are then memorized by the network. We then examine if the network can still recognize and recall the input patterns.

Tables I and II tabulate the results of Symmetric and Canonical orthogonalizations on a set of 5 10-dimensional vectors whose components are generated randomly. The corresponding SW matrices are also shown. The input vectors $\boldsymbol{v}_k$'s represent the information that need to be lodged in the memory, and correspond to the vectors

$\boldsymbol{\xi}^{\nu}$'s of the earlier chapters. $\boldsymbol{\phi}_{\kappa}$'s and $\boldsymbol{\lambda}_{\kappa}$'s represent the orthogonal vectors after Symmetric and Canonical orthogonalizations respectively. The model network might focus equally and impartially on all the incoming patterns, or it might rank them, focusing more on some patterns and less on others. That is, the network can perform either Symmetric or Canonical orthogonalization based on the situation. This choice is in turn reflected in the basis vectors and the corresponding SW matrix. These vectors are then used to calculate the synaptic weights.

The issues of recognition and recall arise once again, as in the case of Gram-Schmidt orthogonalization, leading to the question: are the input patterns recognized by the network, and are they still content-addressable when Löwdin orthogonalization schemes are used in the memorization process? The answer to both these questions is yes. Not just the input patterns themselves, but even something similar can lead the network to retrieve the original input patterns, proving that the network is indeed still capable of recognition as well as associative recall.

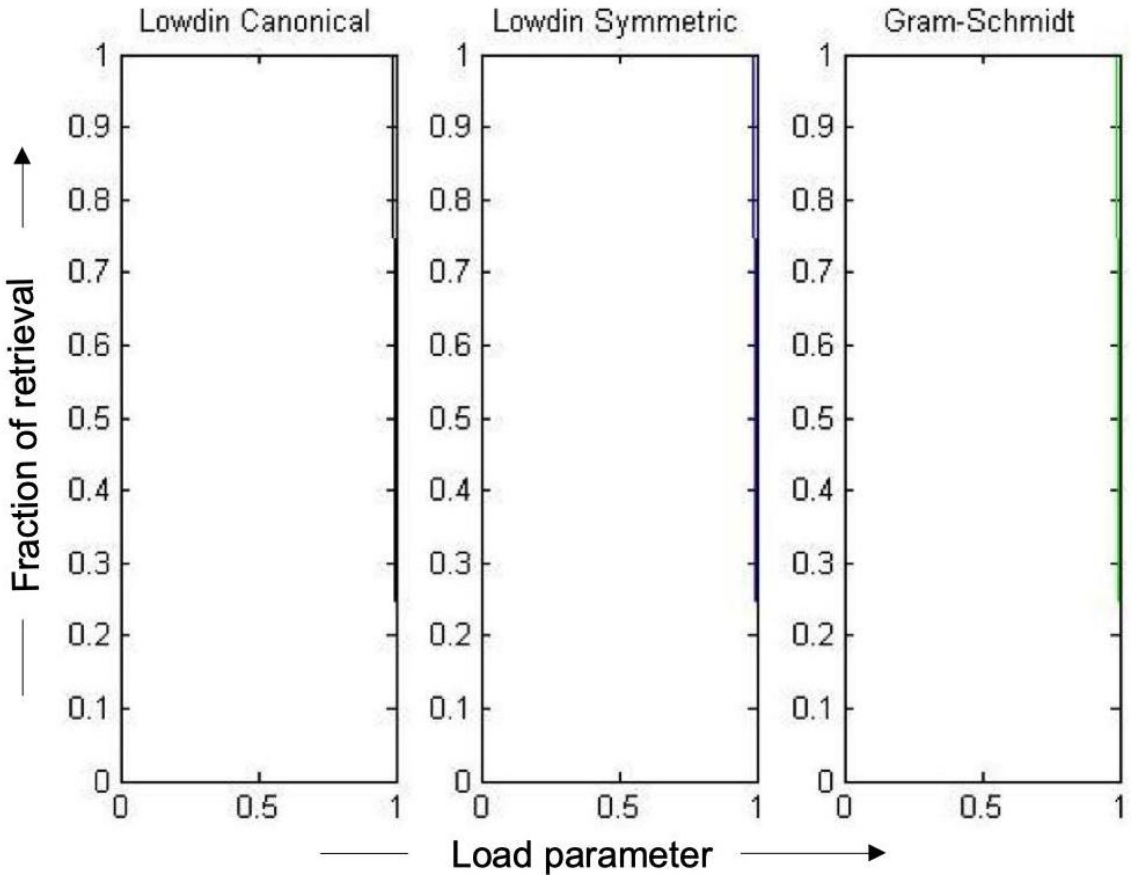


FIGURE 5.2: Plot showing the fraction of stable patterns for different values of the load parameter following Symmetric, Canonical and Gram-Schmidt orthogonalization schemes in a network of size $N = 100$. The data pertains to 3 trials.

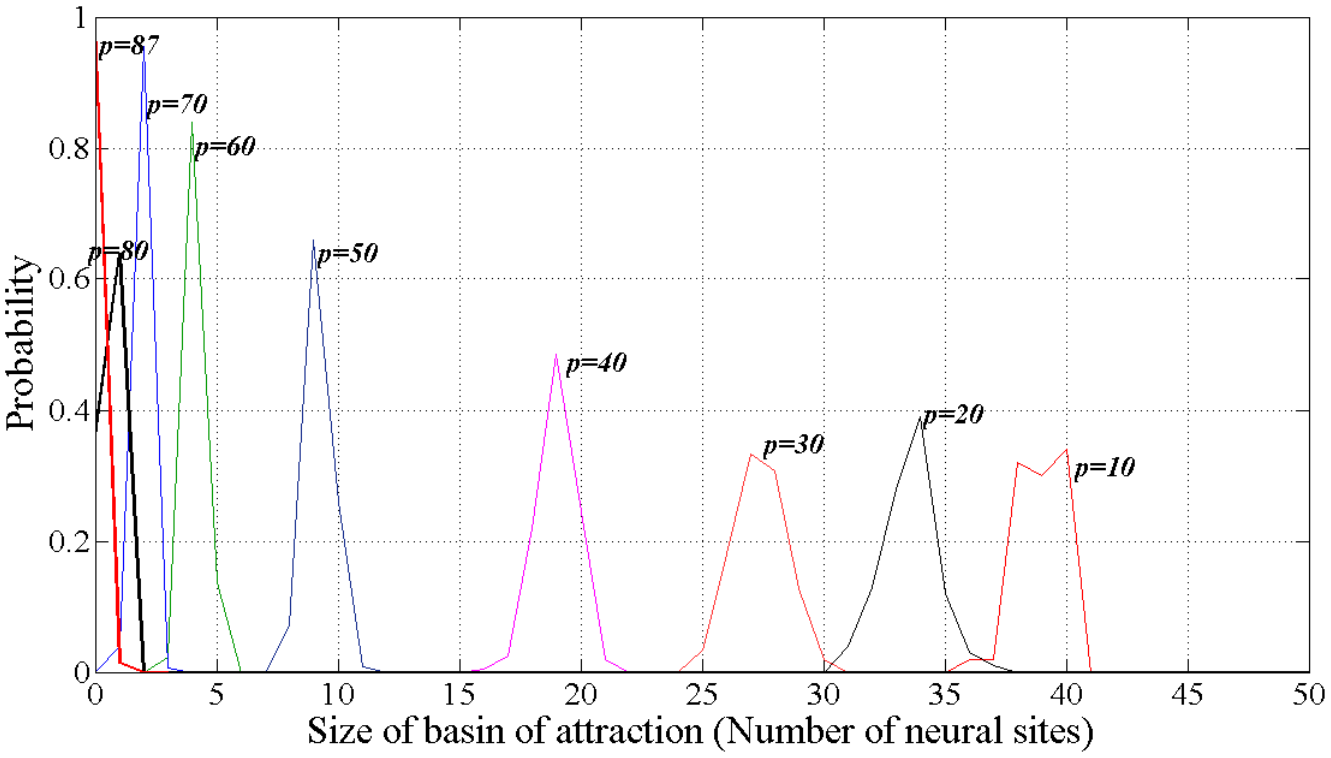


FIGURE 5.3: Probability of finding basin of attraction of a certain size for $p = 10 - 87$ (in steps of 10) following Symmetric or Canonical orthogonalization in a network of size $N = 100$. (Refer to AppendixD for details on the equivalence of the orthogonalization schemes.) The data pertains to 5 trials, and the basins are averaged over 10 samples.

The memory capacity of the network for recognition is $p = N-1$, as shown in Fig.5.2. We examine in some detail the results pertaining to recall which is relevant to episodic memory. We find that the network can associate with the input patterns other patterns which are fairly similar to them. It is worth noting that as the memory load increases, the recall process would become more attuned to the commonalities shared by the presented patterns and the memorized ones, that is, the proximity between the presented and the input patterns. This is evident in the basins of attraction (calculated following the protocol described in AppendixB) plotted in Fig.5.3. The radii of the basins of attraction reduce with increase in p. As memory load goes up, there is also a rise in the number of patterns with zero basins of attraction.

5.5 In short

We have seen that Löwdin orthogonalization schemes can be applied to address problems in computational neuroscience. If validated through experiments, our claims here could provide novel insights into Löwdin orthogonalizations as the physiological mechanism involved in episodic memory wherein an episode gets stored in the brain along with its context.

In this chapter, we have discussed the probable cognitive meanings of the orthonormal bases and their components which are greatly altered each time a new information is added to the memory. Also, the Schweinler-Wigner matrix helps in interpreting in gross terms the processes of orthogonalizations.

We have discussed some preliminary results of our study in this chapter. Further studies using Löwdin orthogonalization schemes could provide insights into complex phenomena observed in the domains of cognitive neuroscience and psycholinguistics.

We also believe that Löwdin orthogonalization schemes might be implicated in dyslexia. It might help us understand the distinct brain state of dyslexics who, despite their learning difficulties, are proficient in identifying patterns in seemingly random collections of data (such as a group of people, objects or numbers) which other non-dyslexic people may not be able to identify as easily.

Chapter 6

Summary and outlook

This thesis focused on the implementation of various orthogonalization schemes in the Hopfield model and also attempted to situate the study in the context of cognition, to understand how the results of the study relate to biological learning and memory. The sequential Gram-Schmidt orthogonalization procedure and the democratic Symmetric and Canonical schemes were invoked in the Hopfield model and the network performance was evaluated by estimating the memory capacity.

The thesis also addressed deeper issues related to memory stability and the associative character of the network. The terms retrieval, recognition and recall were defined precisely and distinguished from each other after stating the rationale behind the distinction and the need for it. A detailed study of the behaviour of the network following orthogonalization was presented and compared to that of the standard Hebb-Hopfield model.

Future work would involve a more in-depth study of Löwdin orthogonalization schemes and their relevance to understanding dyslexia and to pattern identification. Other themes to pursue include formulating a representation for forgetting within our framework, and ways to combine a short-term memory model with our long-term memory models.

Another avenue to be explored is a model where the network can implement different orthogonalization schemes according to the situation and capture different aspects of the incoming information.

Further work also includes addressing some of the limitations of the Hopfield model, beyond the low memory capacity that we have addressed here, such as the (possibly)

arbitrarily large values the synaptic efficacies can take. This can be resolved by using orthogonalization as discussed in the previous chapters. An alternate solution is to restrict the range of values synaptic efficacies can take, a process which also takes the model closer to biological realism. This situation, where the weights are regulated to always lie within certain limits can be referred to as *learning within bounds* [19, 83, 84, 85]. We take our cue from a series of studies[22, 86, 87] and consider a network of bounded synapses (following [22] , referred to as "original" model hereafter) and study its efficiency in terms of memory lifetime, or how long memorized information remains in the system and is distinguishable from the noise. We wished to incorporate synaptic nature in the network - biologically, synapses are either excitatory or inhibitory in nature and retain this nature in the face of modifications during the learning process. We elaborate upon this in AppendixF, but preliminary results show that our modified network shows a slight increase in the initial signal-to-noise ratio, while memory lifetimes are roughly half compared to the original model.

Another criticism of the Hebb-Hopfield network is its use of dense patterns to represent information, with roughly half of the total neurons active in each pattern, whereas only a fraction of neurons in the brain fire or are active in response to any information entering the brain. Such sparse patterns of activity can be implemented computationally through a model proposed by David Willshaw et al[23, 88].

The Willshaw model consists of a network of neurons connected by synapses which are all initially inactive and get activated only if both the neurons connected by the synapse fire simultaneously. The model is biologically realistic in terms of the sparseness in the information. However, unlike biological systems, it is fully connected and the synaptic efficacies are symmetric. We attempt to address this issue by modifying the learning prescription such that a synapse can take one of three values (positive/active, 0 and negative/inactive), rather than two (active/inactive). With this simple modification, we can introduce dilution and asymmetry in the network dynamics, as well as vary inhibition and/or excitation and study the effects of varying amounts of dilution, as discussed in AppendixG.

The network performance in this case is measured in terms of the number of patterns the network can memorize and retrieve accurately. Initial results show favourable results when inhibition is maintained at a constant level and the excitation varied. This result could be of relevance to biological systems (see [89] related to Alzheimer's disease, for instance), as well neural networks[90]. However, the results were less

favourable when both inhibition and excitation are varied. Further work remains to be done in improving the model.

In sum, in this thesis, we have studied the Hebb-Hopfield network following the implementation of various orthogonalization schemes. We have analyzed network efficacy using memory capacity, pattern stability and associativity. We have also discussed some possible avenues of further research.

Appendix A

Methods

A.1 Method - Chapter 2

The network was built on GNU Octave/MATLAB and FORTRAN 90. The size of the network was fixed at $N = 100$. The patterns were randomly-generated N-dimensional vectors whose components were ± 1. The patterns were stored using the learning rule of eq.(2.3) and presented to the network for retrieval. The fractions of stable and unstable patterns were then calculated. The percentage of stable and unstable patterns were then plotted against the number of patterns. The data was obtained from 50 trials (sets of patterns) and plotted using GNUPlot.

A.2 Method - Chapter 3

The network was built on GNU Octave/MATLAB and FORTRAN 90. The number of neurons in the network were $N = 100$ and $N = 1000$. The patterns were randomly-generated N-dimensional vectors whose components were ± 1. The patterns were then orthogonalized and normalized following the Gram-Schmidt procedure described in sec. 3.3.1 and stored using the learning rule of eq. (3.12). They were then presented to the network for retrieval. The fractions of patterns were then calculated. The percentage of stable patterns were then plotted against the number of patterns. The fraction of retrieval was then plotted for different values of the load parameter. The data was obtained from 50 trials (sets of patterns) and plotted using GNUPlot.

A.3 Method - Chapter 4

The network was built on GNU Octave/MATLAB and plots generated using GNU-Plot. The size of the network was fixed at $N = 100$. The patterns were randomly-generated N-dimensional vectors with components having different signs but the same magnitude, namely 1. The patterns were stored using the learning rule of eq.(2.3) and tested for retrieval following eq.(2.9). The quality of convergence was also checked for different values of N ($N = 100, 500, 1000$). The data was obtained always from 50 sets of patterns (or trials), collated and then arranged into bins. That is, the fraction of retrieval (number of retrieved patterns/total number of patterns) for each bin (here, representing the load parameter (p/N)) was counted for all the trials and the resulting set plotted. Recall of a learnt pattern was tested by calculating its basin of attraction – in a learnt pattern chosen to test for recall the signs of its components, visualized as spins, are flipped systematically in a random sequence (which is called a sample) and convergence is checked at each step; the process is stopped when there is no more convergence. The number of flipped spins gives the Hamming distance for that sample. The process is repeated with different samples and the set of all Hamming distances gives the basin of attraction for that pattern. This protocol is described in greater detail in the box in the appendix B.

A.4 Method - Chapter 5

The network was built on GNU Octave/MATLAB. The size of the network was fixed at $N = 100$. The patterns were randomly-generated N-dimensional vectors with components having different signs but the same magnitude, namely 1 and normalized. The patterns were stored using the learning rule of eq.(2.3) and tested for retrieval and recognition and recall following eq.(2.9). The data was obtained always from 50 sets of patterns (or trials), collated and then arranged into bins. The network performance was evaluated in terms of capacity for retrieval and recall, i.e., estimation of basins of attraction following the same procedure as in the previous chapter.

Appendix B

More on basins of attraction

B.1 How to calculate a basin of attraction

1. Select the pattern $\boldsymbol{\xi}^{(\nu)}$ whose basin we wish to calculate. Present the test pattern $X = \boldsymbol{\xi}^{(\nu)}$ to the network for retrieval following eq.(2.9). Let the recovered pattern be X'. If $X' = X$, then the network has retrieved X. Present X' back to the network using

$$X' = sgn(h')|1| \tag{B.1}$$

where h' is calculated from eq.(2.4) for pattern X' (as the patterns presented to the network have components with magnitude ± 1) to obtain X''. If $X'' = X'$, then we call it convergence. If $X'' = X' = X$, then we can say X'' and X' converge to X. Pattern $X = \boldsymbol{\xi}^{(\nu)}$ is hence an attractor.

2. Make a sample ($S1$) by choosing a random sequence of $N/2$ elements. This sequence contains the indices of the elements of the test pattern to be flipped. Sequentially flip or change the signs of the components of the test pattern corresponding to these indices. For instance, for $N = 10$,

S1	10	7	5	3	6

Flip the element of X whose index is the first element of $S1$ to get pattern X_1 and present it to the network for retrieval to get pattern X_1'. If $X_1' \neq X_1$, then present X_1' back to the network to get $X_1^{(1)}$, where the 1 in the superscript refers to an *iteration*.

X	1	-1	-1	-1	1	-1	1	-1	*1*
X_1	1	-1	-1	-1	1	-1	1	-1	*-1*
X_1'	1	-1	-1	-1	1	-1	1	-1	*1*
$X_1^{(1)}$	1	-1	-1	-1	1	-1	1	-1	*1*

Now, if $X_1^{(1)} = X_1'$, then we say that X_1' has converged. Additionally, if $X_1^{(1)} = X$, then we say that X_1 converges to X in one iteration. Hence, pattern X_1 which is one flip away from X is associated with X.

If $X_1^{(1)} \neq X_1$, then we present it for one more iteration, and repeat until either convergence or the maximum number of iterations (fixed at 10, say) is reached.

Convergence happens only at an attractor, so if X is an attractor, $X_1' = X_1$ would not be possible.

3. If X_1 converges to X, then prepare pattern X_2 by flipping two elements of X whose indices correspond to the first two elements of $S1$. Now present X_2 to the network and check for convergence to X.

X	1	-1	-1	-1	1	*-1*	1	-1	*1*
X_2	1	-1	-1	-1	1	*1*	1	-1	*-1*
$X_2^{(1)}$	*-1*	*1*	-1	-1	1	-1	1	-1	*-1*
$X_2^{(2)}$	1	-1	*1*	-1	1	*1*	1	-1	*1*
$X_2^{(3)}$	1	-1	-1	-1	1	*-1*	1	-1	*1*
$X_2^{(4)}$	1	-1	-1	-1	1	*-1*	1	-1	*1*

Here, we see that X_2 converges to X, but in 3 iterations. Hence, X_2 which is 2 flips away from X is also associated with X.

4. Repeat the process as long as there is convergence. If test pattern X_b with b flips converges to X, then present X_{b+1} with $b+1$ flips and check for convergence. If there is no convergence to X even after 10 iterations, then stop the process.

5. The maximum value of b for which X_b converges to X within the specified number of iterations gives the Hamming distance b_1 for sample $S1$.

6. Repeat the above steps with more samples (for upto $s = 10$ samples, say), checking for convergence to X at each step, to get the Hamming distances $b_2, b_3, \ldots b_{10}$ for samples $S2$, $S3$, ..., $S10$.

7. Get the basin of attraction of pattern $X = \boldsymbol{\xi}^{(\nu)}$ as $B = \{b_1, b_2, \ldots, b_s\}$ for s samples.

After orthogonalization, the process is repeated to obtain the basins of attraction of the raw input patterns $\boldsymbol{\xi}^{(\nu)}$s'. However, to calculate the basins of attraction of the $\boldsymbol{\eta}$'s , we start with $X = \boldsymbol{\eta}^{(\nu)}$. For further iterations, we present X'

$$X' = sgn(h')|X|. \tag{B.2}$$

with h' calculated from eq.(3.13) using $\boldsymbol{\eta}^{(\nu)}$ and the magnitude of X.

B.2 Memory capacity and average basin size

We can now see how the memory capacity of the network is related to the average basin size. The average radius of the basin of attraction of the Hopfield network can be expressed in terms of Hamming distance, and provides an estimate of the limit on the number of patterns that can be learnt by the network. The Hamming distance $d_H^{(\mu,\nu)}$ between two patterns μ and ν is given by the number of sites q such that $\{\xi_i^{(\mu)} \neq \xi_i^{(\nu)}\}$ for q values of i randomly distributed between 1 and N. We inscribe p patterns in the network and present a random test pattern $\boldsymbol{\xi}^{(\boldsymbol{t})}$ for retrieval. Then,

$$\begin{aligned} h_i^{(t)} &= \sum_{\substack{j=1 \\ j\neq i}}^{N} J_{ij}\xi_j^{(t)} \\ &= \frac{1}{N}\left[\sum_{\mu=1}^{p} \xi_i^{(\mu)} \boldsymbol{\xi}^{(\boldsymbol{\mu})} \cdot \boldsymbol{\xi}^{(\boldsymbol{t})} - p\xi_i^{(t)}\right]. \end{aligned} \tag{B.3}$$

If $d_H^{(t,\mu)}$ is the Hamming distance between the test pattern $\boldsymbol{\xi^{(t)}}$ and the μ^{th} pattern $\boldsymbol{\xi^{(\mu)}}$, then,

$$\boldsymbol{\xi^{(\mu)}} \cdot \boldsymbol{\xi^{(t)}} = N - 2{d_H}^{(t,\mu)}, \tag{B.4}$$

since $\boldsymbol{\xi^{(\mu)}} \cdot \boldsymbol{\xi^{(\mu)}} = N$ and $\boldsymbol{\xi^{(t)}}$ differs from $\boldsymbol{\xi^{(\mu)}}$ on $d_H^{(t,\mu)}$ elements.

The average Hamming distance between $\boldsymbol{\xi^{(\mu)}}$ and $\boldsymbol{\xi^{(t)}}$ is,

$$\tilde{d}_H^{(t,\mu)} = \frac{d_H^{(t,1)} + d_H^{(t,2)} + \ldots + d_H^{(t,p)}}{p}. \tag{B.5}$$

We can now rewrite (B.3) as:

$$h_i^{(t)} = \frac{1}{N}\left[N\sum_{\mu=1}^{p}\xi_i^{(\mu)} - 2p\sum_{\mu=1}^{p}\xi_i^{(\mu)}\tilde{d}_H^{(t,\mu)} - p\xi_i^{(t)}\right]. \tag{B.6}$$

If the test pattern is within the basin of attraction of a stored pattern, it must converge to that pattern, and so must satisfy the retrieval criterion (eqn. (2.9)), i.e.

$$sgn\left(\xi_i^{(\mu)}\right) = sgn\left(h_i^{(t)}\right), \text{ for all } i\text{'s.}$$

This holds as long as the average distance between the test pattern and a stored pattern is[91],

$$\tilde{d}_H^{(t,\mu)} < \frac{(N-1)}{2p}. \tag{B.7}$$

This is a useful relation showing how a typical basin size can estimate the memory capacity, or vice versa.

B.3 Basins of attraction after orthogonalization

We now estimate the radii of basins of attraction following orthogonalization to study the effects of noise elimination on basin size.

We present a test pattern $\boldsymbol{\xi^{(t)}}$ for retrieval such that its first b elements are the same as those of $\boldsymbol{\xi^{(\nu)}}$ and the remaining $N-b$ elements are the inverses of the corresponding elements of $\boldsymbol{\xi^{(\nu)}}$. That is,

$$\boldsymbol{\xi^{(t)}} = \sum_{i=1}^{N-b}\left(\xi_i^{(\nu)}\right) - \sum_{i=N-b+1}^{N}\left(\xi_i^{(\nu)}\right) = \boldsymbol{\xi^{(\nu)}} - \sum_{i=N-b+1}^{N}\left(2\xi_i^{(\nu)}\right). \tag{B.8}$$

The maximum value of b gives a Hamming distance in the basin of attraction of $\boldsymbol{\xi^{(\nu)}}$. Now, calculating $h_i^{(\nu)}$ with the new weights calculated from the set $\{\boldsymbol{\eta^{(\mu)}}\}$,

$$\begin{aligned} h_i^{(t)} &= \sum_{\substack{j=1 \\ j\neq i}}^{N} J_{ij}\xi_j^{(t)} \\ &= \sum_{\mu=1}^{p} \eta_i^{(\mu)} \left[\sum_j \eta_j^{(\mu)}\xi_j^{(t)} - \eta_i^{(\mu)}\xi_i^{(t)} \right] \end{aligned} \tag{B.9}$$

Substituting for $\boldsymbol{\xi^{(t)}}$ using (B.8),

$$h_i^{(t)} = \sum_{\mu=1}^{p} \eta_i^{(\mu)} \left[\sum_{j=1}^{N-b} \eta_j^{(\mu)}\xi_j^{(t)} + \sum_{j=N-b+1}^{N} \eta_j^{(\mu)} \left(\xi_j^{(\nu)} - 2\xi_j^{(t)} \right) - \eta_i^{(\mu)}\xi_i^{(t)} \right]. \tag{B.10}$$

The complete set of h_i's corresponding to pattern ν is:

$$\begin{aligned} \boldsymbol{h^{(t)}} = \{h_i^{(t)}\} &= \sum_{\mu=1}^{p} \left[\boldsymbol{\eta^{(\mu)}} \left(\boldsymbol{\eta^{(\mu)}} \cdot \boldsymbol{\xi^{(\nu)}} \right) - 2\boldsymbol{\eta^{(\mu)}} \sum_{j=N-b+1}^{N} \eta_j^{(\mu)}\xi_j^{(\nu)} - \left(\boldsymbol{\eta^{(\mu)}}\right)^2 \xi_i^{(t)} \right] \\ &= \boldsymbol{\xi^{(\nu)}} - 2\boldsymbol{\hat{\eta}^{(\nu)}} \sum_{j=N-b+1}^{N} \underset{\sim}{\eta_j}^{(\mu)}\xi_j^{(\nu)} - \sum_{\mu=1}^{p} \left(\boldsymbol{\hat{\eta}^{(\mu)}}\right)^2 \boldsymbol{\xi^{(t)}}, \end{aligned} \tag{B.11}$$

since $\sum_{\mu=1}^{p} \boldsymbol{\hat{\eta}^{(\mu)}} \left(\boldsymbol{\hat{\eta}^{(\mu)}} \cdot \boldsymbol{\xi^{(\nu)}} \right) = \boldsymbol{\xi^{(\nu)}}$ for normalized $\boldsymbol{\eta}$'s. We can now calculate the stabilization parameter as,

$$\begin{aligned} \mathfrak{s}^{(\nu)} &= \boldsymbol{h^{(t)}} \cdot \boldsymbol{\xi^{(\nu)}} \\ &= \boldsymbol{\xi^{(\nu)}} \cdot \boldsymbol{\xi^{(\nu)}} - 2\sum_{\mu=1}^{p} \boldsymbol{\hat{\eta}^{(\nu)}} \cdot \boldsymbol{\xi^{(\nu)}} \sum_{j=N-b+1}^{N} \underset{\sim}{\eta_j}^{(\mu)}\xi_j^{(\nu)} - \sum_{\mu=1}^{p} \left(\boldsymbol{\hat{\eta}^{(\mu)}}\right)^2 \boldsymbol{\xi^{(t)}} \cdot \boldsymbol{\xi^{(\nu)}} \end{aligned} \tag{B.12}$$

We can evaluate the first term of this equation explicitly as $\boldsymbol{\xi^{(\nu)}} \cdot \boldsymbol{\xi^{(\nu)}} = N$ for $\boldsymbol{\xi^{(\nu)}}$'s with components $\xi_i^{(\nu)} = \pm 1$. The dot product in the last term would yield $N - 2b$, and the whole term will thus be $p(N - 2b)$. As $\boldsymbol{\xi^{(\nu)}}$ does not project onto $\boldsymbol{\eta^{(\mu)}}$ for $\mu > \nu$, the summation in the second term is only upto the first ν patterns. Thus,

$$\mathfrak{s}^{(\nu)} = N - 2\sum_{\mu=1}^{\nu} \boldsymbol{\hat{\eta}^{(\nu)}} \cdot \boldsymbol{\xi^{(\nu)}} \sum_{j=N-b+1}^{N} \underset{\sim}{\eta_j}^{(\mu)}\xi_j^{(\nu)} - p(N - 2b). \tag{B.13}$$

Thus, the stabilization parameter will always be positive upto $p = N - 1$, ensuring the recognition of patterns within that limit. Thus, the $\boldsymbol{\xi}$'s will all be energy minima

for upto $p = N - 1$, but as p approaches $N - 1$, although $\mathfrak{s}^{(\nu)} > 0$, its magnitude decreases, and so the energy of the patterns goes up.

Appendix C

Comparison of the H-H and H-H-GS models - some considerations

C.1 Comparison of the H-H and H-H-GS models

We reiterate in Table C.1 below the various network parameters of the H-H and H-H-GS networks and present them side-by-side for ease of comparison.

TABLE C.1: Table summarizing various parameters of the H-H and H-H-GS models for comparison

Parameter	H-H model	H-H-GS model
Synaptic efficacy (J_{ij})	$\frac{1}{N}\sum_{\mu=1}^{p}\left(\xi_i^{(\mu)}\xi_j^{(\mu)} - \delta_{ij}\xi_i^{(\mu)}\xi_i^{(\mu)}\right)$	$\frac{1}{N}\sum_{\mu=1}^{p}\left(\hat{\eta}^{(\mu)}\hat{\eta}_j^{(\mu)} - \delta_{ij}\hat{\eta}_i^{(\mu)}\hat{\eta}_i^{(\mu)}\right)$
Post-synaptic potential/PSP ($\boldsymbol{h}^{(\nu)}$)	$\left(1-\mathcal{O}\left(\frac{p}{N}\right)\right)\boldsymbol{\xi}^{(\nu)} - \frac{1}{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p}\xi_i^{(\mu)}\left(\boldsymbol{\xi}^{(\mu)}.\boldsymbol{\xi}^{(\nu)}\right)$	$\left(1-\mathcal{O}\left(\frac{p}{N}\right)\right)\boldsymbol{\xi}^{(\nu)}$
Stabilization parameter ($s^{(\nu)}$)	$N - p - \frac{1}{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p}\left(\boldsymbol{\xi}^{(\mu)}.\boldsymbol{\xi}^{(\nu)}\right)^2$	$N - \mathcal{O}\left(\frac{p}{N}\right)N$
Pattern energy ($E\left(\boldsymbol{\xi}^{(\nu)}\right)$)	$-\frac{N}{2} + \frac{p}{2} + \frac{1}{2N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p}\left(\boldsymbol{\xi}^{(\mu)}.\boldsymbol{\xi}^{(\nu)}\right)^2$	$-\frac{N}{2} + \frac{N}{2}\left(\mathcal{O}\left(\frac{p}{N}\right)\right)$

C.2 Range of basins of attraction - comparison

TABLE C.2: The following table shows the extent of basins of attraction in the H-H and H-H-GS(shown within parentheses) networks of size $N = 100$. The Hamming distances constituting the basins of attraction are calculated for 50 trials with 10 samples per pattern for different values of p. The Hamming distances for all the samples for a particular value of p are examined to get the possible extent of a basin of attraction at that value of p. For instance, for $p = 2$, with 10 samples for each pattern and for 50 trials, there would $50 * 2 * 10 = 1000$ Hamming distances, the minimum and maximum among which give the range the basin of attraction of any of a pair of (/the $p = 2$) patterns can take. The tabulated values show the ranges for different values of p. The smaller the range, more the isotropy, and greater the range, more the anisotropy. We can see that the basins in the H-H network are initially somewhat isotropic, but become more anisotropic with increasing p. However, after orthogonalization, the corresponding basins for the same values of p are relatively more isotropic and remain so even for higher values of p. Also note that there are 0's in the basins of attraction for p as low as 8 in the H-H model, while the basin size remains large in the case of H-H-GS network, even for higher values of p, such as $p = 30$, shown here.

p	Minimum	Maximum
2	32 (30)	49 (50)
4	22 (23)	49 (49)
6	6 (26)	49 (48)
8	0 (28)	50 (47)
10	0 (25)	50 (48)
12	0 (26)	49 (48)
14	0 (24)	49 (47)
16	0 (24)	49 (46)
18	0 (21)	44 (45)
20	0 (23)	39 (46)
22	0 (21)	39 (43)
24	0 (20)	38 (43)
26	0 (18)	37 (41)
28	0 (18)	33 (41)
30	0 (16)	32 (39)

C.3 Probability of zero basins of attraction

The following figure (Fig. C.1) shows the probability of zero basins of attraction in the H-H and H-H-GS models.

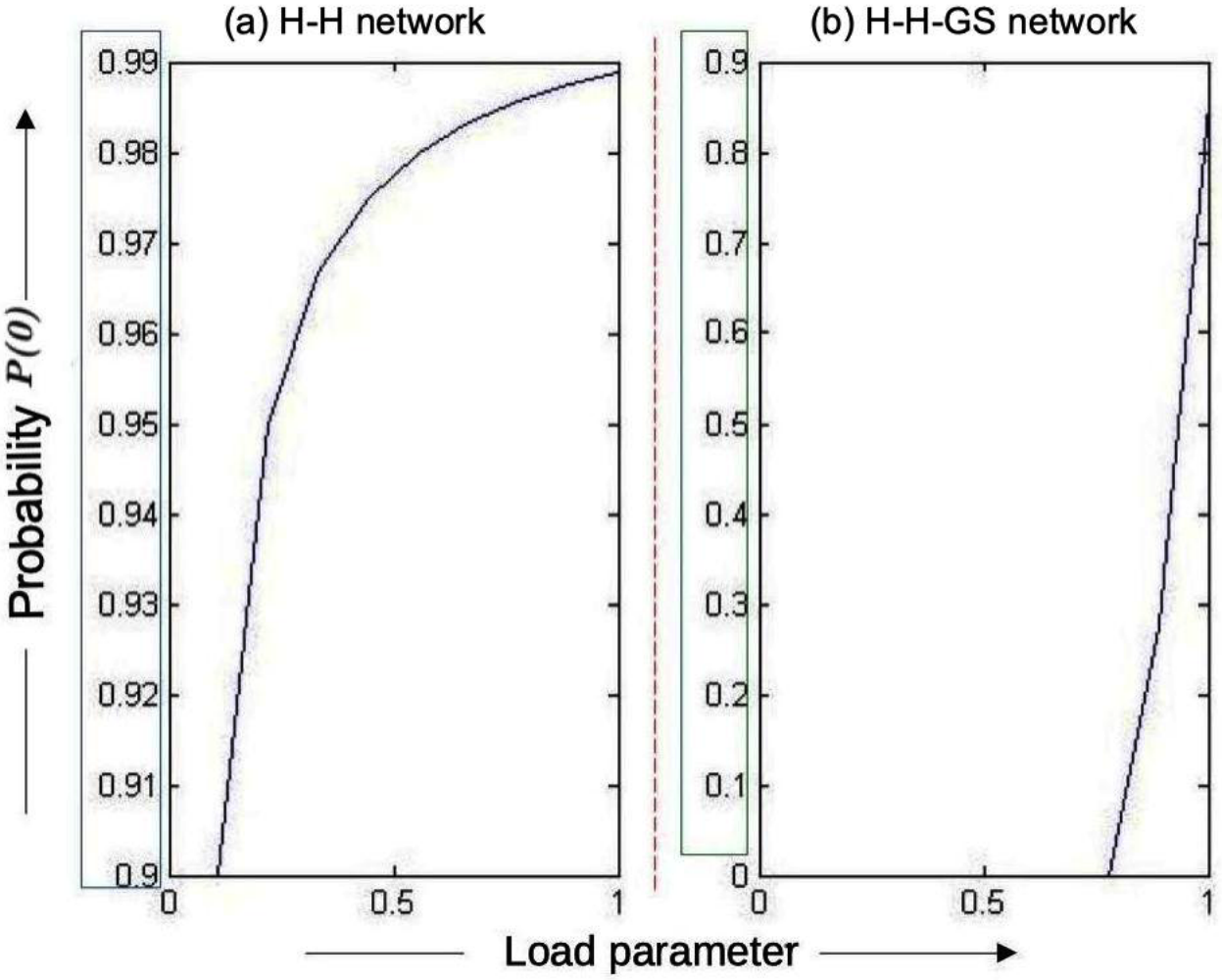


FIGURE C.1: Probability of finding basin of attraction of size zero (*P(0)*) for $p = 1 - 100$ (in steps of 10) in the H-H (a) and H-H-GS (b) networks. The data is shown for a single trial in a network of size $N = 100$. Note the difference in the Y-axis. As we can see from Fig. (a), even for low values of p, around $p = 10$, the probability of finding a zero basin is about 0.9 in the H-H network, while Fig. (b) shows how after orthogonalization, the probability of finding a zero basin is non-zero only much beyond $p = 50$.

Appendix D

Numerical example of the equivalence of various orthogonalization schemes

We find computationally that the weights calculated using the orthonormal bases obtained from the various orthogonalization schemes discussed in this thesis, are the same, even though the bases themselves are very different from each other. We refer to this phenomenon as "equivalence".

D.1 Numerical demonstration

We present here a numerical example of the equivalence. For a set of input vectors ($\mathbf{V}$), we calculate the orthonormal bases using Gram-Schmidt ($\mathbf{G}$) and Löwdin's Symmetric ($\mathbf{S}$) and Canonical ($\mathbf{C}$) orthogonalization schemes. The corresponding constituent vectors are given by $\boldsymbol{v}$'s, $\boldsymbol{g}$'s, $\boldsymbol{s}$'s and $\boldsymbol{c}$'s respectively.We then calculate the weights using each orthonormal basis and compare them.

We first look at the weights when $p = 2$.

TABLE D.1: Unnormalized input patterns

V	1	-1	-1	1	-1	1	1	-1	1	-1
	1	-1	-1	1	-1	-1	-1	-1	1	1
	1	-1	1	1	-1	-1	1	-1	1	-1

TABLE D.2: Orthonormal bases for $p = 2$

G	0.3162	−0.3162	−0.3162	0.3162	−0.3162	0.3162	0.3162	−0.3162	0.3162	−0.3162
	0.2070	−0.2070	−0.2070	0.2070	−0.2070	−0.4830	−0.4830	−0.2070	0.2070	0.4830
S	0.2673	−0.2673	−0.2673	0.2673	−0.2673	0.4082	0.4082	−0.2673	0.2673	−0.4082
	0.2673	−0.2673	−0.2673	0.2673	−0.2673	−0.4082	−0.4082	−0.2673	0.2673	0.4082
C	−0.3780	0.3780	0.3780	−0.3780	0.3780	−0	−0	0.3780	−0.3780	0
	0	0	0	0	0	−0.5774	−0.5774	0	0	0.5774

Calculating weights using $\mathbf{G}$:

$$J = J^1 + J^2 = (\boldsymbol{g}^1)' * (\boldsymbol{g}^1) + (\boldsymbol{g}^2)' * (\boldsymbol{g}^2). \tag{D.1}$$

$$\begin{bmatrix}
0 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 \\
-0.1 & 0 & 0.1 & -0.1 & 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\
-0.1 & 0.1 & 0 & -0.1 & 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\
0.1 & -0.1 & -0.1 & 0 & -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 \\
-0.1 & 0.1 & 0.1 & -0.1 & 0 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\
0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0 & 0.1 & -0.1 & 0.1 & -0.1 \\
0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 & 0 & -0.1 & 0.1 & -0.1 \\
-0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 & -0.1 & 0 & -0.1 & 0.1 \\
0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 & 0.1 & -0.1 & 0 & -0.1 \\
-0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & -0.0429 & -0.0429 & 0.0429 & -0.0429 & -0.1 & -0.1 & -0.0429 & 0.0429 & 0.1 \\
-0.0429 & 0 & 0.0429 & -0.0429 & 0.0429 & 0.1 & 0.1 & 0.0429 & -0.0429 & -0.1 \\
-0.0429 & 0.0429 & 0 & -0.0429 & 0.0429 & 0.1 & 0.1 & 0.0429 & -0.0429 & -0.1 \\
0.0429 & -0.0429 & -0.0429 & 0 & -0.0429 & -0.1 & -0.1 & -0.0429 & 0.0429 & 0.1 \\
-0.0429 & 0.0429 & 0.0429 & -0.0429 & 0 & 0.1 & 0.1 & 0.0429 & -0.0429 & -0.1 \\
-0.1 & 0.1 & 0.1 & -0.1 & 0.1 & 0 & -0.2333 & 0.1 & -0.1 & -0.2333 \\
-0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.2333 & 0 & 0.1 & -0.1 & -0.2333 \\
-0.0429 & 0.0429 & 0.0429 & -0.0429 & 0.0429 & 0.1 & 0.1 & 0 & -0.0429 & -0.1 \\
0.0429 & -0.0429 & -0.0429 & 0.0429 & -0.0429 & -0.1 & -0.1 & -0.0429 & 0 & 0.1 \\
0.1 & -0.1 & -0.1 & 0.1 & -0.1 & -0.2333 & -0.2333 & -0.1 & 0.1 & 0
\end{bmatrix}$$

$$=$$

$$\begin{bmatrix}
0 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
-0.1429 & 0.1429 & 0 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & -0.3333 \\
0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & 0 & -0.3333 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.3333 & -0.3333 & 0 & 0 & 0
\end{bmatrix}$$

Calculating weights using **S**:

$$J = J^1 + J^2 = (\boldsymbol{s}^1)' * (\boldsymbol{s}^1) + (\boldsymbol{s}^2)' * (\boldsymbol{s}^2). \tag{D.2}$$

$$\begin{bmatrix}
0.0714 & -0.0714 & -0.0714 & 0.0714 & -0.0714 & 0.1091 & 0.1091 & -0.0714 & 0.0714 & -0.1091 \\
-0.0714 & 0.0714 & 0.0714 & -0.0714 & 0.0714 & -0.1091 & -0.1091 & 0.0714 & -0.0714 & 0.1091 \\
-0.0714 & 0.0714 & 0.0714 & -0.0714 & 0.0714 & -0.1091 & -0.1091 & 0.0714 & -0.0714 & 0.1091 \\
0.0714 & -0.0714 & -0.0714 & 0.0714 & -0.0714 & 0.1091 & 0.1091 & -0.0714 & 0.0714 & -0.1091 \\
-0.0714 & 0.0714 & 0.0714 & -0.0714 & 0.0714 & -0.1091 & -0.1091 & 0.0714 & -0.0714 & 0.1091 \\
0.1091 & -0.1091 & -0.1091 & 0.1091 & -0.1091 & 0.1667 & 0.1667 & -0.1091 & 0.1091 & -0.1667 \\
0.1091 & -0.1091 & -0.1091 & 0.1091 & -0.1091 & 0.1667 & 0.1667 & -0.1091 & 0.1091 & -0.1667 \\
-0.0714 & 0.0714 & 0.0714 & -0.0714 & 0.0714 & -0.1091 & -0.1091 & 0.0714 & -0.0714 & 0.1091 \\
0.0714 & -0.0714 & -0.0714 & 0.0714 & -0.0714 & 0.1091 & 0.1091 & -0.0714 & 0.0714 & -0.1091 \\
-0.1091 & 0.1091 & 0.1091 & -0.1091 & 0.1091 & -0.1667 & -0.1667 & 0.1091 & -0.1091 & 0.1667
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & -0.0714 & -0.0714 & 0.0714 & -0.0714 & -0.1091 & -0.1091 & -0.0714 & 0.0714 & 0.1091 \\
-0.0714 & 0 & 0.0714 & -0.0714 & 0.0714 & 0.1091 & 0.1091 & 0.0714 & -0.0714 & -0.1091 \\
-0.0714 & 0.0714 & 0 & -0.0714 & 0.0714 & 0.1091 & 0.1091 & 0.0714 & -0.0714 & -0.1091 \\
0.0714 & -0.0714 & -0.0714 & 0 & -0.0714 & -0.1091 & -0.1091 & -0.0714 & 0.0714 & 0.1091 \\
-0.0714 & 0.0714 & 0.0714 & -0.0714 & 0 & 0.1091 & 0.1091 & 0.0714 & -0.0714 & -0.1091 \\
-0.1091 & 0.1091 & 0.1091 & -0.1091 & 0.1091 & 0 & 0.1667 & 0.1091 & -0.1091 & -0.1667 \\
-0.1091 & 0.1091 & 0.1091 & -0.1091 & 0.1091 & 0.1667 & 0 & 0.1091 & -0.1091 & -0.1667 \\
-0.0714 & 0.0714 & 0.0714 & -0.0714 & 0.0714 & 0.1091 & 0.1091 & 0 & -0.0714 & -0.1091 \\
0.0714 & -0.0714 & -0.0714 & 0.0714 & -0.0714 & -0.1091 & -0.1091 & -0.0714 & 0 & 0.1091 \\
0.1091 & -0.1091 & -0.1091 & 0.1091 & -0.1091 & -0.1667 & -0.1667 & -0.1091 & 0.1091 & 0
\end{bmatrix}$$

$$=$$

$$\begin{bmatrix}
0 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
-0.1429 & 0.1429 & 0 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & -0.3333 \\
0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & 0 & -0.3333 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.3333 & -0.3333 & 0 & 0 & 0
\end{bmatrix}$$

Calculating weights using $\mathbf{C}$:

$$J = J^1 + J^2 = (\boldsymbol{c}^1)' * (\boldsymbol{c}^1) + (\boldsymbol{c}^2)' * (\boldsymbol{c}^2). \tag{D.3}$$

$$\begin{bmatrix}
0 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
-0.1429 & 0.1429 & 0 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & -0.3333 \\
0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & 0 & -0.3333 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.3333 & -0.3333 & 0 & 0 & 0
\end{bmatrix}$$

$$=$$

$$\begin{bmatrix}
0 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
-0.1429 & 0.1429 & 0 & -0.1429 & 0.1429 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0 & -0.1429 & 0 & 0 & -0.1429 & 0.1429 & 0 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0 & 0 & 0 & 0.1429 & -0.1429 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & -0.3333 \\
0 & 0 & 0 & 0 & 0 & 0.3333 & 0 & 0 & 0 & -0.3333 \\
-0.1429 & 0.1429 & 0.1429 & -0.1429 & 0.1429 & 0 & 0 & 0 & -0.1429 & 0 \\
0.1429 & -0.1429 & -0.1429 & 0.1429 & -0.1429 & 0 & 0 & -0.1429 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.3333 & -0.3333 & 0 & 0 & 0
\end{bmatrix}$$

We now consider the situation when $p = 3$ and again calculate and compare the weights obtained using the various orthonormal bases.

TABLE D.3: Orthonormal bases for $p = 3$

G	0.3162	−0.3162	−0.3162	0.3162	−0.3162	0.3162	0.3162	−0.3162	0.3162	−0.3162
	0.2070	−0.2070	−0.2070	0.2070	−0.2070	−0.4830	−0.4830	−0.2070	0.2070	0.4830
	0.1157	−0.1157	0.6944	0.1157	−0.1157	−0.5401	0.2700	−0.1157	0.1157	−0.2700
S	0.2195	−0.2195	−0.4493	0.2195	−0.2195	0.5507	0.3209	−0.2195	0.2195	−0.3209
	0.2448	−0.2448	−0.3462	0.2448	−0.2448	−0.3462	−0.4476	−0.2448	0.2448	0.4476
	0.2195	−0.2195	0.5507	0.2195	−0.2195	−0.4493	0.3209	−0.2195	0.2195	−0.3209
C	−0.3921	0.3921	0.1170	−0.3921	0.3921	0.1170	−0.1581	0.3921	−.3921	0.1581
	−0.0498	0.0498	0.3336	−0.0498	0.0498	0.3336	0.6175	0.0498	−0.0498	−0.6175
	0	0	0.7071	0	0	−0.7071	0	0	0	0

The weights are now calculated as:

- Gram-Schmidt

$$J = J^1 + J^2 + J^3 = (\boldsymbol{g}^1)' * (\boldsymbol{g}^1) + (\boldsymbol{g}^2)' * (\boldsymbol{g}^2) + (\boldsymbol{g}^3)' * (\boldsymbol{g}^3), \tag{D.4}$$

- Symmetric

$$J = J^1 + J^2 + J^3 = (\boldsymbol{s}^1)' * (\boldsymbol{s}^1) + (\boldsymbol{s}^2)' * (\boldsymbol{s}^2) + (\boldsymbol{s}^3)' * (\boldsymbol{s}^3), \tag{D.5}$$

- Canonical

$$J = J^1 + J^2 + J^3 = (\boldsymbol{c}^1)' * (\boldsymbol{c}^1) + (\boldsymbol{c}^2)' * (\boldsymbol{c}^2) + (\boldsymbol{c}^3)' * (\boldsymbol{c}^3). \tag{D.6}$$

Calculating weights using **G**:

$$\begin{bmatrix} 0 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 \\ -0.1 & 0 & 0.1 & -0.1 & 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ -0.1 & 0.1 & 0 & -0.1 & 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ 0.1 & -0.1 & -0.1 & 0 & -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 \\ -0.1 & 0.1 & 0.1 & -0.1 & 0 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 \\ 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0 & 0.1 & -0.1 & 0.1 & -0.1 \\ 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 & 0 & -0.1 & 0.1 & -0.1 \\ -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 & -0.1 & 0 & -0.1 & 0.1 \\ 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0.1 & 0.1 & -0.1 & 0 & -0.1 \\ -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & 0 \end{bmatrix}$$

$$+$$

$$\begin{bmatrix} 0 & -0.0429 & -0.0429 & 0.0429 & -0.0429 & -0.1 & -0.1 & -0.0429 & 0.0429 & 0.1 \\ -0.0429 & 0 & 0.0429 & -0.0429 & 0.0429 & 0.1 & 0.1 & 0.0429 & -0.0429 & -0.1 \\ -0.0429 & 0.0429 & 0 & -0.0429 & 0.0429 & 0.1 & 0.1 & 0.0429 & -0.0429 & -0.1 \\ 0.0429 & -0.0429 & -0.0429 & 0 & -0.0429 & -0.1 & -0.1 & -0.0429 & 0.0429 & 0.1 \\ -0.0429 & 0.0429 & 0.0429 & -0.0429 & 0 & 0.1 & 0.1 & 0.0429 & -0.0429 & -0.1 \\ -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & 0 & -0.2333 & 0.1 & -0.1 & -0.2333 \\ -0.1 & 0.1 & 0.1 & -0.1 & 0.1 & -0.2333 & 0 & 0.1 & -0.1 & -0.2333 \\ -0.0429 & 0.0429 & 0.0429 & -0.0429 & 0.0429 & 0.1 & 0.1 & 0 & -0.0429 & -0.1 \\ 0.0429 & -0.0429 & -0.0429 & 0.0429 & -0.0429 & -0.1 & -0.1 & -0.0429 & 0 & 0.1 \\ 0.1 & -0.1 & -0.1 & 0.1 & -0.1 & -0.2333 & -0.2333 & -0.1 & 0.1 & 0 \end{bmatrix}$$

$$+$$

$$\begin{bmatrix} 0 & -0.0133 & 0.0804 & 0.0133 & -0.0133 & -0.0625 & 0.0313 & -0.0133 & 0.0133 & -0.0313 \\ -0.0133 & 0 & -0.0804 & -0.0133 & 0.0133 & 0.0625 & -0.0313 & 0.0133 & -0.0133 & 0.0313 \\ 0.0804 & -0.0804 & 0 & 0.0804 & -0.0804 & -0.3750 & 0.1875 & -0.0804 & 0.0804 & -0.1875 \\ 0.0133 & -0.0133 & 0.0804 & 0 & -0.0133 & -0.0625 & 0.0313 & -0.0133 & 0.0133 & -0.0313 \\ -0.0133 & 0.0133 & -0.0804 & -0.0133 & 0 & 0.0625 & -0.0313 & 0.0133 & -0.0133 & 0.0313 \\ -0.0625 & 0.0625 & -0.3750 & -0.0625 & 0.0625 & 0 & -0.1458 & 0.0625 & -0.0625 & 0.1458 \\ 0.0313 & -0.0313 & 0.1875 & 0.0313 & -0.0313 & -0.1458 & 0 & -0.0313 & 0.0313 & -0.0729 \\ -0.0133 & 0.0133 & -0.0804 & -0.0133 & 0.0133 & 0.0625 & -0.0313 & 0 & -0.0133 & 0.0313 \\ 0.0133 & -0.0133 & 0.0804 & 0.0133 & -0.0133 & -0.0625 & 0.0313 & -0.0133 & 0 & -0.0313 \\ -0.0313 & 0.0313 & -0.1875 & -0.0313 & 0.0313 & 0.1458 & -0.0729 & 0.0313 & -0.0313 & 0 \end{bmatrix}$$

$$=$$

$$\begin{bmatrix} 0 & -0.1563 & -0.0625 & 0.1563 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0.1563 & -0.0313 \\ -0.1563 & 0 & 0.0625 & -0.1563 & 0.1563 & 0.0625 & -0.0313 & 0.1563 & -0.1563 & 0.0313 \\ -0.0625 & 0.0625 & 0 & -0.0625 & 0.0625 & -0.3750 & 0.1875 & 0.0625 & -0.0625 & -0.1875 \\ 0.1563 & -0.1563 & -0.0625 & 0 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0.1563 & -0.0313 \\ -0.1563 & 0.1563 & 0.0625 & -0.1563 & 0 & 0.0625 & -0.0313 & 0.1563 & -0.1563 & 0.0313 \\ -0.0625 & 0.0625 & -0.3750 & -0.0625 & 0.0625 & 0 & 0.1875 & 0.0625 & -0.0625 & -0.1875 \\ 0.0313 & -0.0313 & 0.1875 & 0.0313 & -0.0313 & 0.1875 & 0 & -0.0313 & 0.0313 & -0.4063 \\ -0.1563 & 0.1563 & 0.0625 & -0.1563 & 0.1563 & 0.0625 & -0.0313 & 0 & -0.1563 & 0.0313 \\ 0.1563 & -0.1563 & -0.0625 & 0.1563 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0 & -0.0313 \\ -0.0313 & 0.0313 & -0.1875 & -0.0313 & 0.0313 & -0.1875 & -0.406\,25 & 0.0313 & -0.0313 & 0 \end{bmatrix}$$

Calculating weights using **S**:

$$\begin{bmatrix}
0 & -0.0482 & -0.0986 & 0.0482 & -0.0482 & 0.1209 & 0.0704 & -0.0482 & 0.0482 & -0.0704 \\
-0.0482 & 0 & 0.0986 & -0.0482 & 0.0482 & -0.1209 & -0.0704 & 0.0482 & -0.0482 & -0.0704 \\
-0.0986 & 0.0986 & 0 & -0.0986 & 0.0986 & -0.2474 & -0.1442 & 0.0986 & -0.0986 & 0.1442 \\
0.0482 & -0.0482 & -0.0986 & 0 & -0.0482 & 0.1209 & 0.0704 & -0.0482 & 0.0482 & -0.0704 \\
-0.0482 & 0.0482 & 0.0986 & -0.0482 & 0.0 & -0.1209 & -0.0704 & 0.0482 & -0.0482 & -0.0704 \\
0.1209 & -0.1209 & -0.2474 & 0.1209 & -0.1209 & 0 & 0.1767 & -0.1209 & 0.1209 & -0.1767 \\
0.0704 & -0.0704 & -0.1442 & 0.0704 & -0.0704 & 0.1767 & 0 & -0.0704 & 0.0704 & -0.1030 \\
-0.0482 & 0.0482 & 0.0986 & -0.0482 & 0.0482 & -0.1209 & -0.0704 & 0 & -0.0482 & -0.0704 \\
0.0482 & -0.0482 & -0.0986 & 0.0482 & -0.0482 & 0.1209 & 0.0704 & -0.0482 & 0 & -0.0704 \\
-0.0704 & 0.0704 & 0.1442 & -0.0704 & 0.0704 & -0.1767 & -0.1030 & 0.0704 & -0.0704 & 0
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & -0.0599 & -0.0848 & 0.0599 & -0.0599 & -0.0848 & -0.1096 & -0.0599 & 0.0599 & 0.1096 \\
-0.0599 & 0 & 0.0848 & -0.0599 & 0.0599 & 0.0848 & 0.1096 & 0.0599 & -0.0599 & -0.1096 \\
-0.0848 & 0.0848 & 0 & -0.0848 & 0.0848 & 0.1199 & 0.1550 & 0.0848 & -0.0848 & -0.1550 \\
0.0599 & -0.0599 & -0.0848 & 0 & -0.0599 & -0.0848 & -0.1096 & -0.0599 & 0.0599 & 0.1096 \\
-0.0599 & 0.0599 & 0.0848 & -0.0599 & 0 & 0.0848 & 0.1096 & 0.0599 & -0.0599 & -0.1096 \\
-0.0848 & 0.0848 & 0.1199 & -0.0848 & 0.0848 & 0 & 0.1550 & 0.0848 & -0.0848 & -0.1550 \\
-0.1096 & 0.1096 & 0.1550 & -0.1096 & 0.1096 & 0.1550 & 0 & 0.1096 & -0.1096 & -0.2004 \\
-0.0599 & 0.0599 & 0.0848 & -0.0599 & 0.0599 & 0.0848 & 0.1096 & 0 & -0.0599 & -0.1096 \\
0.0599 & -0.0599 & -0.0848 & 0.0599 & -0.0599 & -0.0848 & -0.1096 & -0.0599 & 0 & 0.1096 \\
0.1096 & -0.1096 & -0.1550 & 0.1096 & -0.1096 & -0.1550 & -0.2004 & -0.1096 & 0.1096 & 0
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & -0.0482 & 0.1209 & 0.0482 & -0.0482 & -0.0986 & 0.0704 & -0.0482 & 0.0482 & -0.0704 \\
-0.0482 & 0 & -0.1209 & -0.0482 & 0.0482 & 0.0986 & -0.0704 & 0.0482 & -0.0482 & -0.0704 \\
0.1209 & -0.1209 & 0 & 0.1209 & -0.1209 & -0.2474 & 0.1767 & -0.1209 & 0.1209 & -0.1767 \\
0.0482 & -0.0482 & 0.1209 & 0 & -0.0482 & -0.0986 & 0.0704 & -0.0482 & 0.0482 & -0.0704 \\
-0.0482 & 0.0482 & -0.1209 & -0.0482 & 0 & 0.0986 & -0.0704 & 0.0482 & -0.0482 & -0.0704 \\
-0.0986 & 0.0986 & -0.2474 & -0.0986 & 0.0986 & 0 & -0.1442 & 0.0986 & -0.0986 & 0.1442 \\
0.0704 & -0.0704 & 0.1767 & 0.0704 & -0.0704 & -0.1442 & 0 & -0.0704 & 0.0704 & -0.1030 \\
-0.0482 & 0.0482 & -0.1209 & -0.0482 & 0.0482 & 0.0986 & -0.0704 & 0 & -0.0482 & -0.0704 \\
0.0482 & -0.0482 & 0.1209 & 0.0482 & -0.0482 & -0.0986 & 0.0704 & -0.0482 & 0 & -0.0704 \\
-0.0704 & 0.0704 & -0.1767 & -0.0704 & 0.0704 & 0.1442 & -0.1030 & 0.0704 & -0.0704 & 0
\end{bmatrix}$$

$$=$$

$$\begin{bmatrix}
0 & -0.1563 & -0.0625 & 0.1563 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0.1563 & -0.0313 \\
-0.1563 & 0 & 0.0625 & -0.1563 & 0.1563 & 0.0625 & -0.0313 & 0.1563 & -0.1563 & 0.0313 \\
-0.0625 & 0.0625 & 0 & -0.0625 & 0.0625 & -0.3750 & 0.1875 & 0.0625 & -0.0625 & -0.1875 \\
0.1563 & -0.1563 & -0.0625 & 0 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0.1563 & -0.0313 \\
-0.1563 & 0.1563 & 0.0625 & -0.1563 & 0 & 0.0625 & -0.0313 & 0.1563 & -0.1563 & 0.0313 \\
-0.0625 & 0.0625 & -0.3750 & -0.0625 & 0.0625 & 0 & 0.1875 & 0.0625 & -0.0625 & -0.1875 \\
0.0313 & -0.0313 & 0.1875 & 0.0313 & -0.0313 & 0.1875 & 0 & -0.0313 & 0.0313 & -0.4063 \\
-0.1563 & 0.1563 & 0.0625 & -0.1563 & 0.1563 & 0.0625 & -0.0313 & 0 & -0.1563 & 0.0313 \\
0.1563 & -0.1563 & -0.0625 & 0.1563 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0 & -0.0313 \\
-0.0313 & 0.0313 & -0.1875 & -0.0313 & 0.0313 & -0.1875 & -0.406\,25 & 0.0313 & -0.0313 & 0
\end{bmatrix}$$

Calculating weights using **C**:

$$\begin{bmatrix}
0 & -0.1538 & -0.0459 & 0.1538 & -0.1538 & -0.0459 & 0.0620 & -0.1538 & 0.1538 & -0.0620 \\
-0.1538 & 0 & 0.0459 & -0.1538 & 0.1538 & 0.0459 & -0.0620 & 0.1538 & -0.1538 & 0.0620 \\
-0.0459 & 0.0459 & 0 & -0.0459 & 0.0459 & -0.0137 & -0.0185 & 0.0459 & -0.0459 & 0.0185 \\
0.1538 & -0.1538 & -0.0459 & 0 & -0.1538 & -0.0459 & 0.0620 & -0.1538 & 0.1538 & -0.0620 \\
-0.1538 & 0.1538 & 0.0459 & -0.1538 & 0 & 0.0459 & -0.0620 & 0.1538 & -0.1538 & 0.0620 \\
-0.0459 & 0.0459 & -0.0137 & -0.0459 & 0.0459 & 0 & -0.0185 & 0.0459 & -0.0459 & 0.0185 \\
0.0620 & -0.0620 & -0.0185 & 0.0620 & -0.0620 & -0.0185 & 0 & -0.0620 & 0.0620 & -0.0250 \\
-0.1538 & 0.1538 & 0.0459 & -0.1538 & 0.1538 & 0.0459 & -0.0620 & 0 & -0.1538 & 0.0620 \\
0.1538 & -0.1538 & -0.0459 & 0.1538 & -0.1538 & -0.0459 & 0.0620 & -0.1538 & 0 & -0.0620 \\
-0.0620 & 0.0620 & 0.0185 & -0.0620 & 0.0620 & 0.0185 & -0.0250 & 0.0620 & -0.0620 & 0
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & -0.0025 & -0.0160 & 0.0025 & -0.0025 & -0.0160 & -0.0307 & -0.0025 & 0.0025 & 0.0307 \\
-0.0025 & 0 & 0.0160 & -0.0025 & 0.0025 & 0.0160 & 0.0307 & 0.0025 & -0.0025 & -0.0307 \\
-0.0160 & 0.0160 & 0 & -0.0160 & 0.0160 & 0.1113 & 0.2060 & 0.0160 & -0.0160 & -0.2060 \\
0.0025 & -0.0025 & -0.0160 & 0 & -0.0025 & -0.0160 & -0.0307 & -0.0025 & 0.0025 & 0.0307 \\
-0.0025 & 0.0025 & 0.0160 & -0.0025 & 0 & 0.0160 & 0.0307 & 0.0025 & -0.0025 & -0.0307 \\
-0.0160 & 0.0160 & 0.1113 & -0.0160 & 0.0160 & 0 & 0.2060 & 0.0160 & -0.0160 & -0.2060 \\
-0.0307 & 0.0307 & 0.2060 & -0.0307 & 0.0307 & 0.2060 & 0 & 0.0307 & -0.0307 & -0.3813 \\
-0.0025 & 0.0025 & 0.0160 & -0.0025 & 0.0025 & 0.0160 & 0.0307 & 0 & -0.0025 & -0.0307 \\
0.0025 & -0.0025 & -0.0160 & 0.0025 & -0.0025 & -0.0160 & -0.0307 & -0.0025 & 0 & 0.0307 \\
0.0307 & -0.0307 & -0.2060 & 0.0307 & -0.0307 & -0.2060 & -0.3813 & -0.0307 & 0.0307 & 0
\end{bmatrix}$$

$$+$$

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -0.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$=$$

$$\begin{bmatrix}
0 & -0.1563 & -0.0625 & 0.1563 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0.1563 & -0.0313 \\
-0.1563 & 0 & 0.0625 & -0.1563 & 0.1563 & 0.0625 & -0.0313 & 0.1563 & -0.1563 & 0.0313 \\
-0.0625 & 0.0625 & 0 & -0.0625 & 0.0625 & -0.3750 & 0.1875 & 0.0625 & -0.0625 & -0.1875 \\
0.1563 & -0.1563 & -0.0625 & 0 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0.1563 & -0.0313 \\
-0.1563 & 0.1563 & 0.0625 & -0.1563 & 0 & 0.0625 & -0.0313 & 0.1563 & -0.1563 & 0.0313 \\
-0.0625 & 0.0625 & -0.3750 & -0.0625 & 0.0625 & 0 & 0.1875 & 0.0625 & -0.0625 & -0.1875 \\
0.0313 & -0.0313 & 0.1875 & 0.0313 & -0.0313 & 0.1875 & 0 & -0.0313 & 0.0313 & -0.4063 \\
-0.1563 & 0.1563 & 0.0625 & -0.1563 & 0.1563 & 0.0625 & -0.0313 & 0 & -0.1563 & 0.0313 \\
0.1563 & -0.1563 & -0.0625 & 0.1563 & -0.1563 & -0.0625 & 0.0313 & -0.1563 & 0 & -0.0313 \\
-0.0313 & 0.0313 & -0.1875 & -0.0313 & 0.0313 & -0.1875 & -0.406\,25 & 0.0313 & -0.0313 & 0
\end{bmatrix}$$

These results can be extended to higher values of p, and remain valid as long as the network is capable of recognition, that is, upto $p = N - 1$.

D.2 Possible implications for cognition

We have seen in the earlier chapters that the brain possesses the capability to orthogonalize information. Here, we have presented a numerical example of equivalence of the orthogonalization procedures. We could hence hypothesize that the brain is capable of performing more than one type of orthogonalization, and chooses the procedure based on the context and what information it wants to store. It would be interesting to study a network which can switch between different orthogonalization schemes and also examine how it can be implemented.

Appendix E

A comment on the orthogonalization schemes

E.1 Orthogonalization with an already orthogonal vector

The tables show the orthonormal bases following Gram-Schmidt (**G**), Symmetric (**S**) and Canonical (**C**) orthogonalization of an set of vectors **V** which contains a vector orthogonal to one or more of the other vectors. Table E.1 shows an example of sets with 3 and 4 vectors in which the last vector is already orthogonal to the rest. Table E.2 shows the orthonormal bases when sets of 4 vectors, in which the fourth vector is orthogonal to one or two of the previous three vectors. The presence of an orthogonal vector is reflected most clearly in the values of the Canonical basis.

TABLE E.1: Orthonormal bases for shown for two sets $\mathbf{V}$ with $p = 3$ and $p = 4$. The last vector in $\mathbf{V}$ is already orthogonal to the remaining vectors in the set.

$p = 3$										
V	−1	1	1	−1	−1	1	−1	1	−1	1
	1	−1	1	−1	1	−1	−1	−1	−1	1
	−1	−1	1	−1	−1	−1	−1	1	1	−1
G	−0.3162	0.3162	0.3162	−0.3162	−0.3162	0.3162	−0.3162	0.3162	−0.3162	0.3162
	0.3162	−0.3162	0.3162	−0.3162	0.3162	−0.3162	−0.3162	−0.3162	−0.3162	0.3162
	−0.2582	−0.3873	0.2582	−0.2582	−0.2582	−0.3873	−0.2582	0.2582	0.3873	−0.3873
S	−0.2887	0.3536	0.2887	−0.2887	−0.2887	0.3536	−0.2887	0.2887	−0.3536	0.3536
	0.3162	−0.3162	0.3162	−0.3162	0.3162	−0.3162	−0.3162	−0.3162	−0.3162	0.3162
	−0.2887	−0.3536	0.2887	−0.2887	−0.2887	−0.3536	−0.2887	0.2887	0.3536	−0.3536
C	0.4082	0	−0.4082	0.4082	0.4082	0	0.4082	−0.4082	0	0
	−0.3162	0.3162	−0.3162	0.3162	−0.3162	0.3162	0.3162	0.3162	0.3162	−0.3162
	0	0.5000	0	0	0	0.5000	0	0	−0.5000	0.5000
$v4$ orthogonal to $v1, v2$ and $v3$										
V	1	1	−1	−1	−1	−1	1	1	−1	1
	−1	1	−1	1	1	1	1	−1	1	1
	1	−1	−1	1	−1	−1	1	−1	1	1
	−1	1	1	1	−1	−1	1	−1	−1	−1
G	0.3162	0.3162	−0.3162	−0.3162	−0.3162	−0.3162	0.3162	0.3162	−0.3162	0.3162
	−0.2582	0.3873	−0.3873	0.2582	0.2582	0.2582	0.3873	−0.2582	0.2582	0.3873
	0.3333	−0.5000	−0.1667	0.3333	−0.3333	−0.3333	0.1667	−0.3333	0.3333	0.1667
	−0.3162	0.3162	0.3162	0.3162	−0.3162	−0.3162	0.3162	−0.3162	−0.3162	−0.3162
S	0.2488	0.4082	−0.3285	−0.3285	−0.2488	−0.2488	0.3285	0.3285	−0.3285	0.3285
	−0.3285	0.4082	−0.3285	0.2488	0.3285	0.3285	0.3285	−0.2488	0.2488	0.3285
	0.3285	−0.4082	−0.2488	0.3285	−0.3285	−0.3285	0.2488	−0.3285	0.3285	0.2488
	−0.3162	0.3162	0.3162	0.3162	−0.3162	−0.3162	0.3162	−0.3162	−0.3162	−0.3162
C	−0.4714	0	0.2357	0.2357	0.4714	0.4714	−0.2357	−0.2357	0.2357	−0.2357
	0	0	−0.4082	0.4082	0	0	0.4082	−0.4082	0.4082	0.4082
	−0.3162	0.3162	0.3162	0.3162	−0.3162	−0.3162	0.3162	−0.3162	−0.3162	−0.3162
	0.2357	−0.7071	0.2357	0.2357	−0.2357	−0.2357	−0.2357	−0.2357	0.2357	−0.2357

TABLE E.2: Orthonormal bases for sets of 4 vectors each containing an already orthogonal vector. Examples are shown with sets where a vector is already orthogonal to 1 or 2 other vectors.

$v4$ orthogonal to $v1$										
V	1	1	−1	−1	−1	−1	1	1	−1	1
	−1	1	−1	1	1	1	1	−1	1	−1
	1	1	−1	1	−1	−1	1	−1	1	1
	−1	1	1	1	−1	−1	1	−1	−1	−1
G	0.3162	0.3162	−0.3162	−0.3162	−0.3162	−0.3162	0.3162	0.3162	−0.3162	0.3162
	−0.2070	0.4830	−0.4830	0.2070	0.2070	0.2070	0.4830	−0.2070	0.2070	−0.2070
	0.3273	0	0	0.4364	−0.3273	−0.3273	0	−0.4364	0.4364	0.3273
	−0.3273	0.2182	0.4364	0.2182	−0.3273	−0.3273	0.2182	−0.2182	−0.4364	−0.3273
S	0.1772	0.3818	−0.3818	−0.3818	−0.1772	−0.1772	0.3818	0.3818	−0.3818	0.1772
	−0.3018	0.3696	−0.4271	0.1650	0.3594	0.3594	0.3696	−0.1650	0.2225	−0.3018
	0.3594	0.1650	−0.2225	0.3696	−0.3018	−0.3018	0.1650	−0.3696	0.4271	0.3594
	−0.3248	0.2673	0.3823	0.2673	−0.3248	−0.3248	0.2673	−0.2673	−0.3823	−0.3248
C	−0.4362	−0.1995	0.1995	0.1995	0.4362	0.4362	−0.1995	−0.1995	0.1995	−0.4362
	0.1543	−0.4629	0.1543	−0.4629	0.1543	0.1543	−0.4629	0.4629	−0.1543	0.1543
	−0.2887	0	0.5774	0	−0.2887	−0.2887	0	0	−0.5774	−0.2887
	0.2444	−0.3562	0.3562	0.3562	−0.2444	−0.2444	−0.3562	−0.3562	0.3562	0.2444
$v4$ orthogonal to $v1$ and $v3$										
V	1	1	−1	−1	−1	−1	1	1	−1	1
	−1	1	−1	1	1	1	1	−1	1	−1
	1	−1	−1	1	−1	−1	1	−1	1	1
	−1	1	1	1	−1	−1	1	−1	−1	−1
G	0.3162	0.3162	−0.3162	−0.3162	−0.3162	−0.3162	0.3162	0.3162	−0.3162	0.3162
	−0.2070	0.4830	−0.4830	0.2070	0.2070	0.2070	0.4830	−0.2070	0.2070	−0.2070
	0.2777	−0.4320	−0.2160	0.3703	−0.2777	−0.2777	0.2160	−0.3703	0.3703	0.2777
	−0.2722	0.2074	0.4278	0.2852	−0.3760	−0.3760	0.2204	−0.2852	−0.3630	−0.2722
S	0.2451	0.4448	−0.3949	−0.3180	−0.2220	−0.2220	0.3718	0.3180	−0.2949	0.2451
	−0.2451	0.3949	−0.4448	0.2220	0.3180	0.3180	0.3718	−0.2220	0.2949	−0.2451
	0.2949	−0.3679	−0.2718	0.3487	−0.2989	−0.2989	0.2758	−0.3487	0.3448	0.2949
	−0.2949	0.2718	0.3679	0.2989	−0.3487	−0.3487	0.2758	−0.2989	−0.3448	−0.2949
C	−0.4798	0.1405	0.1405	0.3393	0.3393	0.3393	0	−0.3393	0.1988	−0.4798
	0	0.1645	−0.1645	0.3971	−0.3971	−0.3971	0.5615	−0.3971	0	0
	−0.2527	0.4314	0.4314	−0.1787	−0.1787	−0.1787	0	0.1787	−0.6101	−0.2527
	0	−0.5745	0.5745	0.2380	−0.2380	−0.2380	−0.3366	−0.2380	0	0

Appendix F

Some preliminary results from a model with bounded synapses and fixed synaptic type

One of the criticisms of the Hopfield network is that the synapses can take arbitrarily large values. This issue can be addressed by restricting the range of values synaptic efficacies can take, a process which also takes the model closer to biological realism. It has been shown that incrementing each J_{ij} by a small amount during the learning process only reduces the memory capacity slightly [92]. The processes of *discretization* and *clipping* also have a similar effect on the memory capacity of the network [19]. Discretization refers to the mechanism which prohibits the weights from moving beyond a permitted set of discrete values, while clipping is a means of limiting all the synaptic efficacies to a certain range. This situation, where the weights are regulated to always lie within certain limits can be referred to as *learning within bounds* [19, 83, 84, 85]. The clipping procedure can be applied either after the presentation of a single pattern, in the Hopfield fashion, or following the presentation of a small group of patterns [86].

In this case, the efficiency of the network is better measured in terms of the memory lifetime rather than the number of patterns it can store without any loss of information. Memory lifetime refers to duration of the memorized information being present in the network and being distinguishable from the noise in system. It can in principle be prolonged by increasing the number of states between the limits. A state refers to an interval in the range of the synaptic weights bounded by the limits, with transition

between the states modifying the efficacies by a small amount (inversely proportional to the number of states). However, this improvement is not robust [22]. A better way of increasing memory lifetime is to alter how the limits are enforced.

Turning our attention to a network with bounded synapses [22, 86, 87], we try to refine it and improve its biological plausibility by characterizing each synapse to be either excitatory or inhibitory. While synapses get potentiated or depressed in the course of on-going plasticity,[1] they do not switch their character– excitatory synapses remain excitatory and inhibitory synapses remain inhibitory.

F.0.1 A model with bounded synapses

We take our cue from [22] and consider a network where synaptic efficacies lie within a certain small range, which we will henceforth refer to as "original model". Modifications to the efficacies can occur in the form of *hard* or *soft* bounds. Hard bounds refer to a fixed-step increase or decrease in the synaptic weights, while soft bounds result in modifications based on the current state of the synapse. These will be explained further later on. In this model, a memory can be tracked as a pattern of activities on the synapses, and the memory lifetime is a measure of how long the system retains a particular memory in the face of constant changes due to continuous plasticity.

There are two advantages to studying the activity of the synapses directly, without going into the details of how neuronal activity affects the synapses. The first is that we can check for the presence of a particular memory in the activity of the synapses. The other is that this method can be useful in evaluating the efficiency of ideally performing networks.

We consider a network of n synapses whose weights, w's, can lie in the range $[0, 1]$. Synaptic plasticity results in the potentiation or depression of the synapses, with the amount of modification dictated by the following equation:

$$w \to w + q_+(w) \text{ or } w \to w - q_-(w), \tag{F.1}$$

where w gives the synaptic strength and $q_+(w)$ and $q_+(w)$ are the amounts of potentiation and depression.

[1] Changes in the weights of the synapses due spontaneous activity and also due to the addition of other memories together constitute 'ongoing plasticity'.

We specify the rate, r of ongoing plasticity and estimate the potentiation and depression components constituting the ongoing plasticity. If f_+ and f_- are respectively the probabilities of potentiation and depression due to the plasticity (and $f_+ + f_- = 1$), then rf_+ and rf_- give the corresponding rates of potentiation and depression events.

Starting from the stage where the system is at equilibrium, the introduction of a memory would result in a perturbation of the equilibrium. The time required for the system to settle back into equilibrium gives the lifetime of that memory. During the course of this period, the signal due to the memory is clearly distinguishable from the background noise contributed by the ongoing plasticity. If $\bar{w}$ represents the average synaptic efficacy at equilibrium, then the signal due to a memory trace is given by

$$S = \frac{1}{n}\left(\sum_{i=pot}(w_i - \bar{w}) - \sum_{i=dep}(\bar{w} - w_i)\right). \tag{F.2}$$

The noise term due to ongoing plasticity can be estimated as the standard deviation of the signal.

The performance of the network is evaluated in terms of a memory lifetime τ, and an initial signal-to-noise ratio S_0/N_0. This ratio gives a measure of the versatility of the network in memorizing new information. The value of τ estimates the time during which the signal due to the memory gradually merges with the noise and is given by the time constant of slowest exponential component in convergence to equilibrium distribution.

The synapses in this network can be considered to be continuous variables whose values lie within specified limits. Let α denote the magnitude of a single plasticity event, that is, a single instance of potentiation or depression. Examining two aspects of plasticity, the magnitudes of $q_+(w)$ and $q_-(w)$ and their relation to synaptic strengths would provide insights on whether memory lifetimes can be improved by increasing the number of synaptic states or by modifications in the implementations of the limits.

Various classes of bounds

Hard boundary

Hard bounds refer to the case where potentiation and depression are varied by a fixed amount following a plasticity event, irrespective of the current state of the synapse. This constant is given by

$$q_+(w) = q_-(w) = \alpha, 0 < w < 1, \tag{F.3}$$

and weights that cross the bounds are curtailed.

Increasing the number of states now leads to an improvement in memory lifetimes, but this effect is observed only when potentiation and depression are present in equal amounts. Then, $\tau \approx \frac{1}{r\alpha^2}$, $\tau \propto \frac{1}{\alpha^2}$. Storing a memory in a network in which potentiation and depression are uniform and characterized by hard bounds is tantamount to an unbiased random walk, except at the boundaries. The return to equilibrium as the memory decays is a diffusion process and is invalid at the extremities. Deviating from a balance between the amounts of potentiation and depression results in the loss of enhancement of memory lifetimes if the number of states is higher.

The memory lifetime of the network with small potention step size is given by

$$\tau = \frac{1}{r\left(\sqrt{f_+} - \sqrt{f_-}\right)^2 + \alpha^2\pi^2 r\sqrt{f_+ f_-}}. \tag{F.4}$$

As the effects of potentiation and depression are equivalent, the results pertaining to the subset of synapses which are potentiated are equally valid for the subset of synapses undergoing depression. It is hence sufficient to study one of these two subsets.

Soft bounds

A more lenient way of implementing the bounds is to allow the current efficacy of a synapse dictate the amount of modification, i.e., the amounts of potentiation and depression, $q_+(w)$ and $q_-(w)$ depend on w and vanish at the limits. Such bounds can

be implemented in the following fashion:

$$q_+(w) = \alpha(1-w) \text{ and } q_-(w) = \alpha w. \tag{F.5}$$

In this case, the results are equally applicable to both the balanced and unbalanced cases, and $\tau \propto \frac{1}{\alpha}$.

More generally, soft bounds can be implemented in the form:

$$q_+(w) = \alpha(1-w)^\gamma \text{ and } q_-(w) = \alpha w^\gamma, \tag{F.6}$$

where γ is an odd positive integer.

In this case again $\tau \propto \frac{1}{\alpha}$. While higher the value of γ, longer the memory lifetime, at the same time, the initial signal-to-noise ratio decreases.

Optimal memory lifetimes can be achieved by using the following prescription for the implementation for the bounds:

$$q_+(w) = \frac{\alpha}{2}\left(1-(2w-1)^\gamma\right) \text{ and } q_-(w) = \frac{\alpha}{2}\left(1+(2w-1)^\gamma\right). \tag{F.7}$$

F.0.2 Refinement of the model

It is well-known that biological synapses are of two types, namely excitatory and inhibitory. This character of the synapse is decided by the nature of the neurons connected by the synapse. Synapses emanating from a firing neuron are in general all excitatory or all inhibitory. Though there can be many exceptions to this, it is true in general. The nature of the synapses does not change - a neuron may or may not fire, but synaptic character remains constant and is pre-decided. That is, whether a synapse is excitatory or inhibitory is not governed by the activity of the pre- and post-synaptic neuron. An excitatory synapse will connect two firing neurons, whereas an inhibitory synapse will connect a firing pre-synaptic neuron with a quiescent post-synaptic neuron.

This synaptic character has not been taken into consideration in the models discussed above, where the synapses are treated only as excitatory (since their weights acquire only positive values, between 0 and 1). We propose a refinement to the model by accounting for the fixed nature of a synapse. Recall that (a) there are two types of

synapses, namely excitatory or inhibitory and (b) in the course of on-going plasticity they do not switch their character, i.e while synapses get potentiated or depressed, excitatory synapses remain excitatory and inhibitory synapses remain inhibitory; they retain their nature and transitions between these two synaptic classes are generally not possible.

We consider a network ("our (modified) model" hereafter) of synapses with weights in the range of -0.5 to +0.5, with the weights on either side of the zero representing the two classes of synapses, namely excitatory synapses with positive efficacies and inhibitory synapses with negative weights. The synapses are then allowed to move among a number of states. The system is first allowed to settle down into an equilibrium, or a steady state. It is followed by the introduction of a memory. The system is then allowed to evolve in the presence of on-going plasticity, with the additional restriction of maintaining its type or 'nature', that is, the synapses in the negative interval are not permitted to cross over to the positive interval and vice-versa.

We find from preliminary studies that the initial value of the signal-to-noise ratio was higher for our model, as shown in Fig.F.1. This increase could indicate a greater probability of the signals due to newer memories being detected against the noise in the background. However, the memory lifetime in our model is reduced, as shown in Fig.F.2 . This result can be understood as an effect of the introduction of the synaptic type, which imposes an additional bound on the weights. Moreover, the separation of the synapses by type amounts to considering two sets or classes of synapses with half the number of states in each.

Further research along these lines is needed to establish more significant findings. But such studies are beyond the scope of this thesis and remain to be explored in future.

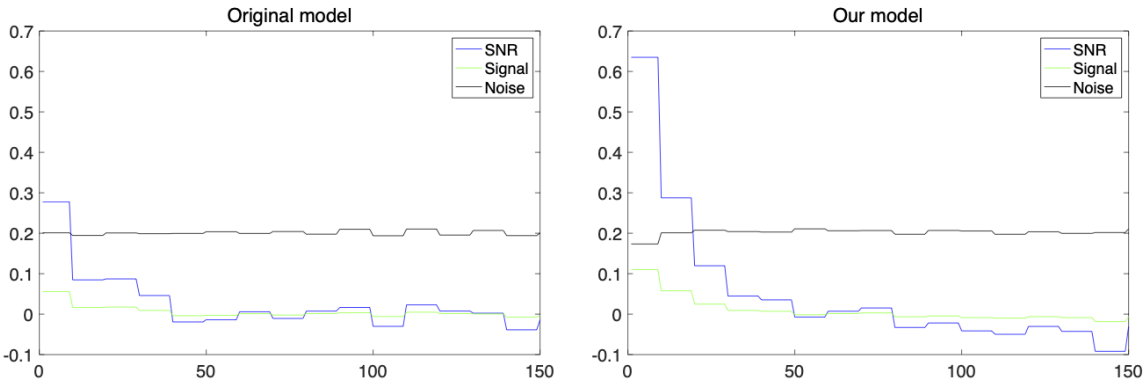


FIGURE F.1: Plot showing the memory lifetime as a function of time in the original model (in red) and our modified model (in black). The data pertains to a network with $n = 300$ synapses and $m = 8$ states and rate $r = 1/10$ (time steps). Soft bounds are implemented following eq.(F.7), with $\alpha = \frac{1}{m}$ and $\gamma = 3$. The initial value of SNR (S_0/N_0) is slightly higher in our model compared to the original model.

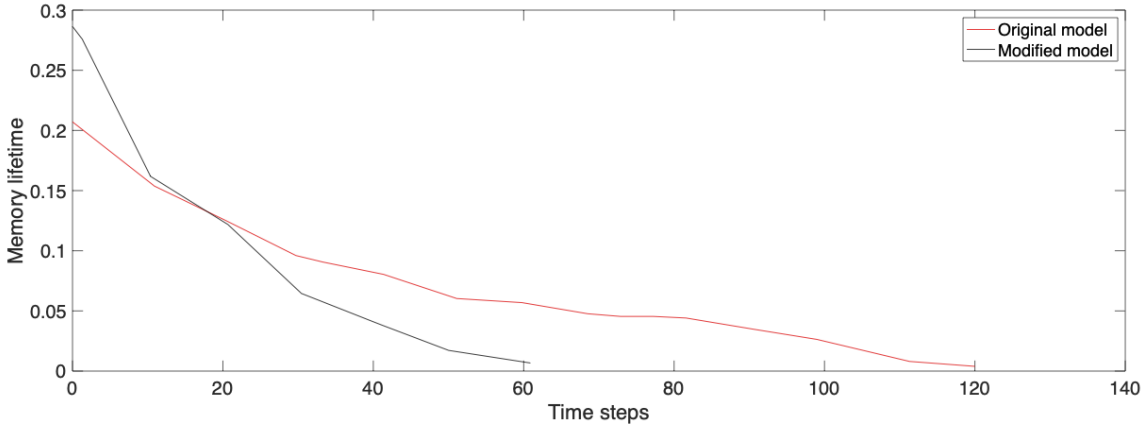


FIGURE F.2: Plot comparing the signal-to-noise ratio in the original model and our refined model. The data pertains to a network with $n = 100$ synapses and $m = 8$ states and rate $r = 1/10$ (time steps). The memory lifetime in our model is reduced to roughly half of that in the original model.

Appendix G

Effects of modifying the learning rule in the Willshaw model

The Hebb-Hopfield model deals with information in the form of dense patterns, however, information is better and more realistically represented by sparse patterns of activity. When the brain encounters some information, only a fraction of neurons fire or are active. This process can be implemented computationally through a model proposed by David Willshaw et al.[23]

G.1 The Willshaw model

The Willshaw model consists of a network of neurons connected by synapses which are all initially inactive (represented by 0's or $-\frac{1}{n}$'s). Information is presented to the network in the form of $N-$dimensional vectors or patterns of neural activities whose components are ± 1. When a pattern is presented to the network for memorization, a synapse can get activated (take value 1 or 0) if both the neurons connected by the synapse fire, or are active simultaneously. Once a synapse is active (excited), it remains in that state forever.

Hebbian learning in the model is implemented by the following learning rule[92]:

$$J_{ij} = \frac{1}{n}\Theta\left(\sum_{\mu=1}^{p}(S_i^\mu + 1)(S_j^\mu + 1)\right) - \frac{1}{n}, \tag{G.1}$$

where $\theta(x) = 1$ if $x > 0$ and 0 otherwise. S_i^μ and S_j^μ represent the activities of the pre- and post-synaptic neurons i and j which are connected by the synapse whose weight is given by J_{ij}. n is the number of neurons in the network. The synapses are all initially at $-\frac{1}{n}$. As more and more patterns are stored, more J_{ij}'s become 0. Unlike in the Hopfield model, here changes in the synaptic matrix are not cumulative. The network requires a large fraction of the efficacies to be zeroes in order to avoid saturation - the network thus deals with sparse information, with only a small fraction m of firing neurons in each pattern. The usage of sparse coding makes the network biologically more realistic.

The learning rule in eq.(G.1) gives rise to the following energy function:

$$E = -\frac{1}{2}\sum_{\substack{i,j=1 \\ i\neq j}}^{N} J_{ij}((S_i^\mu + 1)(S_j^\mu + 1)), \tag{G.2}$$

Retrieval is checked by presenting a pattern to the network and comparing the output pattern with the presented pattern for faithfulness. The prescription for retrieval can be stated as

$$h_i^\nu = sgn\left(J_{ij}(S_j^\nu + 1)\right) \tag{G.3}$$

One measure of the usefulness of a network as memory is its capacity. It can simply be defined as the number of patterns that can be stored in a network of size N. A more precise definition is the ratio of the mutual information between the stored and retrieved patterns to the number of synapses [93].

$$C = \frac{T(\vec{\xi}^{(1)}, \vec{\xi}^{(2)} \ldots \vec{\xi}^{(p)}; \vec{\xi}^{(1')}, \vec{\xi}^{(2')} \ldots \vec{\xi}^{(p')})}{n} \tag{G.4}$$

Alternately, we can analyze the performance of the network using the $0, 1$ binary coding and rewriting the learning rule in the familiar Hopfield fashion [9]:

$$J_{ij} = \mathbf{1}\left(\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu\right), \tag{G.5}$$

where

$$\mathbf{1}(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x \leq 0 \end{cases}$$

Calculating the local field potential on neuron i after the first pattern has been stored,

$$h_i^{(1)} = \sum_j 1 J_{ij} \xi_j^{(1)}. \tag{G.6}$$

We now arrive at an expression for the signal term from eq.(G.5) for $\mu = 1$:

$$h_i^{(1)} = Nm\xi_i^{(1)}, \tag{G.7}$$

as $J_{ij} = 1$ for all j's for $\xi_i^{(1)}\xi_j^{(1)} = 1$.

For a network of size N with m active neurons, we can now set up a threshold Θ whose maximum value is given by

$$\Theta = Nm. \tag{G.8}$$

The capacity of the network under this formulation is now given by

$$C = -\frac{log(1 - N^{--1/mN})}{m^2} \tag{G.9}$$

The memory capacity of the network increases as the patterns become more sparse, i.e., have more 0's in them. However, a decrease in the number of 1's in the patterns leads to a reduction in the amount of information present in them.

Limitations of the model and a new model

The learning rule in eq.(G.2), along with the sparseness of the input patterns together ensure that the levels of inhibition remain constant. (The effects of asymmetry in the network are also described in [88]). The network described above still remains fully connected. We propose here two modified learning rules where synaptic efficacies can now take one of three values : 0, a positive value and a negative value. One advantage of this advantage is that missing synapses characterized by 0's can be identified and distinguished from synapses which become active during the process of memorization. Moreover, this new formulation provides a means of varying the excitation and inhibition levels, keeping one or both constant.

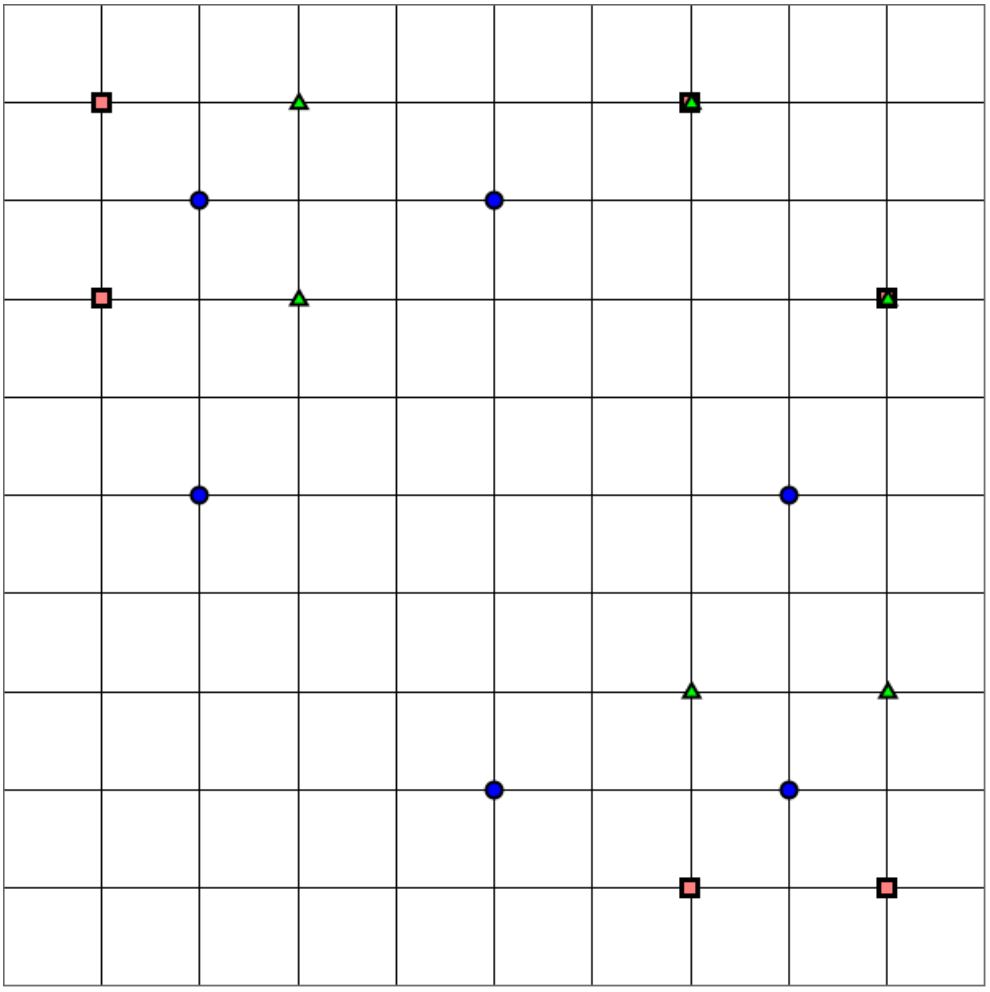


FIGURE G.1: Schematic representation of the synapstic matrix in the Willshaw model with $n = 10$. The (blue) circles show the synapses activated on storing the first pattern ($P1$). The (pink) squares and (green) triangles represent the synapses activated following the memorization of $P2$ and $P3$ respectively. Once activated, a synapse remains in that state forever, even if the same pair of neurons connected by the synapse fire simultaneously in a different pattern.

From Fig.G.1, we can see that once a pair of neurons fire simultaneously, they activate the synapse connecting them. This activated synapse retains that state permanently, even if the same pair of neurons are not simultaneously active in the other patterns being stored. Also, there is no difference in the active state of a synapse if the neurons connected by it fire together in more than one pattern.

The performance of the network is measured in terms of the fraction of retrieval. Retrieval is checked by whether presenting the stored patterns back to the network leads to recovery of the presented patterns.

G.2 Modifications to the learning rule

While the model takes into account the sparseness of the input patterns, the network still remains fully connected. It would be interesting to study the network with different degrees of dilution. Moreover, the symmetry of synaptic weights is biologically inaccurate, and the synapses in the model can be modified to have asymmetric weights. Both dilution and asymmetry can be introduced in the model by assigning positive weights to active synapses, and negative weights to the inactive ones. Under this new scheme, a '0' would represent a missing synapse.

The introduction of the new classification of weights also provides a tool for studying the effects of varying levels of excitation and inhibition on network performance. Their effects on the network with different levels of sparseness can also be studied.

We first examine the case where inhibition is kept constant while varying the excitation levels. The modified learning rule is now given by:

$$J_{ij} = \frac{c}{N}\sum_{\mu=1}^{p}\theta((S_i^\mu + 1)(S_j^\mu + 1)) - \frac{1}{N}, c \in [1, 2] \tag{G.10}$$

In the second scenario, we modify the learning rule such that both inhibition and excitation levels vary. This learning prescription follows the following equation:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\theta((S_i^\mu + 1)(S_j^\mu + 1)) - \frac{c}{N}, c \in (0, 1] \tag{G.11}$$

The excitation-inhibition ratios for these two modified learning rules are plotted in Figs.G.2 and G.3.

G.3 Results of the modified learning rules

Our preliminary study involved simulations with network size $N = 100$ and $N = 1000$. Patterns were generated randomly, with sets of patterns with a uniform number of firing patterns for each value of the sparseness parameter (f). The network was later diluted and a certain number of synapses corresponding to the level of dilution(D) were removed randomly. The dilution was also uniform across the network, i.e., the same number of synapses emanating from each neuron was removed.

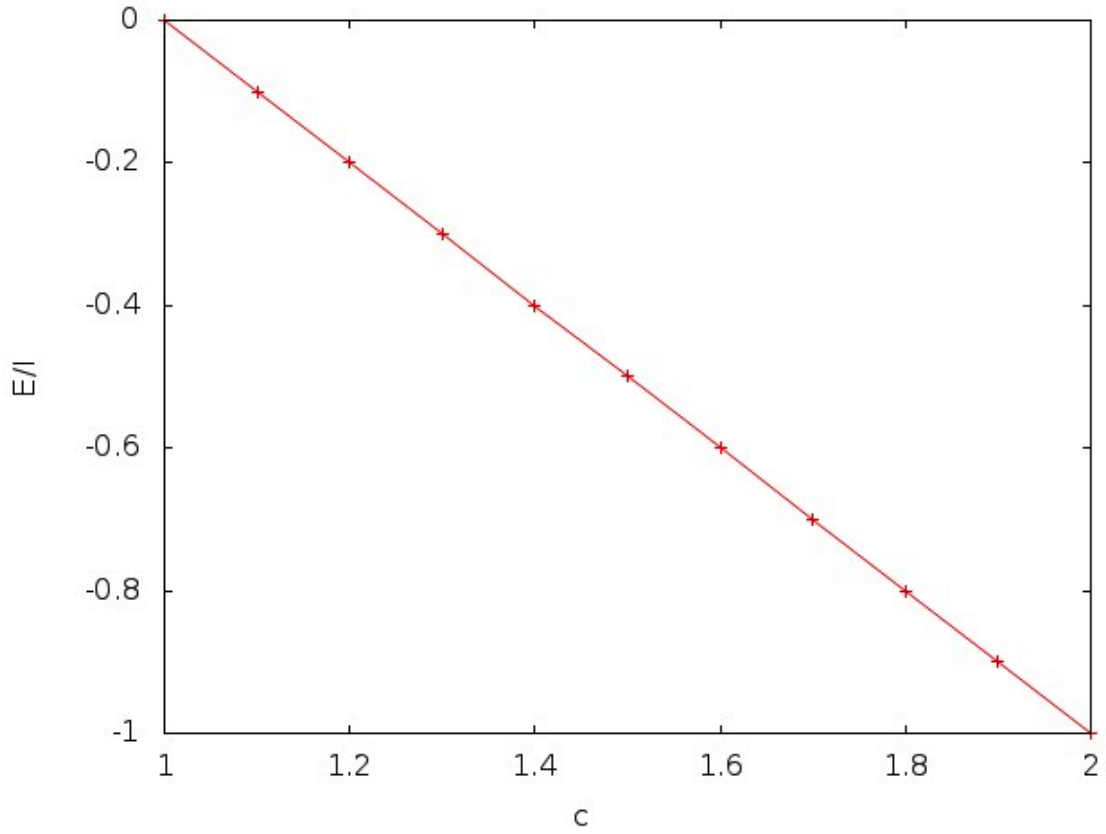


FIGURE G.2: Figure showing the ratio of excitation(E) and inhibition(I) for various values of c in the Willshaw model following the modified learning rule (G.10). Excitation levels vary, while inhibition is at a constant level, resulting in a linear variation in the E/I ratio.

This process of dilution automatically introduces asymmetry into the system. The weights are no longer symmetric, as a synapse from neuron 'i' to 'j' might be present, whereas the synapse from 'j' to 'i' may be missing. The effects of the modified learning rules were then studied.

G.3.1 Case (i)

When patterns were stored in a network of size $N = 1000$ following eq. (G.10), there was complete retrieval upto $f = 0.001$ for $c = 1 - 1.1$. Beyond that, there is a slight deterioration in retrieval quality for $c = 1.2 - 1.3$, with retrieval upto $f = 0.005$. The quality of retrieval degrades further for higher values of c ($c = 1.4, 1.5 \ldots 2$), with patterns being retrieved only when $f < 0.005$.

In a network of size $N = 100$, patterns were retrieved upto $f = 0.1$ for $c = 1, 1.1$. For $c = 1.2, 1.3$, retrieval went down beyond $f = 0.05$. With further increase in c,

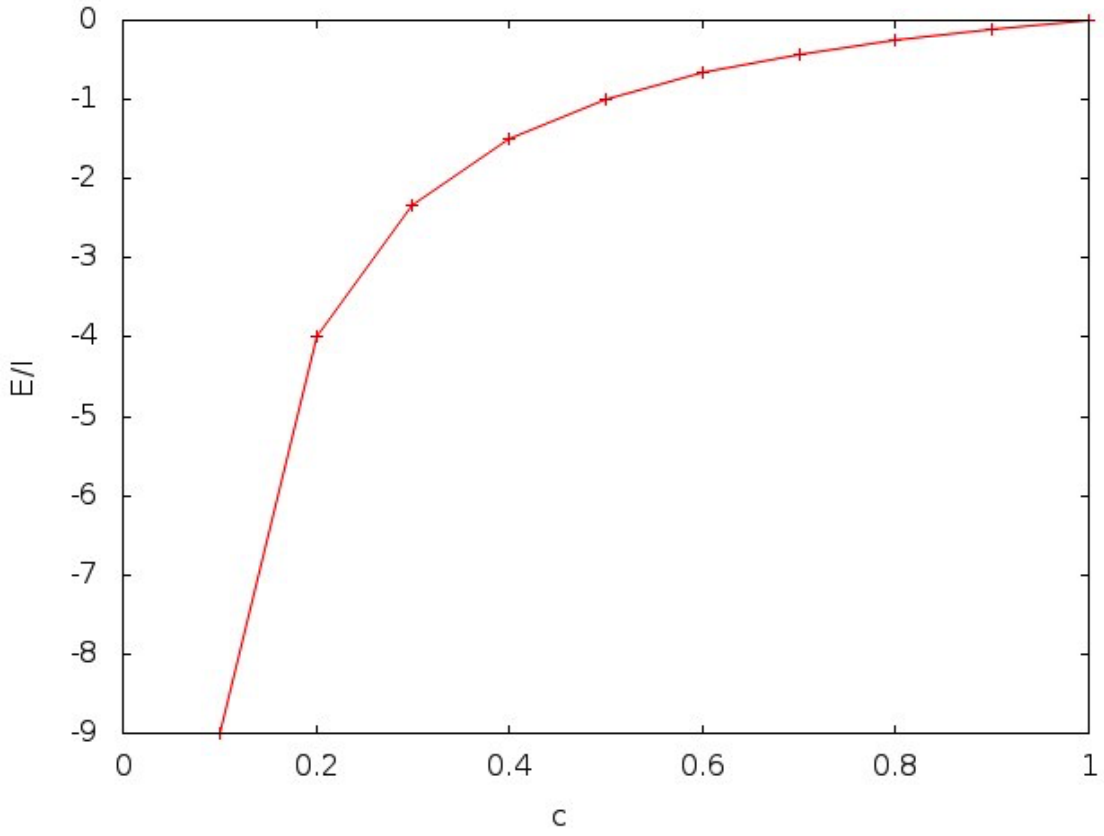


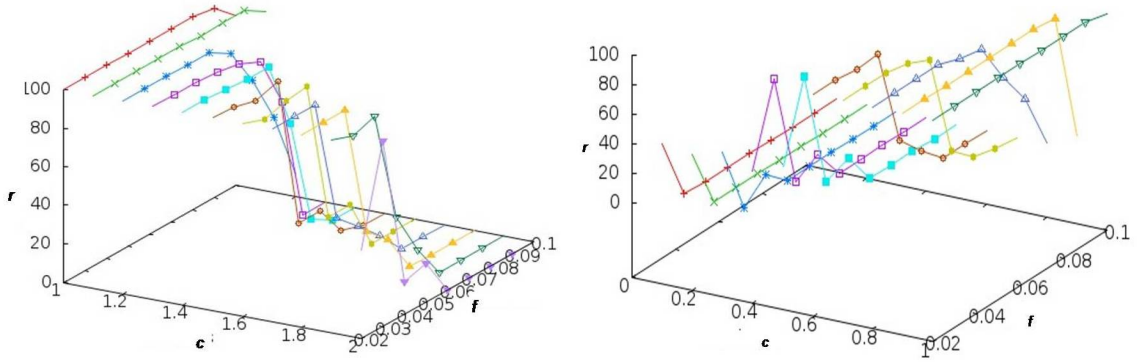
FIGURE G.3: Figure showing the ratio of excitation(E) and inhibition(I) for various values of c in the Willshaw model following the modified learning rule (G.11). Both excitation and inhibition levels vary, resulting in the variation in the E/I ratio following a curve as shown above.

the value of f upto which there was complete retrieval was as low as $f = 0.02$. This can be seen from Fig.G.4a.

G.3.2 Case (ii)

Learning in a network ($N = 1000$) following eq.(G.11) results in retrieval upto $f = 0.07$ for $c = 0.8 - 0.9$. For $c = 0.6 - 0.7$, retrieval quality deteriorates, and patterns are retrieved only upto $f = 0.03$. However, at lower values of c ($c < 0.6$), retrieval quality degrades even further, as can be seen from Fig.G.4b

For $N = 1000$, there was complete retrieval upto $f = 0.009$ for $c = 0.9$. Patterns were retrieved completely upto $f = 0.005$ for $c = 0.7$ and upto $f = 0.006$ for $c = 0.8$. For even values of c ($c = 0.1 - 0.6$), hardly any patterns were retrieved completely.

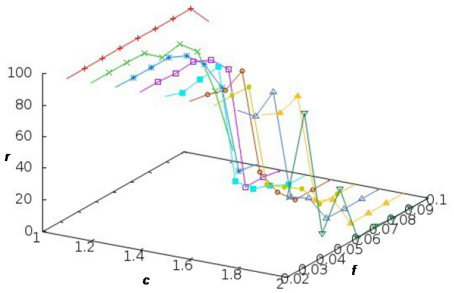


(a) With modified learning rule (i) (b) With modified learning rule (ii)

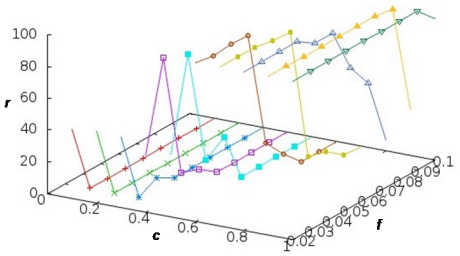
FIGURE G.4: Figure showing the results of the Willshaw model with modified learning rules (i) (Fig.a) and (ii) (Fig.b) for various values of the sparseness parameter f when $p = 100$ patterns are stored in a network of size $N = 100$. r gives the number of patterns retrieved while c is a constant whose range is specified by the learning rule. The network can efficiently retrieve the stored patterns when c remains close to 1 ($c = 1$ is the default Willshaw model).

G.3.3 Effects of dilution

As we can see from Fig.G.5, the retrieval is not significantly affected at low levels of dilution, when 1% or 10% of the synapses were removed at random. However, there is a drastic decrease in network performance when the dilution is 25%.

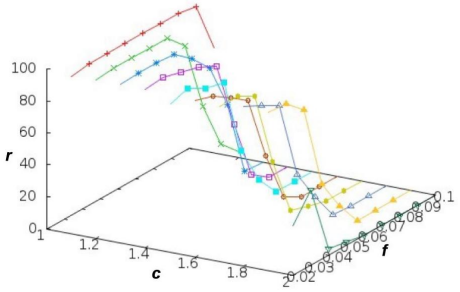


(a) With modified learning rule (*i*)

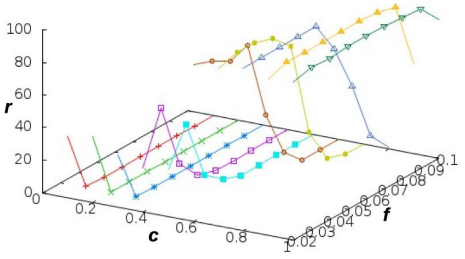


(b) With modified learning rule (*ii*)

At 1% dilution

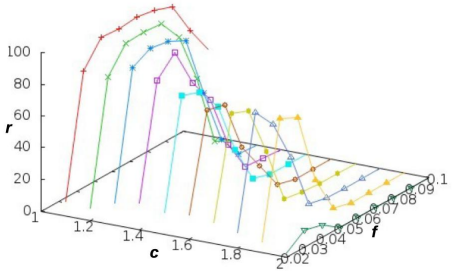


(c) With modified learning rule (*i*)

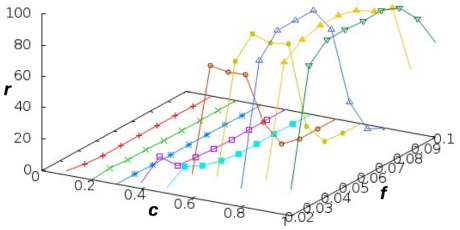


(d) With modified learning rule (*ii*)

At 10% dilution



(e) With modified learning rule (*i*)



(f) With modified learning rule (*ii*)

At 25% dilution

FIGURE G.5: Results from the Willshaw model with modified learning rules with different levels of dilution and for different values of sparseness. The performance of the network is not significantly impeded at $D = 1\%$ or 10%, but worsens at $D = 25\%$.

G.4 Summing up

The model thus addresses some of the drawbacks of the Hopfield network, including its complete connectivity, dense patterns and the symmetry of the efficacies ($J_{ij} = J_{ji}$). The network in our study deals with sparse patterns, and the modified learning rules address the issues of connectivity and symmetry.

We have presented some exploratory results here, but a deeper study of the network remains beyond the scope of this thesis.

Bibliography

[1] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79: 2554–2558, 1982.

[2] D.O.Hebb. *Organization of behaviour*. Wiley, New York, 1949.

[3] Vipin Srivastava. A unified view of the orthogonalization methods. *Journal of Physics A: Mathematical and General*, 33(35):6219, 2000.

[4] Per-Olov Löwdin. A box model of alkali halide crystal. *Ark. Mat. Astr. Fys. A*, 35:30, 1947.

[5] Per-Olov Löwdin. Quantum theory of many-particle systems. i. physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configurational interaction. *Physical Review*, 97(6): 1474, 1955.

[6] Per-Olov Löwdin. Quantum theory of cohesive properties of solids. *Advances in Physics*, 5(17):1–171, 1956.

[7] Daniel J. Amit. *Modeling Brain Function : the World of Attractor Neural Networks*. Cambridge University Press, New York, NY, USA, 1989. ISBN 0-521-36100-1.

[8] Yaneer Bar-Yam. *Dynamics of complex systems*, volume 213. Addison-Wesley Reading, MA, 1997.

[9] Pierre Peretto. *An introduction to the modeling of neural networks*, volume 2. Cambridge University Press, 1992.

[10] Shun-Ichi Amari and Kenjiro Maginu. Statistical neurodynamics of associative memory. *Neural Networks*, 1(1):63–73, 1988.

[11] Wickliffe C Abraham and Anthony Robins. Memory retention–the synaptic stability versus plasticity dilemma. *Trends in neurosciences*, 28(2):73–78, 2005.

[12] Tim VP Bliss and Graham L Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407):31–39, 1993.

[13] Elie L Bienenstock, Leon N Cooper, and Paul W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32–48, 1982.

[14] Leon N Cooper. A possible organization of animal memory and learning. In *How We Learn; How We Remember: Toward An Understanding Of Brain And Neural Systems: Selected Papers of Leon N Cooper*, pages 13–25. World Scientific, 1995.

[15] V Srivastava and Samuel F. Edwards. A model of how the brain discriminates and categorises. *Physica A: Statistical Mechanics and its Applications*, 276:352–358, 2000.

[16] Vipin Srivastava, DJ Parker, and SF Edwards. The nervous system might 'orthogonalize' to discriminate. *Journal of theoretical biology*, 253(3):514–517, 2008.

[17] Vipin Srivastava, Suchitra Sampath, and David J Parker. Overcoming catastrophic interference in connectionist networks using gram-schmidt orthogonalization. *PLoS ONE*, 9(9):e105619, 2014.

[18] Suchitra Sampath and Vipin Srivastava. On stability and associative recall of memories in attractor neural networks. *PloS one*, 15(9):e0238054, 2020.

[19] John A Hertz, Anders S Krogh, and Richard G Palmer. *Introduction to the theory of neural computation*, volume 1. Basic Books, 1991.

[20] Raul Rojas. *Neural Networks: A Systematic Introduction*. Springer Science & Business Media, 1996.

[21] HC Schweinler and Eugene P. Wigner. Orthogonalization methods. *Journal of Mathematical Physics*, 11(5):1693–1694, 1970.

[22] Stefano Fusi and LF Abbott. Limits on the memory storage capacity of bounded synapses. *Nature neuroscience*, 10(4):485–493, 2007.

[23] David J Willshaw, O Peter Buneman, and Hugh Christopher Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960, 1969.

[24] V Srivastava and Samuel F. Edwards. *The brain and orthonormal bases*. Allied, Mumbai, 2009.

[25] Gary Fiskum. Mitochondrial participation in ischemic and traumatic neural cell death. *Journal of neurotrauma*, 17(10):843–855, 2000.

[26] Karim Nader, Glenn E Schafe, and Joseph E Le Doux. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406 (6797):722–726, 2000.

[27] Cyrinne Ben Mamou, Karine Gamache, and Karim Nader. Nmda receptors are critical for unleashing consolidated auditory fear memories. *Nature neuroscience*, 9(10):1237–1239, 2006.

[28] Jean M Barnes and Benton J Underwood. " fate" of first-list associations in transfer theory. *Journal of experimental psychology*, 58(2):97, 1959.

[29] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

[30] Reza Shadmehr and Thomas Brashers-Krug. Functional stages in the formation of human long-term motor memory. *Journal of Neuroscience*, 17(1):409–419, 1997.

[31] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[32] Denis Mareschal, Paul C Quinn, and Robert M French. Asymmetric interference in 3-to 4-month-olds' sequential category learning. *Cognitive Science*, 26(3):377–389, 2002.

[33] Stefano Fusi, Patrick J Drew, and Larry F Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005.

[34] Stefano Fusi and Walter Senn. Eluding oblivion with smart stochastic selection of synaptic updates. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(2):026112, 2006.

[35] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[36] Gerd Kempermann. The neurogenic reserve hypothesis: what is adult hippocampal neurogenesis good for? *Trends in neurosciences*, 31(4):163–169, 2008.

[37] Rafal Bogacz and Malcolm W Brown. An anti-hebbian model of familiarity discrimination in the perirhinal cortex. *Neurocomputing*, 52:1–6, 2003.

[38] Mark CW Van Rossum, Maria Shippi, and Adam B Barrett. Soft-bound synaptic plasticity increases storage capacity. *PLoS Comput Biol*, 8(12):e1002836, 2012.

[39] Wickliffe C Abraham and Mark F Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends in neurosciences*, 19(4):126–130, 1996.

[40] Mark F Bear. Bidirectional synaptic plasticity: from theory to reality. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1432):649–655, 2003.

[41] Gina Turrigiano. Homeostatic signaling: the positive side of negative feedback. *Current opinion in neurobiology*, 17(3):318–324, 2007.

[42] Vipin Srivastava, Meena Vipin, and Enzo Granato. Recall of old and recent information. *Network: Computation in Neural Systems*, 9(2):159–166, 1998.

[43] John K Kruschke. Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, 99(1):22, 1992.

[44] Ken McRae and Phil A Hetherington. Catastrophic interference is eliminated in pretrained networks. In *Proceedings of the 15h Annual Conference of the Cognitive Science Society*, pages 723–728, 1993.

[45] Makoto Yamaguchi. Reassessment of catastrophic interference. *Neuroreport*, 15 (15):2423–2426, 2004.

[46] Stephan Lewandowsky and Shu-Chen Li. Catastrophic interference in neural networks: Causes, solutions, and data. In *Interference and inhibition in cognition*, pages 329–361. Elsevier, 1995.

[47] Robert M French. Selective memory loss in aphasics: An insight from pseudo-recurrent connectionist networks. In *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, pages 183–195. Springer, 1998.

[48] David Marr. A theory of cerebellar cortex. *The Journal of Physiology*, 202(2): 437, 1969.

[49] Yoram Baram. Orthogonal patterns in binary neural networks. techreport NASA-TM-100060, A-88068, NAS 1.15:100060, (NASA Ames Research Center, Moffett Field, CA, United States), 1988.

[50] 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.

[51] Randall D Beer. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509, 1995.

[52] Francesco Donnarumma, Roberto Prevete, and Giuseppe Trautteur. Programming in the brain: a neural network theoretical framework. *Connection Science*, 24(2-3):71–90, 2012.

[53] Vipin Srivastava and Suchitra Sampath. Could the brain function mathematically? *Neurology and Neuroscience Research*, 1(1):4, 2018. doi: 10.24983/scitemed.nnr.2018.00064.

[54] John F Guzowski, James J Knierim, and Edvard I Moser. Ensemble dynamics of hippocampal regions ca3 and ca1. *Neuron*, 44(4):581–584, 2004.

[55] Michael A Yassa and Craig EL Stark. Pattern separation in the hippocampus. *Trends in neurosciences*, 34(10):515–525, 2011.

[56] Joshua P Neunuebel and James J Knierim. Ca3 retrieves coherent representations from degraded input: direct evidence for ca3 pattern completion and dentate gyrus pattern separation. *Neuron*, 81(2):416–427, 2014.

[57] Talya Sadeh, Jason D Ozubko, Gordon Winocur, and Morris Moscovitch. How we forget may depend on how we remember. *Trends in cognitive sciences*, 18(1): 26–36, 2014.

[58] George Mandler. Recognizing: The judgment of previous occurrence. *Psychological review*, 87(3):252, 1980.

[59] Larry L Jacoby. A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of memory and language*, 30(5):513–541, 1991.

[60] Wayne Donaldson. The role of decision processes in remembering and knowing. *Memory & Cognition*, 24(4):523–533, 1996.

[61] William E Hockley and Angela Consoli. Familiarity and recollection in item and associative recognition. *Memory & Cognition*, 27(4):657–664, 1999.

[62] J. C. Dunn. Dual-state models of the remember/know paradigm. In *32nd European Mathematical Psychology Group Meeting, Lisbon, Portugal.*, 2001.

[63] Larry R Squire and John T Wixted. The cognitive neuroscience of human memory since hm. *Annual review of neuroscience*, 34:259–288, 2011.

[64] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[65] Amos J Storkey and Romain Valabregue. The basins of attraction of a new hopfield learning rule. *Neural Networks*, 12(6):869–876, 1999.

[66] Edmund T Rolls. An attractor network in the hippocampus: theory and neurophysiology. *Learning & Memory*, 14(11):714–731, 2007.

[67] César Rennó-Costa, John E Lisman, and Paul FMJ Verschure. A signature of attractor dynamics in the ca3 region of the hippocampus. *PLoS computational biology*, 10(5):e1003641, 2014.

[68] Rafal Bogacz, Malcolm W Brown, and Christophe Giraud-Carrier. Model of familiarity discrimination in the perirhinal cortex. *Journal of computational neuroscience*, 10(1):5–23, 2001.

[69] Rafal Bogacz and Malcolm W Brown. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, 13(4):494–524, 2003.

[70] Endel Tulving. Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American Scientist*, 62(1):74–82, 1974.

[71] Nicolas Brunel. Course 10 - network models of memory. In C.C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, editors, *Methods and Models in Neurophysics*, volume 80 of *Les Houches*, pages 407 – 476. Elsevier, 2005. doi: https://doi.org/10.1016/S0924-8099(05)80016-2. URL http://www.sciencedirect.com/science/article/pii/S0924809905800162.

[72] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, page 201611835, 2017.

[73] Nicolas Y Masse, Gregory D Grant, and David J Freedman. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *arXiv preprint arXiv:1802.01569*, 2018.

[74] Ulises Pereira and Nicolas Brunel. Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron*, 2018.

[75] Frederic Charles Bartlett and Frederic C Bartlett. *Remembering: A study in experimental and social psychology*, volume 14. Cambridge University Press, 1995.

[76] Sam McKenzie and Howard Eichenbaum. Consolidation and reconsolidation: two lives of memories? *Neuron*, 71(2):224–233, 2011.

[77] Vipin Srivastava and Suchitra Sampath. Cognition of learning and memory: What have löwdin's orthogonalizations got to do with that? In *Advances in Quantum Chemistry*, volume 74, pages 299–319. Elsevier, 2017.

[78] Per-Olov Löwdin. On the nonorthogonality problem. *Advances in quantum chemistry*, 5:185, 1970.

[79] Nancy Kanwisher. What's in a face? *Science(Washington)*, 311(5761):617–618, 2006.

[80] Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311 (5761):670–674, 2006.

[81] DR Inglis. Non-orthogonal wave functions and ferromagnetism. *Physical Review*, 46(2):135, 1934.

[82] John C Slater. Cohesion in monovalent metals. *Physical Review*, 35(5):509, 1930.

[83] Giorgio Parisi. A memory which forgets. *Journal of Physics A: Mathematical and General*, 19(10):L617, 1986.

[84] JP Nadal, G Toulouse, JP Changeux, and S Dehaene. Networks of formal neurons and memory palimpsests. *EPL (Europhysics Letters)*, 1(10):535, 1986.

[85] Mirta B Gordon. Memory capacity of neural networks learning within bounds. *Journal de Physique*, 48(12):2053–2058, 1987.

[86] Daniel J Amit and Stefano Fusi. Constraints on learning in dynamic synapses. *Network: Computation in Neural Systems*, 3(4):443–464, 1992.

[87] Daniel J Amit and Stefano Fusi. Learning in neural networks with material synapses. *Neural Computation*, 6(5):957–982, 1994.

[88] Haim Sompolinsky. Statistical mechanics of neural networks. *Physics Today*, 41 (21):70–80, 1988.

[89] Anupam Hazra, Feng Gu, Ahmad Aulakh, Casey Berridge, Jason L Eriksen, and Jokūbas Žiburkus. Inhibitory neuron and hippocampal circuit dysfunction in an aged mouse model of alzheimer's disease. *PloS one*, 8(5):e64318, 2013.

[90] Miguel Maravall. Sparsification from dilute connectivity in a neural network model of memory. *Network: Computation in Neural Systems*, 10(1):15–39, 1999.

[91] Fan Zhang and Xinhong Zhang. The average radius of attraction basin of hopfield neural networks. In *International Symposium on Neural Networks*, pages 253–258. Springer, 2008.

[92] Haim Sompolinsky. The theory of neural networks: The hebb rule and beyond. In *Heidelberg colloquium on glassy dynamics*, pages 485–527. Springer, 1987.

[93] Andreas Knoblauch, Günther Palm, and Friedrich T Sommer. Memory capacities for synaptic and structural plasticity. *Neural Computation*, 22(2):289–341, 2010.

OPEN ACCESS Freely available online

PLOS ONE

Overcoming Catastrophic Interference in Connectionist Networks Using Gram-Schmidt Orthogonalization

Vipin Srivastava[1,2]*, Suchitra Sampath[2], David J. Parker[3]

1 School of Physics, University of Hyderabad, Hyderabad, India, **2** Centre for Neural and Cognitive Sciences, University of Hyderabad, Hyderabad, India, **3** Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, United Kingdom

Abstract

Connectionist models of memory storage have been studied for many years, and aim to provide insight into potential mechanisms of memory storage by the brain. A problem faced by these systems is that as the number of items to be stored increases across a finite set of neurons/synapses, the cumulative changes in synaptic weight eventually lead to a sudden and dramatic loss of the stored information (catastrophic interference, CI) as the previous changes in synaptic weight are effectively lost. This effect does not occur in the brain, where information loss is gradual. Various attempts have been made to overcome the effects of CI, but these generally use schemes that impose restrictions on the system or its inputs rather than allowing the system to intrinsically cope with increasing storage demands. We show here that catastrophic interference occurs as a result of interference among patterns that lead to catastrophic effects when the number of patterns stored exceeds a critical limit. However, when Gram-Schmidt orthogonalization is combined with the Hebb-Hopfield model, the model attains the ability to eliminate CI. This approach differs from previous orthogonalisation schemes used in connectionist networks which essentially reflect sparse coding of the input. Here CI is avoided in a network of a fixed size without setting limits on the rate or number of patterns encoded, and without separating encoding and retrieval, thus offering the advantage of allowing associations between incoming and stored patterns. PACS Nos.: 87.10.+e, 87.18.Bb, 87.18.Sn, 87.19.La

Citation: Srivastava V, Sampath S, Parker DJ (2014) Overcoming Catastrophic Interference in Connectionist Networks Using Gram-Schmidt Orthogonalization. PLoS ONE 9(9): e105619. doi:10.1371/journal.pone.0105619

Editor: Manabu Sakakibara, Tokai University, Japan

Received November 12, 2013; **Accepted** July 26, 2014; **Published** September 2, 2014

Funding: The Royal Society; The Leverhulme Foundation (UK); National Initiative of Research in Cognitive Science by the Department of Science and Technology, Government of India. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: vipinsri02@gmail.com

Introduction

Nervous systems have two basic requirements: they must be stable and thus able to generate reliable specific outputs, while at the same time they must be flexible to allow the output to change during development or as a result of experience. This is the "stability-plasticity dilemma" [1], and it is a concern to both neurobiologists who want to understand how nervous systems cope with constantly changing internal and external conditions, and those working on artificial neural networks. While not exclusively related to it, this problem is often considered in relation to memory. The analysis of memory systems has been a major focus of neuroscience research, but there are still many unanswered questions that need to be addressed at both the experimental and theoretical levels. In terms of the stability-plasticity problem, the question is how a system can store new input patterns across shared components without disturbing previously stored information in those components.

One of the first considerations of this problem was highlighted by Bienenstock, Cooper and Munro [2], who suggested that long-term potentiation (LTP), a proposed mechanism for learning and memory [3], could suffer from an inherent instability (the BCM model). They suggested that in systems with a set threshold for plasticity the potentiation of a synapse by a particular input that exceeded the threshold could leave that synapse open to further potentiation when another, non-salient, input was presented (this has also been referred to as the "ongoing plasticity" problem; see [4]). Due to the initial potentiation of the synapse, non-salient or random inputs caused by a non-stationary environment could exceed the threshold for plasticity, resulting in the potential for run-away cycles of potentiation which would alter the synaptic changes associated with the original memory. This would effectively overwrite the original memory, and in biological systems if left unchecked, excessive activation could also lead to epileptogenic or excitotoxic damage and cell death [5]. The opposite effect could occur with long-term depression, where a synapse is weakened when the input falls below a depression threshold: in this case there could be a positive feedback loop that results in the successive depression of the synapse.

While the exact relationship is not clear, a similar effect may occur in artificial neural networks. When the number of sequentially recorded/stored patterns exceeds a critical value there is a sudden and complete loss of previously stored inputs [6]. This example of retroactive interference is called catastrophic interference (CI) and is caused by the sharing of connections whose weights are changed by the presentation of specific inputs. As more patterns are stored the weights are changed and beyond a critical point new inputs erase the memory of previous inputs. If the memories happen to be overlapping, or correlated, which essentially means that several of their elements are similar (the mathematical meaning is explained in [7], [8]), then a particular

CHAPTER FOURTEEN

Cognition of Learning and Memory: What Have Löwdin's Orthogonalizations Got to Do With That?

Vipin Srivastava[*,†,1], **Suchitra Sampath**[*]
[*]Centre for Neural and Cognitive Sciences, University of Hyderabad, Hyderabad, Telangana, India
[†]School of Physics, University of Hyderabad, Hyderabad, Telangana, India
[1]Corresponding author: e-mail address: vipinsri02@gmail.com

Contents

Abstract

We present some initial results to show that Löwdin's two orthogonalization schemes, namely Symmetric and Canonical, can help us to understand certain important aspects of the brain's competence to learn and memorize. We propose that these orthogonalizations may constitute the physiological actions that the brain may perform to deal with certain types of memories.

1. INTRODUCTION

Converting a given set of linearly independent vectors into a set of mutually orthogonal vectors is an old problem, which has been studied at length in mathematics, physics, and chemistry. The orthonormal basis sets find wide ranging applications in all three disciplines. That the idea and schemes of orthogonalization could have bearing on understanding cognition

Advances in Quantum Chemistry, Volume 74
ISSN 0065-3276
http://dx.doi.org/10.1016/bs.aiq.2016.08.001

Neurology and Neuroscience Research

IDEA AND INNOVATION

Could the Brain Function Mathematically?

Vipin Srivastava, PhD[1]*; Suchitra Sampath, MSc[2]

[1] School of Physics, University of Hyderabad, Hyderabad, India

[2] Centre for Neural and Cognitive Sciences, University of Hyderabad, Hyderabad, India

Abstract

We have put forth a hypothesis that the brain bears the innate capability of performing high-level mathematical computing in order to perform certain cognitive tasks. We give examples of Orthogonalization and Fourier transformation and argue that the former may correspond to the physiological action the brain performs to compare incoming information and put them in categories, while the latter could be responsible for the holographic nature of the long-term memory, which is known to withstand trauma. We plead that this proposal may not be as strange as it may appear, and argue how this line of mathematical modeling can have far-reaching consequences.

Preamble

How the brain processes information, where and how it stores them, and how it retrieves from memory as and when required, are some of the basic questions one is naturally curious about. In spite of neuroscience being an old discipline and the brain having been mapped extensively, one hardly knows much about the physiological mechanisms underlying such basic functions of the brain involving learning and memory. One has begun to develop some understanding on this account in the past few decades due to the efforts by psychologists (e.g. Donald Hebb [1]) and formal approaches by mathematicians, physicists, engineers, and cognitive scientists employing mathematical and cognitive models for certain brain functions. The theoretical approach not only gives crucial insight into how the brain functions, but also helps in designing and planning experiments that would otherwise be difficult and expensive, and in devising ways of processing and storing non-cognitive information. The latter may pertain to information technology. It is our contention that the mathematical and cognitive models of brain processes should give ideas to construct algorithms to handle numerous non-cognitive problems.

A Hypothesis

In this short communication, we summarize one such theoretical approach we have pursued for some time. We have hypothesized since 2000 [2] that in many situations the brain might be functioning in a mathematical manner in that it might be using mathematical functions and transformations (which are otherwise well known to mathematicians and physicists) to perform certain cognitive tasks. A natural question that will arise then is "how would an untrained brain know about these functions and transformations?" To this end, we go on to conjecture that the brain might be hard-wired to do such mathematical functions and transformations, and that these competencies might have been acquired by the brain in the course of evolution while mathematicians and physicists have been only reinventing them. Apparently, a section of modern philosophers also believes so.

The Crux

Let us start with the basic question: how do we learn? — On the basis of certain experimental observations, a psychologist Donald Hebb [1] put forth a hypothesis that the synaptic efficacies, i.e. the nature (whether excitatory or inhibitory) and strength of synaptic connections between numerous neurons in the brain, change as and when an information is registered. The synapses have plastic character, i.e. the modification in their efficacies stay, sometimes for short durations and sometimes over longer periods, and it is through this ongoing process of modifications that we learn and store information in synapses.

Electrical impulses are constantly exchanged by the huge number of neurons ($\approx 10^{11}$) when the brain is active. Suppose, when information comes to be recorded, the neurons are already individually potentiated (or inhibited) to certain levels, which may be a base level or a level reached in the course of assimilating earlier information. The level of potentiation or inhibition will typically vary from one neuron to the other. The new information triggers them and some of them that might have been already near the threshold of firing might fire, i.e. send out electrical impulses, while the others remain quiescent. These impulses are received by other neurons via synapses, which, depending on their chemical character, whether excitatory or inhibitory, will excite or depress the neurons that are the recipients of the impulses. A neuron receives such excitatory and inhibitory inputs from a large number of pre-synaptic neurons and adds them linearly. If the net effect of the combined input makes the recipient neuron cross its threshold, which is pre-assigned to it by nature, then it fires an electrical impulse that is received by a large number of neurons via synapses. Note that the signal or impulse sent out by a neuron is replicated into as many of them as the number of neurons this particular neuron is synaptically connected with.

Thus, we see that the neurons might be already programmed to add linearly. The brain also knows how to multiply as an input from a pre-synaptic neuron goes to a post-synaptic neuron weighted by the synaptic efficacy of the synapse connecting them. The combined capabilities of neurons to add, and the neuron-synapse duos to multiply enable the neuronal network to form memories and the Hebbian plasticity enables them to be stored in the synapses. We further propose that when these competencies are extended over a collection of neurons and synapses, they enable them to also perform mathematical operations of higher levels like 'orthogonalization' and 'Fourier transformation'. We have studied these two mathematical operations, in particular, to propose that the brain might employ them respectively to discriminate between information [2,3] and make the long-term memory robust against trauma [4]. When we categorize information, we compare entities and isolate similarities and differences between them. To acquire this capability, we argue, the brain employs the mathematics involved in orthogonalization. Orthogonalization is a mathematical transformation that converts a given set of vectors into a set of mutually perpendicular or orthogonal vectors. So how is it connected with the brain and its capability to discriminate between information to categorize them? To address this question, we will first prepare the background.

PLOS ONE

RESEARCH ARTICLE

On stability and associative recall of memories in attractor neural networks

Suchitra Sampath[1☯]*, **Vipin Srivastava**[2☯]

1 Centre for Neural and Cognitive Sciences, University of Hyderabad, Hyderabad, Telangana, India, **2** School of Physics, University of Hyderabad, Hyderabad, Telangana, India

☯ These authors contributed equally to this work.
* suchitra.s85@gmail.com

Abstract

Attractor neural networks such as the Hopfield model can be used to model associative memory. An efficient associative memory should be able to store a large number of patterns which must all be stable. We study in detail the meaning and definition of stability of network states. We reexamine the meanings of retrieval, recognition and recall and assign precise mathematical meanings to each of these terms. We also examine the relation between them and how they relate to memory capacity of the network. We have shown earlier in this journal that orthogonalization scheme provides an effective way of overcoming catastrophic interference that limits the memory capacity of the Hopfield model. It is not immediately apparent whether the improvement made by orthgonalization affects the processes of retrieval, recognition and recall equally. We show that this influence occurs to different degrees and hence affects the relations between them. We then show that the conditions for pattern stability can be split into a necessary condition (recognition) and a sufficient one (recall). We interpret in cognitive terms the information being stored in the Hopfield model and also after it is orthogonalized. We also study the alterations in the network dynamics of the Hopfield network upon the introduction of orthogonalization, and their effects on the efficiency of the network as an associative memory.

Citation: Sampath S, Srivastava V (2020) On stability and associative recall of memories in attractor neural networks. PLoS ONE 15(9): e0238054. https://doi.org/10.1371/journal.pone.0238054

Editor: Ginestra Bianconi, Queen Mary University of London, UNITED KINGDOM

Received: January 15, 2020

Accepted: August 10, 2020

Published: September 17, 2020

Data Availability Statement: All relevant data are within the manuscript.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Associativity is a fundamental feature of learning and memory. When some information is learnt or memorized, it can be recalled not just when the same information is encountered again, but also by similar or partial information. The brain thus forms *associations* between the various information it learns and memorizes with those it encounters externally. This kind of associative memory can be modeled mathematically using some ideas from physics and mathematics which can be adapted to neuronal networks [1–6]. Such models of networks can help us gain insights into the mechanisms underlying learning and memory.

An attractor neural network (ANN) can be used to model associative memory [7]. Information is presented to the network as vectors, which can be sequences of numbers, usually ±1's forming a variety of patterns [1, 8, 9]. A set of patterns learnt by the network should form fixed

Abstract Submitted
for the MAR15 Meeting of
The American Physical Society

On basins of attraction in attractor neural networks SUCHITRA SAMPATH, VIPIN SRIVASTAVA[1], Centre for Neural and Cognitive Sciences, University of Hyderabad, Hyderabad -500046. India — We present an in-depth study of basin of attraction for patterns of ± 1 inscribed following Hebbian hypothesis [1] in a spin-glass like neural network. The aim is to investigate if basin of attraction being *non-zero* is a sufficient condition for the stability of an inscribed state when the necessary condition is that the inscribed state should be retrieved without any error. While this is true for Hopfield model [1], we find that the following model is an exception in that as many as $p=N-1$ stored patterns (N being the number of neurons in a fully connected network) can be retrieved without error while their basins of attraction consistently reduce in size as p increases and become zero around $p=0.8N$. The model proposes that the information that comes to be recorded in the brain is first orthogonalized (as in Gram-Schmidt orthogonalization) and then inscribed in synaptic weights. While the orthogonalized versions of input vectors with ± 1 components are stored in the model brain, the original vectors/patterns are retrieved exactly when checked for retrieval. Simulations are presented that give insight into the energy landscape in the space spanned by the network states.

[1] J.J. Hopfield,*PNAS* **79**, 2554(1982)

[1]Vipin Srivastava is also at School of Physics, University of Hyderabad and a life member of Indian Physics Association.

Vipin Srivastava
Centre for Neural and Cognitive Sciences & School of Physics,
University of Hyderabad, India.

Date submitted: 14 Nov 2014 Electronic form version 1.4

university of groningen

Proceedings of ICCM 2015

13th International Conference on Cognitive Modeling

April 9-11, Groningen, The Netherlands

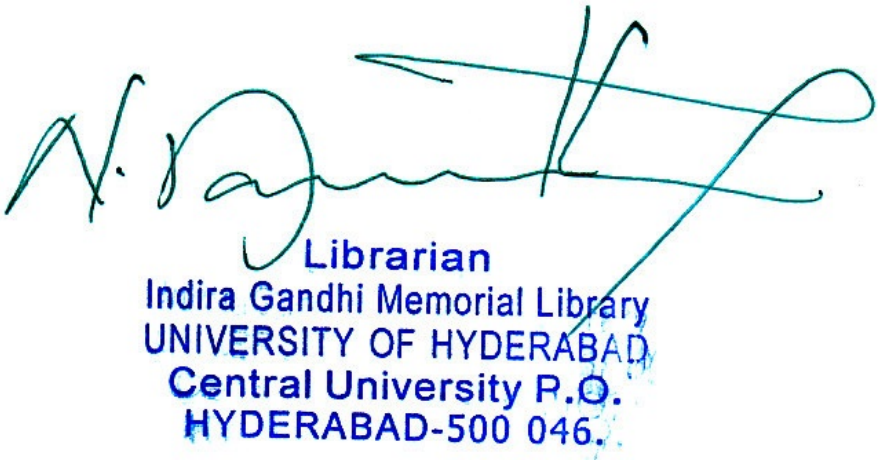
Edited by

Niels A. Taatgen
Marieke K. van Vugt
Jelmer P. Borst
Katja Mehlhorn

Table of Contents

A study on the memory capacity pattern stability and associativity in Attractor Neural Networks

by Suchitra S



A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks

ORIGINALITY REPORT

13%	10%	10%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	5%
2	journals.plos.org Internet Source	1%
3	dspace.rri.res.in Internet Source	1%
4	Suchitra Sampath, Vipin Srivastava. "On stability and associative recall of memories in attractor neural networks", PLOS ONE, 2020 Publication	1%
5	Vipin Srivastava, Suchitra Sampath. "Cognition of Learning and Memory", Elsevier BV, 2017 Publication	1%
6	Submitted to University of Salford Student Paper	<1%
7	tel.archives-ouvertes.fr Internet Source	<1%
8	link.springer.com Internet Source	<1%

Vipin Srivastava

This is to certify that out of 13% similarity 9% is from the students own publications, leaving 4% similarity

Prof. VIPIN SRIVASTAVA
SCHOOL OF PHYSICS
UNIVERSITY OF HYDERABAD
HYDERABAD-500 046,INDIA

9	Vipin Srivastava, S.F. Edwards. "A mathematical model of capacious and efficient memory that survives trauma", Physica A: Statistical Mechanics and its Applications, 2004 Publication	<1 %
10	www.uni-potsdam.de Internet Source	<1 %
11	Vipin Srivastava. Journal of Physics A Mathematical and General, 09/08/2000 Publication	<1 %
12	www.biorxiv.org Internet Source	<1 %
13	"Direct Methods in the Calculus of Variations", Springer Science and Business Media LLC, 2007 Publication	<1 %
14	Ellison, A.J.. "Raman study of potassium silicate glasses containing Rb^+, Sr^2^+, Y^3^+ and Zr^4^+: Implications for cation solution mechanisms in multicomponent silicate liquids", Geochimica et Cosmochimica Acta, 199404 Publication	<1 %
15	Vipin Srivastava, A. Ramesh Naidu. "New classes of orthogonal polynomials",	<1 %

	International Journal of Quantum Chemistry, 2006 Publication	
16	ddd.uab.cat Internet Source	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		

Certified that
out of 13% similarity
9% is due to the student's
own papers. That leaves only
4% similarity.

Prof. VIPIN SRIVASTAVA
SCHOOL OF PHYSICS
UNIVERSITY OF HYDERABAD
HYDERABAD-500 046,INDIA

A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks

by Suchitra Sampath

A study on the memory capacity, pattern stability and associativity in Attractor Neural Networks

ORIGINALITY REPORT

14%	10%	12%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	5%
2	journals.plos.org Internet Source	3%
3	Vipin Srivastava, Suchitra Sampath. "Cognition of Learning and Memory", Elsevier BV, 2017 Publication	1%
4	Suchitra Sampath, Vipin Srivastava. "On stability and associative recall of memories in attractor neural networks", PLOS ONE, 2020 Publication	1%
5	"Direct Methods in the Calculus of Variations", Springer Science and Business Media LLC, 2007 Publication	<1%
6	www.repository.cam.ac.uk Internet Source	<1%
7	www.biorxiv.org Internet Source	<1%

Out of 14% similarity 10% is from the student's own papers, leaving 4% similarity. This is to certify that only 4% similarity exists.

Prof. VIPIN SRIVASTAVA
SCHOOL OF PHYSICS
UNIVERSITY OF HYDERABAD
HYDERABAD-500 046,INDIA

8	Vipin Srivastava. Journal of Physics A Mathematical and General, 09/08/2000 Publication	<1 %
9	cuuduongthancong.com Internet Source	<1 %
10	Physics of Neural Networks, 1995. Publication	<1 %
11	dspace.rri.res.in Internet Source	<1 %
12	tel.archives-ouvertes.fr Internet Source	<1 %
13	beta-win.blogspot.com Internet Source	<1 %
14	www.nature.com Internet Source	<1 %
15	Adam J. H. Newton, Alexandra H. Seidenstein, Robert A. McDougal, Alberto Pérez-Cervera et al. "26th Annual Computational Neuroscience Meeting (CNS*2017): Part 3", BMC Neuroscience, 2017 Publication	<1 %
16	Submitted to Bocconi University Student Paper	<1 %
17	Shao-You Zhao, Wen-Li Yang, Yao-Zhong Zhang. "Determinant Representations of	<1 %

	Correlation Functions for the Supersymmetric t-J Model", Communications in Mathematical Physics, 2006 Publication	
18	Till Frank. "Determinism and Self-Organization of Human Perception and Performance", Springer Science and Business Media LLC, 2019 Publication	<1 %
19	Young Jin Suh. "Recurrent real hypersurfaces in complex two-plane Grassmannians", Acta Mathematica Hungarica, 2006 Publication	<1 %
20	link.springer.com Internet Source	<1 %
21	"Neural Information Processing", Springer Science and Business Media LLC, 2017 Publication	<1 %
22	Jianfeng Feng. Journal of Physics A Mathematical and General, 05/21/1997 Publication	<1 %
23	Ellison, A.J.. "Raman study of potassium silicate glasses containing Rb^+, Sr^2^+, Y^3^+ and Zr^4^+: Implications for cation solution mechanisms in multicomponent silicate liquids", Geochimica et Cosmochimica Acta, 199404 Publication	<1 %

24	L F Abbott. "Limits on the memory storage capacity of bounded synapses", Nature Neuroscience, 03/11/2007 Publication	<1 %
25	Marcelo A Montemurro, Francisco A Tamarit. "An efficient dilution strategy for constructing sparsely connected neural networks", Physica A: Statistical Mechanics and its Applications, 2001 Publication	<1 %
26	Lathi, B. P.. "Linear Systems and Signals", Oxford University Press Publication	<1 %
27	Xinmin Liu, Zongli Lin. "On stabilization of nonlinear systems affine in control", 2008 American Control Conference, 2008 Publication	<1 %
28	Srivastava, V.. "A mathematical model of capacious and efficient memory that survives trauma", Physica A: Statistical Mechanics and its Applications, 20040215 Publication	<1 %
29	docslide.us Internet Source	<1 %
30	idoc.pub Internet Source	<1 %
	Submitted to University of Salford	

31	Student Paper	<1%
32	vuir.vu.edu.au Internet Source	<1%
33	dokumen.pub Internet Source	<1%
34	A. Ramesh Naidu. "Löwdin's canonical orthogonalization: Getting round the restriction of linear independence", International Journal of Quantum Chemistry, 2004 Publication	<1%
35	www.imes.boj.or.jp Internet Source	<1%
36	JOEL FELDMAN, HORST KNÖRRER, EUGENE TRUBOWITZ. "SINGLE SCALE ANALYSIS OF MANY FERMION SYSTEMS PART 2: THE FIRST SCALE", Reviews in Mathematical Physics, 2011 Publication	<1%
37	www.mdpi.com Internet Source	<1%
38	www.uni-potsdam.de Internet Source	<1%
39	epdf.pub Internet Source	<1%

40	www.springerprofessional.de Internet Source	<1 %
41	Sebastian Grijalva, Jacopo De Nardis, Véronique Terras. "Open XXZ chain and boundary modes at zero temperature", SciPost Physics, 2019 Publication	<1 %
42	www.scribd.com Internet Source	<1 %
43	www.slideshare.net Internet Source	<1 %
44	"CS2-PC-22_HR", ActEd Publication	<1 %
45	Vipin Srivastava, S.F. Edwards. "A mathematical model of capacious and efficient memory that survives trauma", Physica A: Statistical Mechanics and its Applications, 2004 Publication	<1 %
46	nozdr.ru Internet Source	<1 %
47	D. Bollé, J.Busquets Blanco, G.M. Shim, T. Verbeiren. "The Blume–Emery–Griffiths neural network: dynamics for arbitrary temperature", Physica A: Statistical Mechanics and its Applications, 2004 Publication	<1 %

48	M.S Mainieri, R Erichsen. "Retrieval and chaos in extremely diluted non-monotonic neural networks", Physica A: Statistical Mechanics and its Applications, 2002 Publication	<1%
49	Peter Eichelsbacher, Matthias L&#65533;we. "Moderate Deviations for the overlap parameter in the Hopfield model", Probability Theory and Related Fields, 2004 Publication	<1%
50	Submitted to University of Birmingham Student Paper	<1%
51	Zak, Stanislaw H.. "Systems and Control", Oxford University Press Publication	<1%
52	www.springeropen.com Internet Source	<1%
53	S. Bhatnagar, H.L. Prasad, L.A. Prashanth. "Stochastic Recursive Algorithms for Optimization", Springer Science and Business Media LLC, 2013 Publication	<1%
54	Steiner, Erich. "The Chemistry Maths Book", The Chemistry Maths Book, 2008 Publication	<1%
55	scholarspace.manoa.hawaii.edu Internet Source	<1%

56	"Elements of Geometric Quantization", Graduate Texts in Contemporary Physics, 2005 Publication	<1%
57	David J. Aldous, Charles Bordenave, Marc Lelarge. "Dynamic Programming Optimization over Random Data: The Scaling Exponent for Near-Optimal Solutions", SIAM Journal on Computing, 2009 Publication	<1%
58	Emmanuel Jerez Usuga. "Group cohomology based on partial representations", Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2020 Publication	<1%
59	Vipin Srivastava, A. Ramesh Naidu. "New classes of orthogonal polynomials", International Journal of Quantum Chemistry, 2006 Publication	<1%
60	ddd.uab.cat Internet Source	<1%

Certified that out of 14% similarity 10% is due to the student's own publications leaving only 4% similarity.

Prof. VIPIN SRIVASTAVA
SCHOOL OF PHYSICS
UNIVERSITY OF HYDERABAD
HYDERABAD-500 046,INDIA

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	Off		