Università degli Studi di Padova

DIPARTIMENTO DI MATEMATICA “TULLIO LEVI-CIVITA”
Corso di Laurea Magistrale in Matematica

Towards a continuous dynamic model of the Hopfield theory on neuronal interaction and memory storage

Supervisor:
Prof. Franco Cardin

Author:
Laura Meneghetti
1155338

October 12, 2018
Academic Year 2017-2018

What is, then, that makes scientists wander about in universe of ideas and experimentation? It may be the search for knowledge or, in more mundane terms, simple curiosity. Nagging questions; the pressing need to figure something out and the inability to do anything else until the answer is found; the tingling feeling that a discovery may be just around the corner; the intuition that a puzzle is starting to take shape, until eventually one reaches the answer and feels the thrilling joy of understanding.

—Rodrigo Quian Quiroga
Borges and Memory: Encounters with the Human Brain

Contents

Introduction ix

Introduzione xiii

1 Memory in Human Brain 1

1.1 Human Brain . . . 1

1.1.1 The cerebrum . . . 2

1.1.2 The Limbic system and Memory . . . 4

1.2 Memory . . . 6

1.2.1 The multi-store model . . . 8

1.2.2 Memory processes . . . 10

1.3 Neurons and Synapses . . . 15

1.3.1 The action potential . . . 16

1.3.2 Synapses . . . 17

2 Hopfield Model 19

2.1 Towards the Hopfield model . . . 19

2.2 Hopfield model: an example of Content Addressable Memory . . . 21

2.2.1 The discrete Hopfield model . . . 21

2.2.2 Hamming distance and overlap measure . . . 25

2.2.3 The interconnection matrix and Hebb's rule . . . 26

2.2.4 Fixed stable points . . . 27

2.2.5 The energy function . . . 32

2.2.6 Spurious states and eigenanalysis . . . 37

2.2.7 The Attraction Basin . . . 40

2.3 A continuous deterministic model . . . 41

2.3.1 Additive model . . . 42

2.3.2 The energy function . . . 44

2.3.3 Relation between the stable states of the two models . . . 46

2.3.4 Cohen-Grossberg Theorem . . . 48

2.4 Improvements of the Hopfield model . . . 49

2.4.1 Asymmetric neural networks . . . 50

2.4.2 Biased Patterns . . . 53

2.4.3 Inferring Learning rule from *In Vivo* data . . . 60

2.5 A probabilistic approach . . . 62

2.5.1 Symmetric case . . . 63

2.5.2 General asymmetric case . . . 65

3 Kuramoto Model 71

3.1 The role of synchronization in nature . . . 71

3.2 The Kuramoto model . . . 73

3.2.1 Phase method . . . 73

3.2.2 Derivation of the model . . . 76

3.2.3 Analysis of the Kuramoto model 77
3.2.4 Order Parameter 78
3.2.5 Synchronization 79
3.2.6 Initial Growth of K_c 86
3.2.7 The bimodal case 88
3.2.8 Summary of the Kuramoto Model 89
3.3 Structural stability of the Kuramoto model 90
3.3.1 Malkin Theorem 91
3.4 Improvements of the Kuramoto model 94
3.4.1 Time delays 95
3.4.2 Dynamically changing coupling strength 96
3.5 Learning with oscillators 96
3.5.1 Application to the Kuramoto model 98
3.6 From Hopfield to Kuramoto model 99
3.6.1 Integrate and fire model 99
3.6.2 Lighthouse model 100
3.6.3 From IFM to LM 102
3.6.4 Neurons as oscillators 103
3.6.5 From Hopfield to Kuramoto model 106

A Proof of Eq. 2.48 **113**

B Discretization of Eq.2.69 **115**

C Spectral method **117**

D Proof of $\rho v = constant$ **123**

E Residual method **127**

F Proof of Theorem 3.6.3 **131**

Bibliography **133**

Ringraziamenti

Arrivata finalmente alla fine di questa avventura, desidero ringraziare tutte le persone che hanno dato un contributo a questo lavoro di tesi e che mi hanno supportata in questi ultimi due anni.

Ringrazio il mio relatore Franco Cardin, che mi ha sostenuto durante questi mesi e ha sempre apprezzato il grande entusiasmo che ho messo in questo lavoro. Sono inoltre grata nei suoi confronti perchè mi ha proposto un'argomento molto accattivante e stimolante, mettendomi alla prova con un contesto completamente nuovo e all'avanguardia per la ricerca. Grazie a questa tesi, infatti, è nato in me il desiderio di continuare con un dottorato e di esplorare campi sempre nuovi come quello delle reti neurali. Lo ringrazio infine per avermi dato la possibilità di apprendere come il mondo della ricerca non sia sempre così facile, ma a volte ci siano periodi duri in cui non si trova l'idea giusta per proseguire o anzi in cui si capisce che l'obiettivo prefissato è più difficile da raggiungere del previsto. Sicuramente tutto questo mi sarà molto utile nel mio prossimo futuro.

Ringrazio poi i Junior Math Days per avermi dato la possiblità di conoscere la SISSA e tanti nuovi amici. Sono grata quindi verso la SISSA per avermi accolto sin dal primo giorno come un nuovo membro della famiglia e avermi fatto innamorare di questo ambiente ricco di idee e di novità che sarà la mia casa per i prossimi anni.

Un grazie va infine ai miei genitori, a mia sorella e ai miei amici che mi fanno sentire ogni giorno fortunata e amata. Li ringrazio per avermi sopportata e aiutata in questi anni, per aver ascoltato le mie note vocali infinite, le mie lamentale e le mie gioie e spero resteranno al mio fianco anche per le future preoccupazioni e soddisfazioni.

Introduction

In the last decades a growing amount of researchers have focused their attention on the study of brain and human memory, because it is widely believed that a full understanding of the mechanisms of memorization, recognition and retrieval of objects is necessary to understand the other functions of the brain.

Human brain is usually described as a highly complex, nonlinear, parallel computer (information-processing system), that has the capability to organize its structural constituents so as to perform certain computations (e.g. pattern recognition, perception and motor control) many times faster than the fastest digital computer in existence today. Its complex structure made of interconnected neurons and suitable to memorize our memories has led to the creation of a mathematical model, called *neural network*, that could simulate the biological one in the structure and in the great computational skills.

Therefore in analogy with the characteristics of the brain, a *neural network* can be defined as a massively parallel distributed processor made up of simple processing units (the neurons), which has a natural propensity for storing experiential knowledge and making it available for use. As proposed in Haykin's book [25], the most important characteristics of neural networks are:

- Knowledge is acquired by the network from its environment through a learning process.
- Interneuron connection strengths, known as *synaptic weights*, are used to store the acquired knowledge.
- A popular paradigm of learning, called *learning with a teacher* or *supervised learning*, involves modification of the synaptic weights by applying a set of labeled *training samples* or *task examples.* Each example consists of a unique input signal and a corresponding desired response. The network is presented with an example picked at random from the set and the synaptic weights of the network are modified to minimize the difference between the desired response and the actual response produced by the input signal in accordance with an appropriate statistical criterion. The training of the network is repeated for many examples in the set until it reaches a steady state where there are no further significant changes in the synaptic weights. Thus the network learns from the examples by constructing an *input-output mapping* for the problem at hand.
- *Adaptivity.* Neural networks have a built-in capability to *adapt* their synaptic weights to changes in the surrounding environment. In particular, a neural network trained to operate in a specific environment can be easily retrained to deal with minor changes in the operating environmental conditions.
- Knowledge is represented by the very structure and activation state of a neural network. Every neuron is potentially affected by the global activity of all other neurons in the network. Consequently, *contextual information* is dealt with naturally by a neural network.
- A neural network has the potential to be inherently *fault tolerant* or capable of robust computation in the sense that its performance degrades gracefully under adverse

operating conditions. For example, if a neuron or its connecting links are damaged, recall of stored pattern is impaired in quality. However, due to the distributed nature of information stored in the network, the damage has to be extensive before the overall response of the network is degraded seriously. In other words, a neural network is error-correcting, in the sense that it can override inconsistent information in the cues presented it.

- The design of a neural network is motivated by analogy with the brain, which is a living proof that fault tolerant parallel processing is not only physically possible but also fast and powerful. Neurobiologists look to neural networks as a research tool for the interpretation of neurobiological phenomena. On the other hand, engineers look to neurobiology for new ideas to solve problems more complex than those based on conventional hard-wired design techniques.

To get a full understanding of the biological model in chapter 1 we present a brief description of the human brain and memory-related brain structures. We will therefore try to understand the different stages of the memory process and how neurons communicate with each other through the synapses.
With this inspiration in mind, a starting point is represented by Hopfield article of 1982 [28], where a new model for a fully connected neural network that generalized the previous one of McCullock and Pitts for artificial neurons is proposed. The description and full understanding of *Hopfield model* is the main topic of chapter 2, that starts with the introduction of the discrete model, as proposed by Hopfield in [28].
We consider N neurons, that can take on either bipolar values $V_i = -1$ when it "not firing" and $V_i = 1$ if it is "firing at maximum rate". The idea of the model is the following: we have p patterns, ξ^μ with $\mu = 1, \dots, p$ and we want to store and retrieve them when necessary using a suitable updating algorithm. The dynamics of the i-th neuron is therefore given by the following discrete-time equation

$$V_i^{(t+1)} = \operatorname{sgn}\Big(\sum_{j=1}^{N} T_{ij} V_j^{(t)}\Big) = \begin{cases} +1, & \text{if } \sum_j T_{ij} V_j^{(t)} > 0 \\ -1, & \text{if } \sum_j T_{ij} V_j^{(t)} < 0 \end{cases}$$

where T_{ij} represents the synaptic matrix, defined through *Hebb's rule*

$$T_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} (\xi_i^\mu \xi_j^\mu - \delta_{ij})$$

T is characterized by the strong assumption of being symmetrical, that leads to the definition of an energy function. The presence of this function is very important since it describes the dynamics of the network and in particular can determine the convergence to memorized patterns. In this context it seems plausible to think that these vectors are exactly the stable states of the dynamics and then initial states that are close in Hamming distance to a particular stable state and far from all others will tend to terminate in that nearby stable state. Unfortunately this property is not always so easy to verify, in fact there can be two cases: the stored vectors form an orthogonal basis or not.

- In the former case it can be proved that the ξ^μ are eigenvectors with positive eigenvalues for T and then in according to Hopfield [28] the energy function

$$E = -\frac{1}{2} \sum_i \sum_j T_{ij} V_i V_j$$

is a Lyapunov function.

- In the latter case, the presence of a crosstalk term, that represents a sort of noise, breaks this harmony and the fundamental memories are just approximate eigenvectors. Only a probabilistic handling can be carried out in this context, giving conditions to obtain network stability.

Using as main reference Hopfield article [29], a continuous version of the model, that maintains the same salient characteristics of the other, can be constructed:

$$\frac{du_j}{dt} = \sum_i T_{ji} V_i - u_j + I_j$$

Hence, the extension of the analysis to this domain leads to a more reasonable model, since neurons are more likely to be continuous variables than an all-or-none basis.
Hopfield model is thus essential as a starting point to understand the process of memorization in the brain, but it has many weaknesses that make it scarcely biologically plausible. The hypothesis of having symmetrical connections will be strongly used throughout chapter 2, but this is not the case of human brain. From medical studies it is well known that for a given neuron the synapses are of only one type (excitatory or inhibitory) and with the requirement of symmetry this would imply two populations of neurons not connected to each other. So following Amit in [4] and Sompolinsky in [55], we will present several ways to account for this property without losing the main features of the original model. Even if, from the point of view of the number of spurious states and network capacity, the new model works well, it lacks a fundamental characteristic: a Lyapunov energy function that guarantees convergence.
A different approach can thus be carried out in the study of neuronal dynamics: instead of following the individual deterministic trajectories of every neuron, we can prefer its probabilistic evolution. Starting from the continuous Hopfield equation, we will derive the associated Fokker-Planck equation, inspired by [18] and [52]:

$$\frac{\partial P(u,t)}{\partial t} = -\sum_{j=1}^{N} \frac{\partial}{\partial u_j}[F_j(u)P(u,t)] + \mathcal{D}\sum_{j=1}^{N}\sum_{i=1}^{N} \frac{\partial^2}{\partial u_j \partial u_i}[G_{ji}(u,t)P(u,t)]$$

where $G(u)$ represents the scaled diffusion matrix, $F(u)$ is the driving force for the dynamics of the underlying neural networks, that in our case is given by

$$F_j(u) = -u_j + \sum_i J_{ji} V_i + I_j \qquad \text{for } j = 1, \ldots, N$$

and J_{ij} the asymmetric interconnection matrix. We will thus be able to determine a Lyapunov energy function, that will characterize and describe the convergence to stable states also in this case.
Beside this description of the neuronal dynamics linked to the transmission of impulses (spikes) between neurons, there is a description in terms of oscillators. In the third chapter of this thesis, we have thus studied *Kuramoto model* for coupled oscillators, following Strogatz [58], Mori and Kuramoto [41] and Acebroń et al. [2].
We consider a set of N oscillators, whose states are defined by phase alone, $\phi_i(t)$. The actual time evolution of these phases is governed by a system of coupled differential equations with phase-interaction terms between the oscillators and is given by

$$\dot{\phi}_i(t) = \omega_i + \frac{K}{N}\sum_{j=1}^{N} \sin(\phi_i - \phi_j) \qquad i = 1, 2, \ldots, N$$

where ω_i is the natural frequency of the i-th oscillator and K is the coupling strength between all oscillator pairs. As was famously proved analytically by Kuramoto, the sine-coupling term drives the oscillators towards synchronization, but only if the coupling constant K exceeds a critical value of K_c. In fact, under appropriate hypotheses, the model exhibits a transition behavior in which a critical coupling constant K_c separates two types of regime: one in which the system is incoherent ($K < K_c$) and the other in which there is a partial and progressive synchronization. In our work, we have then investigated the plausibility of this model; in this regard, the condition of existence that regulates the stability of the interacting oscillator systems is exposed. The focal point of our elaboration consists in the structural verification of the model, which works despite the generating function is perturbed, as can be seen using Malkin Theorem.

From the biological point of view we know that oscillations occur in the brain, thus Kuramoto model seems to be a good candidate to describe the memorization process and neural networks. Following Hoppensteadt and Izhikevich [30], we will show how Kuramoto model has interesting neurocomputational properties. In fact, the existence of a convergence theorem guarantees how it can be used to store new patterns that coincide with phase-locked oscillations to which the network converges. Then, thanks to the contribution of Haken, [24], we can move from the neuronal description through spikes to a phase model, connecting the neurons to the oscillators:

$$\phi_i = \pi(u_i + 1)$$

in this way each oscillator i can be seen as a point on a circumference that emits an action potential every time its phase ϕ_i is equal to 2π. In this context, therefore, the synchronization status is reached if neuron i transmits a spikes to neuron j, to which it is connected, when its phase is close to 2π, thus increasing the possibility of triggering an action potential in neuron j. This situation then occurs when the frequencies of the start and arrival neurons are outof phase by an arbitrary small factor $\tilde{\epsilon}$, which takes into account the delay due to the finite propagation speed in an electrical resistance environment.

Our aim is therefore to understand how and why neurons can be seen as oscillators and to establish a strong link between this model and Hopfield approach. Inspiring from Haken's book [24] and Hoppensteadt and Izhikevich [30], we will trace a path that will take us from Hopfield model to that of Kuramoto, exploiting the realistic hypothesis of weakly connections.

Introduzione

Negli ultimi anni un gran numero di ricercatori ha focalizzato la loro attenzione sullo studio del cervello e della memoria umana, poichè si ritiene che una comprensione completa dei meccanismi di memorizzazione, riconoscimento e recupero degli oggetti rappresentanti i nostri ricordi sia necessaria per capire le altre funzioni del cervello.

Il cervello umano di solito è descritto come un computer altamente complesso e non lineare (un sistema processante informazioni), che ha la capacità di organizzare i suoi costituenti strutturali in modo da eseguire determinati calcoli (e.g. il riconoscimento di pattern, la percezione e il controllo motorio) molto più velocemente del più veloce computer digitale esistente oggi. La sua struttura complessa fatta di neuroni interconnessi e adatta a memorizzare i nostri ricordi ha portato alla creazione di un modello matematico, chiamato *rete neurale*, che potesse simulare il modello biologico nella struttura e nelle grandi capacità computazionali.

In analogia con le caratteristiche del cervello, una *rete neurale* può quindi essere definita come un processore distribuito parallelamente fatto di singole unità processanti (i *neuroni*), che ha una naturale predisposizione a immagazzinare la conoscenza esperienziale e a renderla disponibile all'uso. Come proposto da Haykin in [25], le più importanti caratteristiche delle reti neurali sono:

- La conoscenza è acquisita della rete dall'ambiente circostante tramite un processo di apprendimento.

- Le forze delle connessioni interneuronali, note come *pesi sinaptici*, sono usate per immagazzinare la conoscenza acquisita.

- Un paradigma popolare per l'apprendimento, chiamato *apprendimento supervisionato*, coinvolge modificazioni dei pesi sinaptici applicando un insieme di esempi di addestramento etichettati o esempi di attività. Ogni esempio è costituito da un segnale di ingresso unico e una corrispondente risposta desiderata. La rete viene inizializzata con un esempio scelto a caso dall'insieme e i pesi sinaptici della rete vengono modificati per ridurre al minimo la differenza tra la risposta desiderata e la risposta effettiva prodotta dal segnale di ingresso in conformità con un criterio statistico appropriato. L'addestramento della rete viene ripetuto per molti esempi dell'insieme fino a raggiungere uno stato stazionario in cui non vi sono ulteriori cambiamenti significativi nei pesi sinaptici. Quindi la rete impara dagli esempi costruendo una *mappa di input-output* per il problema in questione.

- *Adattività.* Le reti neurali hanno una capacità integrata di adattare i loro pesi sinaptici ai cambiamenti dell'ambiente circostante. In particolare, una rete neurale addestrata per operare in un ambiente specifico può essere facilmente riqualificata per far fronte a piccoli cambiamenti nelle condizioni ambientali operative.

- La conoscenza è rappresentata dalla struttura e dallo stato di attivazione di una rete neurale. Ogni neurone è potenzialmente influenzato dall'attività globale di tutti gli altri neuroni nella rete. Di conseguenza, *l'informazione contestuale* è trattata naturalmente da una rete neurale.

- Una rete neurale ha il potenziale per essere *tollerante agli errori* o capace di un calcolo robusto nel senso che le sue prestazioni si degradano con garbo in condizioni operative avverse. Ad esempio, se un neurone o i suoi collegamenti sono danneggiati, il richiamo del pattern memorizzato è compromesso in termini di qualità. Tuttavia, a causa della distribuzione naturale delle informazioni memorizzate nella rete, il danno deve essere molto esteso prima che la risposta complessiva della rete sia seriamente compromessa. In altre parole, una rete neurale è in grado di correggere gli errori, nel senso che può ignorare e oltrepassare le informazioni incoerenti nei segnali presentati.

- La costruzione di una rete neurale è motivata dall'analogia con il cervello, che è una prova vivente che l'elaborazione parallela tollerante agli errori non è solo fisicamente possibile, ma anche veloce e potente. Da un lato, i neurobiologi guardano alle reti neurali come uno strumento di ricerca per l'interpretazione dei fenomeni neurobiologici. D'altra parte, gli ingegneri guardano alla neurobiologia per nuove idee per risolvere problemi più complessi di quelli basati su tecniche di progettazione convenzionali.

Per ottenere una completa conoscenza del modello biologico, nel capitolo 1 presentiamo una breve descrizione del cervello umano e delle sue strutture collegate alla memoria. Cercheremo quindi di capire i vari passaggi del processo di memorizzazione e come i neuroni comunichino tra di loro attraverso le sinapsi.
Con questa ispirazione in mente, un punto di inizio è rappresentato dall'articolo di Hopfield del 1982 [28], dove è proposto un nuovo modello per una rete neurale completamente connessa, che generalizza il lavoro di McCullock e Pitts. La descrizione e completa comprensione del *modello di Hopfield* rappresenta l'argomento principale del capitolo 2, che inizia introducendo il modello discreto come proposto da Hopfield in [28].
Si considerino N neuroni, i cui stati possono assumere due diversi valori: $V_i = -1$ quando non è acceso e $V_i = 1$ quando è accesso. L'idea del modello è la seguente: si hanno p pattern ξ^μ con $\mu = 1, \ldots, p$ e vogliamo immagazzinarli e recuperarli quando necessario usando un appropriato algoritmo di aggiornamento. La dinamica del neurone i-esimo è quindi data dalla seguente equazione a tempo discreto

$$V_i^{(t+1)} = \operatorname{sgn}\Big(\sum_{j=1}^{N} T_{ij} V_j^{(t)}\Big) = \begin{cases} +1, & \text{if } \sum_j T_{ij} V_j^{(t)} > 0 \\ -1, & \text{if } \sum_j T_{ij} V_j^{(t)} < 0 \end{cases}$$

dove T_{ij} reppresenta la matrice sinaptica, definita attraverso la *regola di Hebb*

$$T_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} (\xi_i^\mu \xi_j^\mu - \delta_{ij})$$

T è caratterizzata dal forte presupposto di essere simmetrica, che porta alla definizione di una funzione energia. La presenza di questa funzione è molto importante in quanto descrive la dinamica della rete e in particolare può determinare la convergenza ai pattern memorizzati. In questo contesto sembra plausibile pensare che questi vettori siano esattamente gli stati stabili della dinamica e quindi gli stati iniziali che sono vicini nella distanza di Hamming a un particolare stato stabile e lontani da tutti gli altri tenderanno a terminare in quello stato stazionario vicino. Purtroppo questa proprietà non è sempre così facile da verificare, infatti ci possono essere due casi: i vettori memorizzati formano una base ortogonale o meno.

- Nel primo caso può essere provato che i vari ξ^μ sono autovettori con autovalore positivo per T e che la funzione energia

$$E = -\frac{1}{2}\sum_i \sum_j T_{ij} V_i V_j$$

è una funzione di Lyapunov, in accordo a quanto discusso da Hopfield in [28].

- Nell'ultimo caso, la presenza di un "termine di crosstalk", che rappresenta una sorta di rumore per il modello, rompe questa armonia e i vari vettori memorizzati sono solo autovettori approssimati. In questo contesto si può quindi portare avanti soltanto una trattazione di tipo probabilistico per ottenere condizioni sulla stabilità della rete.

Usando come riferimento maggiore l'articolo di Hopfield [29], una versione continua del modello, che mantenga le stesse caratteristiche salienti dell'altro, può essere costruita:

$$\frac{du_j}{dt} = \sum_i T_{ji} V_i - u_j + I_j$$

Quindi, l'estensione dell'analisi a questo dominio porta a un modello più ragionevole, poichè i neuroni hanno più probabilità di essere variabili continue piuttosto che di tipo "tutto o niente".
Il modello di Hopfield è quindi essenziale come punto di partenza per comprendere il processo di memorizzazione nel cervello, ma ha molte debolezze che lo rendono poco plausibile dal punto di vista biologico. L'ipotesi di avere connessioni simmetriche sarà fortemente utilizzata in tutto il capitolo 2, ma questa non è verificata nel caso del cervello umano. Dagli studi medici è noto che per un dato neurone le sinapsi sono di un solo tipo (eccitatorio o inibitorio) e con il requisito della simmetria ciò implicherebbe due popolazioni di neuroni non collegate tra loro. Quindi, seguendo Amit in [4] e Sompolinsky in [55], presenteremo diversi modi per introdurre questa proprietà senza perdere le caratteristiche principali del modello originale. Anche se, dal punto di vista del numero di stati spuri e della capacità di rete, il nuovo modello funziona bene, manca una caratteristica fondamentale: una funzione energia di Lyapunov che garantisce la convergenza.
Un approccio diverso può quindi essere portato avanti nello studio della dinamica neuronale: invece di seguire le traiettorie individuali deterministiche di ogni neurone, possiamo concentraci sulla sua evoluzione probabilistica. Iniziando dall'equazione di Hopfield continua, deriveremo l'equazione di Fokker-Planck associata, prendendo ispirazione da [18] e [52]:

$$\frac{\partial P(u,t)}{\partial t} = -\sum_{j=1}^{N} \frac{\partial}{\partial u_j}[F_j(u)P(u,t)] + \mathcal{D}\sum_{j=1}^{N}\sum_{i=1}^{N} \frac{\partial^2}{\partial u_j \partial u_i}[G_{ji}(u,t)P(u,t)]$$

dove $G(u)$ rappresenta la matrice di diffusione scalata, $F(u)$ è la forza che guida la dinamica della rete neurale sottostante, che nel nostro caso è espressa da

$$F_j(u) = -u_j + \sum_i J_{ji} V_i + I_j \qquad \text{for } j = 1, \ldots, N$$

e J_{ij} la matrice di connessioni asimmetrica. Saremo quindi capaci di determinare una funzione energia di Lyapunov, che caratterizzerà e descriverà la convergenza verso gli stati stabili anche un questo caso.
Oltre a questa descrizione della dinamica neuronale collegata alla trasmissione di impulsi (spikes) tra i neuroni, c'è una descrizione in termini di oscillatori. Nel terzo capitolo di questa

tesi, abbiamo quindi studiato il *modello di Kuramoto* per oscillatori accoppiati, seguendo Strogatz [58], Mori e Kuramoto [41] e Acebroń et al. [2].
Consideriamo un insieme di oscillatori N, i cui stati sono definiti dalla sola fase, $\phi_i(t)$. L'evoluzione temporale di queste fasi è regolata da un sistema di equazioni differenziali accoppiate con termini di interazione di fase tra gli oscillatori ed è data da

$$\dot{\phi}_i(t) = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\phi_i - \phi_j) \qquad i = 1, 2, \ldots, N$$

dove ω_i è la frequenza naturale dell'oscillatore i-esimo e K è la forza di accoppiamento tra tutte le coppie di oscillatori. Come è stato provato analiticamente da Kuramoto, il termine di accoppiamento sinusoidale guida gli oscillatori verso la sincronizzazione, ma solo se la costante di accoppiamento K supera un valore critico di K_c. Di fatto, in ipotesi appropriate, il modello mostra un comportamento di transizione in cui una costante di accoppiamento critica K_c separa due tipi di regime: uno in cui il sistema è incoerente ($K < K_c$) e l'altro in cui vi è una sincronizzazione parziale e progressiva. Nel nostro lavoro, abbiamo quindi studiato la plausibilità di questo modello; a questo proposito, viene esposta la condizione di esistenza che regola la stabilità dei sistemi di oscillatori interagenti. Il punto focale della nostra elaborazione consiste quindi nella verifica strutturale del modello, che funziona malgrado la perturbazione della funzione generatrice, come vede grazie al Teorema di Malkin.
Dal punto di vista biologico sappiamo che le oscillazioni si verificano nel cervello, quindi Il modello di Kuramoto sembra essere un buon candidato per descrivere il processo di memorizzazione e le reti neurali. Seguendo Hoppensteadt e Izhikevich [30], mostreremo come il modello di Kuramoto abbia interessanti proprietà neurocomputazionali. Infatti, l'esistenza di un teorema di convergenza garantisce come possa essere utilizzato per memorizzare nuovi pattern che coincidono con le oscillazioni a fasi bloccate a cui la rete converge. Quindi, grazie al contributo di Haken, [24], possiamo passare dalla descrizione neuronale attraverso gli spikes a un modello per le fasi, collegando i neuroni agli oscillatori:

$$\phi_i = \pi(u_i + 1)$$

in questo modo ogni oscillatore i può essere visto come un punto su una circonferenza che emette un potenziale d'azione ogni volta che la sua fase ϕ_i è uguale a 2π. In questo contesto, quindi, lo stato di sincronizzazione viene raggiunto se il neurone i trasmette un picco al neurone j, a cui è connesso, quando la sua fase è vicina a 2π, aumentando così la possibilità di innescare un potenziale d'azione nel neurone j. Questa situazione si verifica quando le frequenze dei neuroni di inizio e di arrivo sono fuori fase di un piccolo fattore arbitrario $\tilde{\epsilon}$, che tiene conto del ritardo dovuto alla velocità di propagazione finita in un ambiente di resistenza elettrica.
Il nostro obiettivo è quindi capire come e perchè i neuroni possono essere visti come oscillatori e stabilire un forte legame tra questo modello e l'approccio di Hopfield. Traendo ispirazione dal libro di Haken [24] e Hoppensteadt e Izhikevich [30], tracciamo un percorso che ci porterà dal modello di Hopfield a quello di Kuramoto, sfruttando l'ipotesi realistica di connessioni deboli.

CHAPTER 1 Memory in Human Brain

In this chapter we present some basic notions about neurons, brain and memory from the biological and physiological point of view following the books of Klinke et al. [33], Haykin [25], Haken [24] and Quiroga [50].

1.1 Human Brain

The human nervous system may be viewed as a three-stage system, as depicted in the block diagram of Fig. 1.1. Central to the system is the *brain*, represented by the *neural (nerve) net*, which continually receives information, perceives it and makes appropriate decisions. Two sets of arrows are shown in the figure. Those pointing from left to right indicate the *forward* transmission of information-bearing signals through the system. The arrows pointing from right to left signify the presence of *feedback* in the system. The *receptors* convert stimuli from the human body or the external environment into electrical impulses that convey information to the neural net (brain). The *effectors* convert electrical impulses generated by the neural net into discernible responses as system outputs.
The human brain is hugely interconnected but three major components can be identified: the *cerebrum*, the *cerebellum* and the *brainstem*.

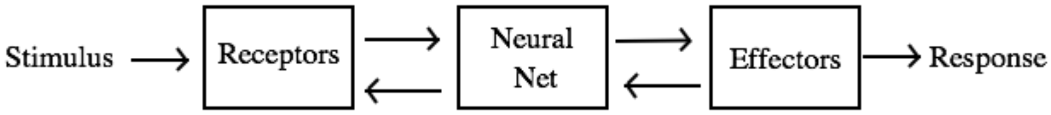


Figure 1.1: Block diagram representation of nervous system. Figure taken from the book of Arbib and Bonaiuto, [7].

- The *brainstem*, which includes the medulla, the pons and the midbrain, controls breathing, digestion, heart rate and other autonomic processes, as well as connecting the brain with the spinal cord and the rest of the body.
- The *cerebellum* plays an important role in balance, motor control, but is also involved in some cognitive functions such as attention, language, emotional functions (e.g. regulating fear and pleasure responses) and in the processing of procedural memories.
- The *cerebrum* (or *forebrain*) is a large part of the brain containing the cerebral cortex (of the two cerebral hemispheres), as well as several subcortical structures, including the hippocampus, basal ganglia and olfactory bulb. In the human brain, the cerebrum is the uppermost region of the central nervous system and is involved in many cognitive functions such as movements, sensory processing, language, communication, learning and memory.

Since most of the memory processes take place in the cerebrum we will now focus on the description of its structure and functionality.

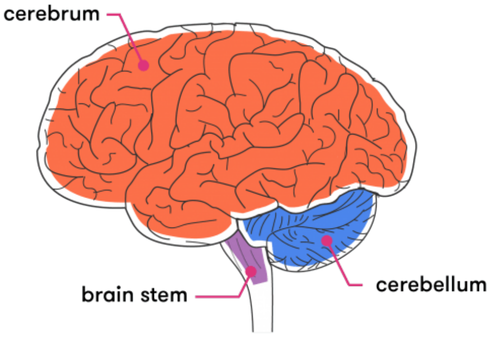


Figure 1.2: The three major components of the human brain.

1.1.1 The cerebrum

The cerebrum is the largest and, for most people, the most easily recognizable part of the brain, where processes such as perception, decision-making, thought, judgement and imagination occur.

- The cerebrum is divided into nearly symmetrical left and right hemispheres by a deep groove, the *longitudinal fissure*. The hemispheres are connected by five commissures that span the longitudinal fissure, the largest of these is the corpus callosum. The surface of the brain is folded into ridges (*gyri*) and grooves (*sulci*), many of which are named, usually according to their position, such as the frontal gyrus of the frontal lobe or the central sulcus separating the central regions of the hemispheres (Fig. 1.3).

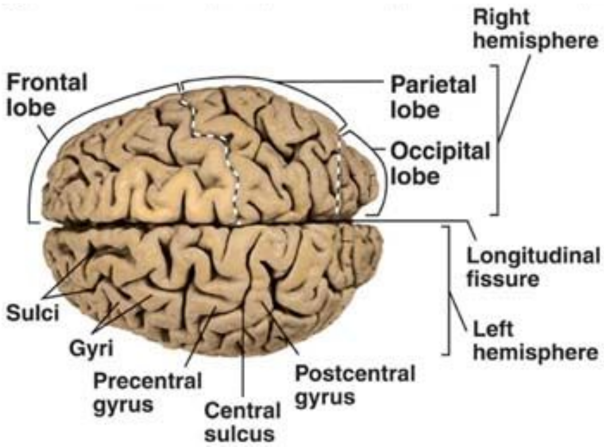


Figure 1.3: Superior view of the brain.

As seen in Fig. 1.3 each hemisphere is then conventionally divided into four lobes: the *frontal lobe, parietal lobe, temporal lobe* and *occipital lobe*, named according to the skull bones that overlie them. Though there is some functional overlap between the lobes, each of them is associated with one or two specialized functions, that can be summarized as follows and are represented in Fig. 1.4:

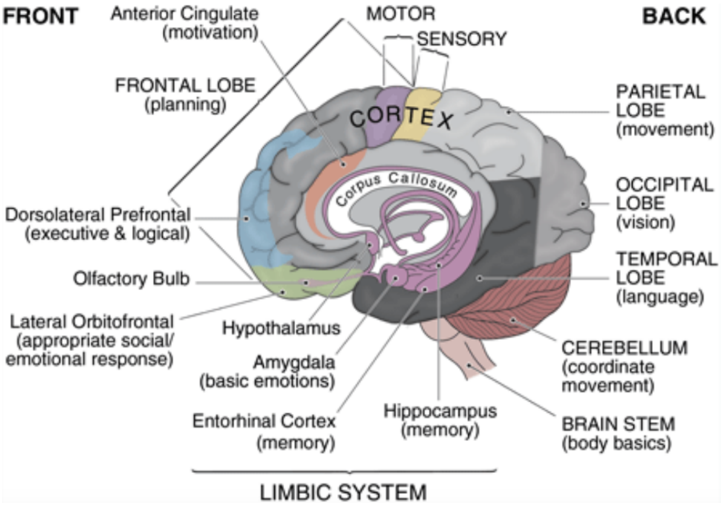

FRONT
Anterior Cingulate (motivation)
MOTOR
SENSORY
BACK
FRONTAL LOBE (planning)
CORTEX
PARIETAL LOBE (movement)
Corpus Callosum
OCCIPITAL LOBE (vision)
Dorsolateral Prefrontal (executive & logical)
TEMPORAL LOBE (language)
Olfactory Bulb
Lateral Orbitofrontal (appropriate social/ emotional response)
Hypothalamus
Amygdala (basic emotions)
CEREBELLUM (coordinate movement)
Entorhinal Cortex (memory)
Hippocampus (memory)
BRAIN STEM (body basics)
LIMBIC SYSTEM

Figure 1.4: Medial vision of the brain. In this figure the functions associated with each lobe and other parts of the brain are emphasized.

- The main functions of the frontal lobe are to control attention, abstract thinking, behavior, problem solving tasks, physical reactions and personality.
- The occipital lobe is the smallest lobe; its main functions are visual reception, visual-spatial processing, movement and color recognition.
- The temporal lobe controls auditory and visual memories, language and some hearing and speech. We then highlight how the inferior temporal lobe is involved in long term memory.
- The parietal lobe integrates sensory information among various modalities, in particular spatial sense and navigation. It includes also several areas important in language processing.

- The outer part of the cerebrum is the *cerebral cortex*, made up of grey matter arranged in layers. It is 2 to 4 millimeters thick and deeply folded to give a convoluted appearance. Its largest part is the *neocortex*, which has six neuronal layers, whereas the rest is called *allocortex*, which has three or four layers. Then, beneath the cortex there is the white matter of the brain.
  The cortex is mapped by divisions into about fifty different functional areas known as *Brodmann's areas*, that are distinctly different when seen under a microscope. They can then be grouped in the following main subdivisions according to their function:
  - The *primary motor cortex*, which sends axons down to motor neurons in the brainstem and spinal cord, occupies the rear portion of the frontal lobe, directly in front of the somatosensory area.
  - The *primary sensory areas* receive signals from the sensory nerves and the tracts by way of relay nuclei[1] in the thalamus. Primary sensory areas include the visual

[1] Relay nuclei represent the group of nuclei to which we normally refer when considering the thalamus as a

cortex of the occipital lobe, the auditory cortex in parts of the temporal lobe and insular cortex and the somatosensory cortex in the parietal lobe.

  - The remaining parts of the cortex are called the *association areas*. These areas receive input from the sensory areas and lower parts of the brain and are involved in the complex cognitive processes of perception, thought and decision-making.

- The cerebrum also contains several subcortical structures, including the *hippocampus* (involved in memory function), the *basal ganglia* (involved in coordination of movement), the *thalamus* (involved in the regulation of consciousness, sleep and alertness), the *hypothalamus* (involved in the control of certain metabolic processes and other autonomic activities such as body temperature, hunger thirst and circadian cycles) and the *olfactory bulb* (involved in smell). They can be grouped to form a collection of brain structures, such as the *diencephalon* or the *limbic system*. We now focus our attention on the latter, since it is connected to memory.

1.1.2 The Limbic system and Memory

The *limbic system* (also known as the "paleomammalian brain") is a collection of brain structures located in the middle of the brain, anatomically related but varying greatly in function. Collectively we can think at the limbic system as the centre for emotional responsiveness, motivation, memory formation and integration, olfaction and the mechanisms to keep ourselves safe. There exists no universal agreement on the entire list of structures composing the system, but the regions considered include the *olfactory bulbs*, *hippocampus*, *hypothalamus, amygdala, columns of fornix, mammillary body, septum, cingulate gyrus* and *entorhinal cortex*, as can be seen in Fig. 1.4 and more in detail in Fig. 1.5.

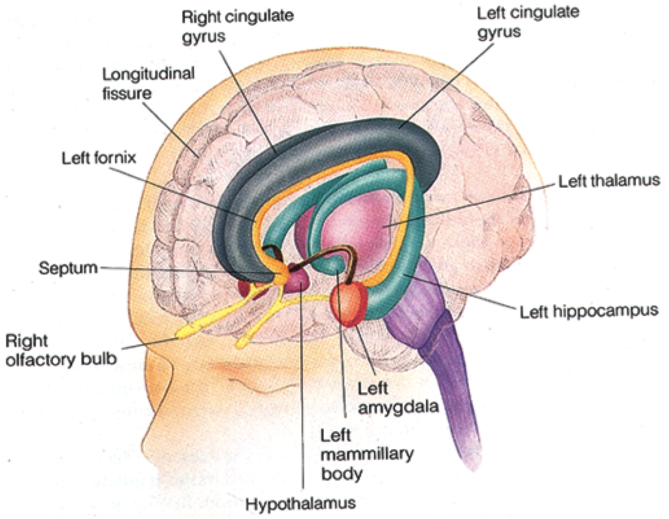


Figure 1.5: Major structure of the limbic system. Note how the olfactory bulbs feed directly into the amygdala, giving smell a particularly important role in emotional memory and evaluation of circumstances. It is also striking that the size of these structures has little correlation to their power and importance in this amazing system.

synaptic station for the sensory pathways. The specific relay nuclei project to a specific area of the cerebral cortex and in particular to the primary cortical areas.

One of the main function connected with the structures that are part of the limbic system is *memory*. In particular the components involved with the various storage and retrieval processes are: the amygdala, the hipppocampus and the entorhinal cortex, that we are now going to describe briefly.

- The *amygdala* is one of two almond-shaped groups of nuclei[2] located deep and medially within the temporal lobes of the brain in complex vertebrates, including humans. It performs a primary role in the processing of memory, decision-making and emotional responses (including fear, anxiety and aggression).

- The *hippocampus* is a major component of the brain of humans and other vertebrates. Humans and other mammals have two hippocampi, one in each side of the brain. It plays important roles in the consolidation of information from short-term memory to long-term memory and in spatial memory that enables navigation.
  The hippocampus contains two main interlocking parts: the *hippocampus proper* (also called *Ammon's horn*) and the *dentate gyrus*, as represented in Fig. 1.6.

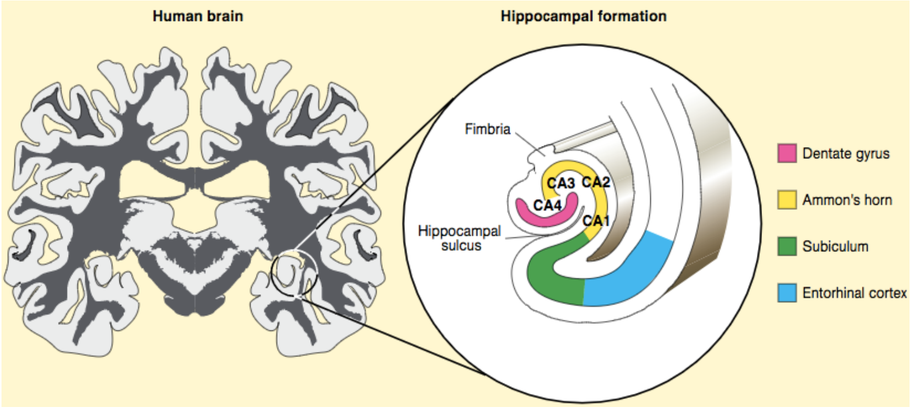


Figure 1.6: The hippocampal formation is a compound structure in the medial temporal lobe of the brain. There is no consensus concerning which brain regions are encompassed by the term, with some authors defining it as the dentate gyrus, the hippocampus proper and the subiculum; and others including also the entorhinal cortex. On the left of the figure, there is a transverse section of the human brain, indicating the location of the hippocampal formation. On the right, the topographical zones of the hippocampal formation are represented. Figure taken from Parkin's article [45].

  - The *hippocampus proper* refers to the actual structure of the hippocampus which is made up of four regions or subfields forming a neural circuit. The subfields CA1, CA2, CA3 and CA4 use the initials of *Cornu Ammonis*, an earlier name of the hippocampus. CA3 is then implicated in a number of working theories on memory and hippocampal learning processes, since its role in this process is still not clear.

  - The *dentate gyrus* DG is a part of the hippocampus (part of the hippocampal formation), that is thought to contribute to the formation of new episodic memories, the spontaneous exploration of novel environments and to play a role in depression.

[2] A *nucleus* (plural form: nuclei) is a cluster of neurons in the central nervous system, located deep within the cerebral hemispheres and brainstem.

- The *entorhinal cortex* (ento = interior, rhino = nose, entorhinal = interior to the rhinal sulcus) is an area of the brain located in the medial temporal lobe and functioning as a hub in a widespread network for memory and navigation. The EC is the main interface between the hippocampus and neocortex. Furthermore, the EC-hippocampus system plays an important role in declarative (autobiographical/episodic/semantic)[3] memories and in particular spatial memories including memory formation, memory consolidation and memory optimization in sleep.

Now that we have understood which parts of the brain are involved in the process of memory, we will present the classification of memories, that is usually done in the field of medicine and psychology, following [33]. In this discussion we will also describe more in detail how the different brain structures mentioned are related to memory processes and what role they play.

1.2 Memory

The ability to learn from the experience of memorizing what has been learned and recovering it when appropriate is one of the most surprising features of the brain, without which many cognitive operation would not be possible. *Memory* can thus be defined as the faculty of the mind by which information is encoded, stored and retrieved.
It can be divided into different categories. A first classification can be made at a qualitative level: in fact, there are two categories of qualitatively distinct memories:

- *Declarative memory* or *explicit memory* is the conscious, intentional recollection of factual information, previous experiences and concepts that are explicitly stored and retrieved in the brain. It includes factual knowledge concerning, for example, objects, people, names, numbers, events and their meaning, which can normally be expressed by language.
  Declarative memory can be further sub-divided into:
  - *semantic memory*, concerning principles and facts taken independent of context. It allows the encoding of abstract knowledge about the world, such as “Paris is the capital of France”.
  - *episodic memory*, concerning information specific to a particular context, such as time and place. It is used for more personal memories, such as sensations, emotions and personal associations with a particular place or time. Autobiographical memory, i.e. memory for particular events within one’s own life, is generally viewed as either equivalent to, or a subset of, episodic memory.
- *Non-declarative memory* or *implicit memory* is not based on the conscious recall of information, but on implicit learning. It concerns the ability to perform certain tasks, or to associate them with each other, whose recovery often happens in a non-conscious way. Research into this type of memory indicates that it operates through a different mental process from explicit memory.
  Implicit memory is usually subdivided in 4 subgroups:
  - *Procedural memory*, the slow and gradual learning of skills that often occurs without conscious attention to learning. In daily life, people rely on implicit

[3] We will discuss the differences about these types of memory in section 1.2.

memory every day in the form of procedural memory, the type of memory that allows people to remember how to tie their shoes or ride a bicycle without consciously thinking about these activities.

- *Priming memory* is the process of subliminally arousing specific responses from memory, as the completion of objects or words presented in an incomplete way.
- *Non-associative learning* refers to a relatively permanent change in the strength of response to a single stimulus due to repeated exposure to that stimulus.
- *Associative learning* is a learning process in which a new response becomes associated with a particular stimulus.

In addition to the qualitative categories just described, the memory can be subdivided in time categories (Fig. 1.7):

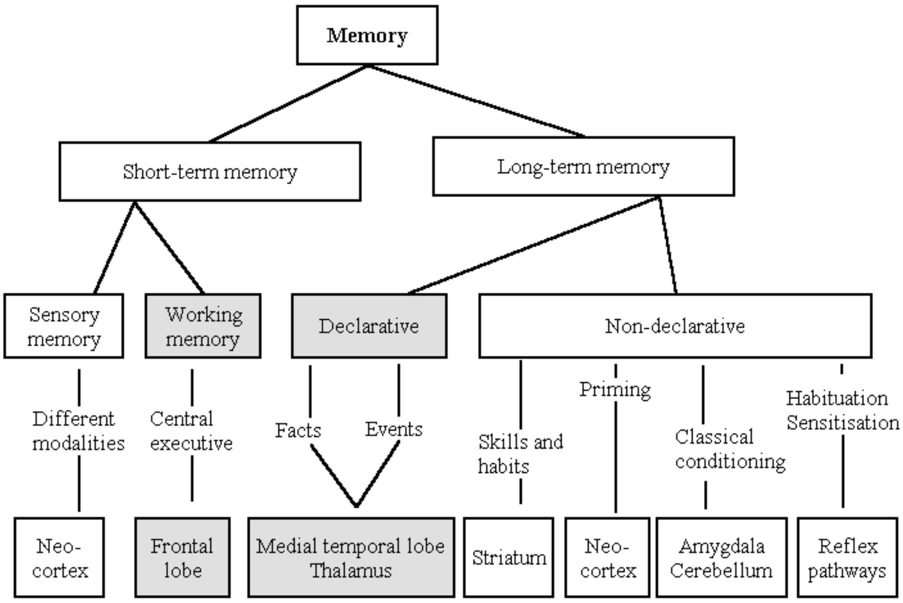


Figure 1.7: Schematic illustration of human memory processes and the anatomical structures involved. The medial temporal lobe includes a system of anatomically related structures that are essential for declarative memory (conscious memory for facts and events). The system consists of the hippocampal region (CA fields, dentate gyrus and subicular complex) and the adjacent entorhinal cortex. Then, the striatum is a part of the basal ganglia.

- *Short-term memory* allows recall of informations for a period of several seconds to a minute without rehearsal. It is subdivided into two main groups: *sensory memory* and *working memory.*
  - *Sensory memory* is our ability to briefly store informations related to the different senses, which have been perceived for a short durations. Each of the senses has a different “memory store” such as iconic (visual), echoic (audial) and haptic (touch). When we focus our attention on a sensory memory it moves into working memory. The capacity of this register is not very high and the information is constantly updated as the most recent information overwrites the oldest one. In fact, it holds sensory information less than one second after an item is perceived.

  - *Working memory*[4] is a low-capacity memory register, which is the bridge between the present and the most recent past. In fact, it is limited to holding between five and seven items in the mind at a time for up to about 30 seconds each. Working memory then represents the way we process the sensory information we are actively thinking about. Furthermore, it serves as an encoding and retrieval processor for informations that are stored and thus belong to long-term memory.

- *Long-term memory* is a high-capacity memory system, where information can be held indefinitely. As discussed before, the storage in short-term memory generally has a strictly limited capacity and duration, which means that information is not retained indefinitely. By contrast, long-term memory can store much larger quantities of information for potentially unlimited duration (sometimes a whole life span) since it has an immeasurable capacity.
  The process of storing new information in long-term memory is called *consolidation*. Consolidation is stimulated by the recirculation of information in the memory (exercise, repetition) and is particularly facilitated when the facts are related to categories of information already present in the memory (association). Then, the process that transfers information into a lasting form is called *encoding* and the physical representation of the memory of the corresponding sensory event ("memory trace") is the *engram*.
  There are then two types of long-term memory: declarative and non-declarative memory, which have been presented previously.

Short-term memory is supported by transient patterns of neuronal communication, dependent on regions of the frontal lobe (especially dorsolateral prefrontal cortex) and of the neo-cortex. Long-term memory, on the other hand, is maintained by more stable and permanent changes in neural connections widely spread throughout the brain. The hippocampus, for example, is essential for learning new information and for consolidating information from short-term to long-term memory, although it does not seem to store information itself.

1.2.1 The multi-store model

In section 1.2 we have presented the classification that can be done for the different types of memory and in particular outlined the temporal process related to the storage and retrieval of a memory, that we are now going to explain better following the contribute of Atkinson and Shiffrin [8].
The *multi-store model* (also known as *Atkinson-Shiffrin memory model*) describes the permanent, structural features of the memory system. The basic structural division is into the three components diagrammed in Fig. 1.8: the sensory, the working and the long-term memory.

- When a stimulus is presented, there is an immediate registration of that stimulus within the appropriate sensory dimensions. Our brain retains sensory memory information for a very short period of time. So we have to decide (consciously or unconsciously) which of this information can catch our attention and be stored in working memory, the others will be forgotten.

- A stimulus must first be perceived, then in order for the information to be further processed, attention is key. *Attention* is a cognitive process that allows us to focus on particular environmental stimuli. Once perceived, paying attention to something allows

[4] Sometimes short-term memory is used as a synonymous for working memory, so there are three different types of memory: sensory, working and long-term memory. In what follows we will use this assumption, especially in figures that are taken from different sources.

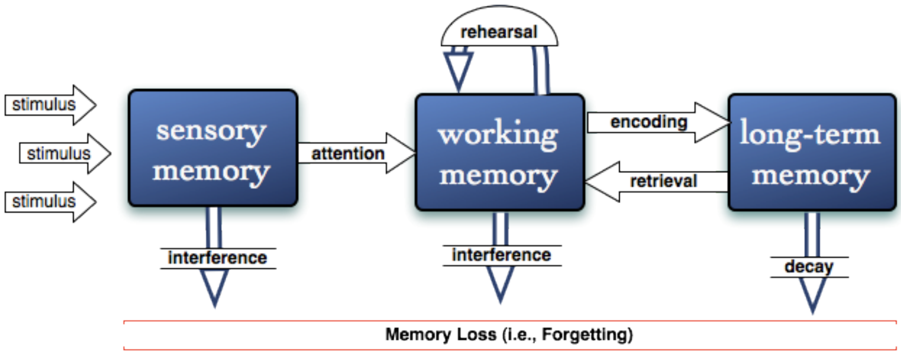


Figure 1.8: Structure of the memory system based on the multi-store model of Atkinson and Shiffrin [8].

information to pass from sensory memory into working memory. Thus, attention serves as a filter for stimuli from our environment. By selectively determining what will "get through" for further examination and what will not, attention allows us to focus only on the necessary stimuli.

- The second basic component of our system is the short-term store (STS) or working memory. The character of the information in the working memory does not depend necessarily upon the form of the sensory input (for example if it is an image or a sound). We will use the abbreviation a-v-l to stand for *auditory-verbal-linguistic store*, a triple term introduced because it is not easy to separate these three functions.
  Information entering the short-term store is assumed to decay and disappear completely, but the time required for the information to be lost is considerably longer than for the sensory register. The exact rate of decay of information in the short-term store is difficult to estimate because it is greatly influenced by subject-controlled processes. In the a-v-l mode, for example, the subject can invoke rehearsal mechanisms that maintain the information in STS and thereby complicate the problem of measuring the structural characteristics of the decay process. However, the available evidence suggests that information represented in the a-v-l mode decays and is lost within a period of about 15-30 seconds. Storage of information in other modalities is less well understood and thus it is difficult to assign values to their decay rates.

- Short-term memories can become long-term memory through the process of *consolidation*, involving rehearsal and meaningful association. Unlike short-term memory (which relies mostly on an acoustic, and to a lesser extent a visual, code for storing information), long-term memory encodes information for storage semantically (i.e. based on meaning and association). However, there is also some evidence that long-term memory does also encode to some extent by sound.
  Therefore, a process that is fundamental in the transition from working memory to long term is *encoding*. Encoding is the first crucial step to create a new memory. It allows the perceived item of interest to be converted into a construct that can be stored within the brain and then recalled later from long-term memory.

- The last major component of our system is the long-term store. This store differs from the preceding ones in that information stored here does not decay and become lost in the same manner. All information eventually is completely lost from the sensory register and the short-term store, whereas information in the long-term store is relatively

permanent (although it may be modified or made temporarily irretrievable as the result of other incoming information). *Forgetting* then occurs in long-term memory when the formerly strengthened synaptic connections among the neurons in a neural network become weakened, or when the activation of a new network is superimposed over an older one, thus causing interference in the older memory.
Most experiments in the literature dealing with long-term store have been concerned with storage in the a-v-l mode, but it is clear that there is long-term memory in each of the other sensory modalities, as demonstrated by an ability to recognize stimuli presented to these senses. There may even be information in the long-term store which is not classifiable into any of the sensory modalities, the prime example being temporal memory.

- After an item has been stored, we may want to retrieve it in a future moment. The connected process is called *recall* or *retrieval* of memory. It refers to the subsequent re-accessing of events or information from the past, which have been previously encoded and stored in the brain. In common parlance, it is known as remembering.

1.2.2 Memory processes

In the previous section, we have discussed the different stages of memory formation (from perception to sensory memory, then to short-term memory and finally to long-term memory). We now want to look at the overall processes involved.
Memory is the ability to encode, store and recall information. The three main processes involved in human memory are therefore *encoding, storage* and *recall (retrieval)*, as summarized in Fig. 1.9. Additionally, the process of memory consolidation (which can be considered to be either part of the encoding process or the storage process) is treated here as a separate process in its own right. Some of the physiological and neurological concepts involved in these processes are highly complex and technical (and some of them still not completely understood) and lie largely outside the remit of this thesis, although at least a general introduction is given here.

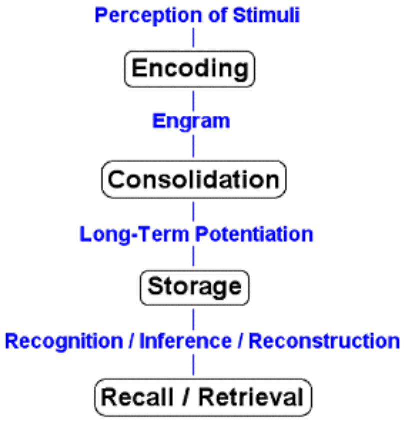


Figure 1.9: Fundamental memory processes.

Memory Encoding

Encoding is a biological event beginning with perception through the senses. The process of laying down a memory begins with attention (regulated by the thalamus and the frontal lobe), in which a memorable event causes neurons to fire more frequently, making the experience more intense and increasing the likelihood that the event is encoded as a memory. Emotion tends to increase attention and the emotional element of an event is processed on an unconscious pathway in the brain leading to the amygdala, where the actual sensations derived from an event are processed.

The perceived sensations are then decoded in the various sensory areas of the cortex and then combined in the brain's hippocampus into one single experience. The hippocampus is thus responsible for analyzing these inputs and ultimately deciding if they will be committed to long-term memory. It acts as a kind of sorting centre where the new sensations are compared and associated with previously recorded ones. The various threads of information are then stored in various different parts of the brain, although the exact way in which these pieces are identified and recalled later remains largely unknown. The key role that the hippocampus plays in memory encoding has been highlighted by examples of individuals who have had their hippocampus damaged or removed and can no longer create new memories (as the case of patient HM, that is treated in detail for example in Quiroga's book [50]). It is also one of the few areas of the brain where completely new neurons can grow.

Although the exact mechanism is not completely understood, encoding occurs on different levels, the first step being the formation of short-term memory from the ultra-short term sensory memory, followed by the conversion to a long-term memory by a process of memory consolidation. The process begins with the creation of a memory trace or engram in response to the external stimuli. An engram is a hypothetical biophysical or biochemical change in the neurons of the brain, hypothetical in the sense that no-one has ever actually seen, or even proved, the existence of such a construct. The hippocampus then receives connections from the primary sensory areas of the cortex, as well as from associative areas and the rhinal and entorhinal cortexes. While these anterograde connections converge at the hippocampus, other retrograde pathways emerge from it, returning to the primary cortexes. Thus, a neural network of cortical synapses effectively records the various associations which are linked to the individual memory.

There are three or four main types of encoding:

- *Acoustic encoding* is the processing and encoding of sound, words and other auditory input for storage and later retrieval. This is aided by the concept of the phonological loop, which allows input within our echoic memory to be sub-vocally rehearsed in order to facilitate remembering.

- *Visual encoding* is the process of encoding images and visual sensory information. Visual sensory information is temporarily stored within the iconic memory before being encoded into long-term storage. The amygdala fulfills an important role in visual encoding, as it accepts visual input in addition to input from other systems and encodes the positive or negative values of conditioned stimuli[5]. It seems in fact that distinct neurons respond to positive and negative stimuli, but there is no clustering of these distinct neurons into clear anatomical nuclei.

[5] *Classical* or *Pavlovian conditioning* is an example of associative learning, in which a neutral stimulus (e.g. a luminous stimulus) is paired with a significant stimulus (*unconditioned stimulus*, e.g. a high intensity noise that causes fright). After a series of sessions of learning, the neutral stimulus alone (now become *conditional stimulus*) triggers an answer (in the previous example the light stimulus induces a fear response). We then say that the conditioned stimuli are positive (negative) if they evoke positive (negative) feelings.

- *Tactile encoding* is the encoding of how something is perceived, normally through the sense of touch. Physiologically, neurons in the primary somatosensory cortex of the brain react to vibrotactile stimuli caused by the feel of an object.
- *Semantic encoding* is the process of encoding sensory input that has a particular meaning or can be applied to a particular context, rather than deriving from a particular sense.

It is believed that, in general, encoding for short-term memory storage in the brain relies primarily on acoustic encoding, while encoding for long-term storage is more reliant (although not exclusively) on semantic encoding.

Human memory is fundamentally associative, meaning that a new piece of information is remembered better if it can be associated with previously acquired knowledge that is already firmly anchored in memory. The more personally meaningful the association, the more effective the encoding and consolidation. Elaborate processing, that emphasizes familiar meaning and associations, tends to lead to improved recall. On the other hand, information that a person finds difficult to understand cannot be readily associated with already acquired knowledge and so will usually be poorly remembered, or may even be remembered in a distorted form due to the effort to comprehend its meaning and associations.

Because of the associative nature of memory, encoding can be improved by a strategy of organization of memory called *elaboration*, in which new pieces of information are associated with other information already recorded in long-term memory, thus incorporating them into a broader, coherent narrative which is already familiar. An example of this kind of elaboration is the use of mnemonics, which are verbal, visual or auditory associations with other easy-to-remember constructs, which can then be related back to the data that is to be remembered. When we use mnemonic devices, we are effectively passing facts through the hippocampus several times, so that it can keep strengthening the associations and therefore improve the likelihood of subsequent memory recall.

Memory Consolidation and Synaptic Plasticity

Consolidation is the process of stabilizing a memory trace after the initial acquisition. It may perhaps be thought as a part of the process of encoding or of storage, or it may be considered as a memory process in its own right. It is usually considered to consist of two specific steps: *synaptic consolidation* (which occurs within the first few hours after learning or encoding) and *system consolidation* (where hippocampus-dependent memories become independent of the hippocampus over a period of weeks to years). In consolidation are involved also other mechanisms, that can reinforce or weaken it:

- Neurologically, the process of consolidation utilizes a phenomenon called *long-term potentiation*, which allows a synapse to increase in strength as increasing numbers of signals are transmitted between the two neurons. *Potentiation* is the process by which synchronous firing of neurons makes those neurons more inclined to fire together in the future. Thus, long-term potentiation occurs when the same group of neurons fire together so often that they become permanently sensitized to each other.

  At its simplest, whenever something is learned, circuits of neurons in the brain, known as *neural networks*, are created, altered or strengthened. These neural circuits are composed of a number of neurons that communicate with one another through special junctions called *synapses*. Through a process involving the creation of new proteins within the body of neurons and the electrochemical transfer of neurotransmitters across synapse gaps to receptors, the communicative strength of certain circuits of neurons in the brain is reinforced. With repeated use the efficiency of these synaptic connections

increases, facilitating the passage of nerve impulses along particular neural circuits, which may involve many connections to the visual cortex, the auditory cortex, the associative regions of the cortex, etc. In this way, the brain organizes and reorganizes itself in response to experiences, creating new memories prompted by experience, education or training. The ability of the connection, or synapse, between two neurons to change in strength, and for lasting changes to occur in the efficiency of synaptic transmission, is known as *synaptic plasticity* or *neural plasticity* and it is one of the most important neurochemical foundations of memory and learning.
It should be remembered that each neuron makes thousands of connections with other neurons and that memories and neural connections are mutually interconnected in extremely complex ways. Unlike the functioning of a computer, each memory is embedded in many connections and each connection is involved in several memories. Thus, multiple memories may be encoded within a single neural network by different patterns of synaptic connections. Conversely, a single memory may involve simultaneously the activation of different groups of neurons in completely different parts of the brain.

- The inverse of long-term potentiation, known as *long-term depression*, can also take place, whereby the neural networks involved in erroneous movements are inhibited by the silencing of their synaptic connections. This can occur in the cerebellum, which is located towards the back of the brain, in order to correct our motor procedures when learning how to perform a task (procedural memory), but also in the synapses of the cortex, the hippocampus, the striatum and other memory-related structures. Contrary to long-term potentiation, which is triggered by high-frequency stimulation of the synapses, long-term depression is produced by nerve impulses reaching the synapses at very low frequencies, leading them to undergo the reverse transformation from long-term potentiation and, instead of becoming more efficient, the synaptic connections are weakened. It is still not clear whether long-term depression contributes directly to the storage of memories in some way, or whether it simply makes us forget the traces of some things learned long ago so that new things can be learned.

- *Sleep* (particularly slow-wave or deep sleep during the first few hours) is also thought to be important in improving the consolidation of information in memory and the activation of patterns in the sleeping brain, which mirror those recorded during the learning of tasks from the previous day, suggesting that new memories may be solidified through such reactivation and rehearsal[6].

- Another important process that takes place is that of re-consolidation. *Memory re-consolidation* is the process of previously consolidated memories being recalled and then actively consolidated all over again in order to maintain, strengthen and modify memories that are already stored in long-term memory. Several retrievals of memory may be needed for these memories to last for many years, depending on the depth of the initial processing. The very act of re-consolidation, though, may change the initial memory. As a particular memory trace is reactivated, the strengths of the neural connections may change, the memory may become associated with new emotional or environmental conditions or subsequently acquired knowledge, expectations rather than actual events may become incorporated into the memory, etc.

[6] The interested reader can find more explanations in [51].

Memory Storage

Storage is more or less a passive process of retaining information in the brain, either in sensory memory, working memory or the more permanent long-term memory. Each of these different stages of human memory functions as a sort of filter that helps to protect us from the flood of information that confront us on a daily basis, avoiding an overload of information and helping to keep us sane. As previously discussed, the more the information is repeated or used, the more likely it is to be retained in long-term memory through consolidation.
Since the early neurological work of Karl Lashley and Wilder Penfield in the 1950s and 1960s, it has become clear that long-term memories are not stored in just one part of the brain, but are widely distributed throughout the cortex. After consolidation, long-term memories are stored throughout the brain as groups of neurons that are primed to fire together in the same pattern that created the original experience and each component of a memory is stored in the brain area that initiated it (e.g. groups of neurons in the visual cortex store a sight, neurons in the amygdala store the associated emotion, etc). Indeed, it seems that they may even be encoded redundantly, several times, in various parts of the cortex, so that, if one engram (or memory trace) is wiped out, there are duplicates, or alternative pathways, elsewhere, through which the memory may still be retrieved. Therefore, contrary to the popular notion, memories are not stored in our brain like books on library shelves, but must be actively reconstructed from elements scattered throughout various areas of the brain by the encoding process. Memory storage is therefore an ongoing process of reclassification resulting from continuous changes in our neural pathways and parallel processing of information.
In the absence of disorders due to trauma or neurological disease, the human brain has the capacity to store almost unlimited amounts of information indefinitely. *Forgetting*, therefore, is more likely to be the result from incorrectly or incompletely encoded memories and/or problems with the recall/retrieval process. It is a common experience that we may try to remember something one time and fail, but then remember that same item later. The information is therefore clearly still there in storage, but there may have been some kind of a mismatch between retrieval cues and the original encoding of the information. Hence, forgetting can be better thought of as the temporary or permanent inability to retrieve a piece of information or a memory that had previously been recorded in the brain. It typically follows a logarithmic curve, so that information loss is quite rapid at the start, but becomes slower as time goes on. However, theorists disagree over exactly what becomes material that is forgotten. Some hold that long-term memories do actually decay and disappear completely over time; others hold that the memory trace remains intact as long as we live, but the bonds or cues that allow us to retrieve the trace become broken, due to changes in the organization of the neural network, new experiences, etc.

Memory Retrieval

There are two main methods of accessing memory: recognition and recall.
Recognition is the association of an event or physical object with one previously experienced or encountered and involves a process of comparison of information with memory, e.g. recognizing a known face, true/false or multiple choice questions, etc. Recognition is a largely unconscious process and the brain even has a dedicated face-recognition area, which passes information directly through the limbic areas to generate a sense of familiarity, before linking up with the cortical path, where data about the person's movements and intentions are processed.
Recall involves remembering a fact, event or object that is not currently physically present (in the sense of retrieving a representation, mental image or concept) and requires the direct

uncovering of information from memory, e.g. remembering the name of a recognized person, fill-in the blank questions, etc. Memories then are not frozen in time and new information and suggestions may become incorporated into old memories over time. Thus, remembering can be thought of as an act of creative re-imagination. Because of the way memories are encoded and stored, memory recall is effectively an on-the-fly reconstruction of elements scattered throughout various areas of our brain. Memory retrieval therefore requires re-visiting the nerve pathways the brain formed when encoding the memory and the strength of those pathways determines how quickly the memory can be recalled. The information pattern is then re-stored back in the memory, thus re-consolidating and strengthening it.

We have presented a biological model for storing and retrieving memories emphasizing the parts of the brain involved in the various processes. Unfortunately, we still do not know a reliable model regarding many aspects of memory, as has already been mentioned. For example, it is not yet clear how the hippocampus influences the formation of memories and how its components CA1 and CA3 act for this purpose. Neuroscientists are doing a lot of studies about it, as Kesner and Rolls [32] and Treves [62], but they are still far from achieving complete knowledge. Obviously, if everything were clear, it would be even simpler to create a biologically plausible mathematical model that represents the human brain and in particular the mechanism of memory. For these reasons, in chapter 2 we are going to introduce a very simplified model for this dynamics, which does not take into account the difference between short-term and long-term memory or other properties whose operation has not yet been fully understood. We will therefore restrict ourselves to presenting a basic model for this dynamics that maintains its salient characteristics, such as synaptic plasticity, the ability to remember and store memories, but without the complications of the biological model. At the end of chapter 2, various studies will be presented to extend Hopfield model to include some missing aspects or improvements, but there is still no global mathematical model that includes everything that has been discussed in this chapter.

1.3 Neurons and Synapses

The struggle to understand the brain has been made easier because of the pioneering work of Ramón y Cajál in 1911, who introduced the idea of neurons as structural constituents of the brain. Then, in 1990 it has been estimated by Shepherd and Koch that there are approximately 10 billion neurons in the human cortex and 60 trillion synapses or connections. *Neurons* are the basic units that make up the brain and nervous system. They are specialized cells that act like telegraph wires carrying messages in the form of electrochemical impulses throughout the body. Though there are about 20 different types of neurons, their structure is basically the same. A neuron is composed of its *soma*, its *dendrites* that quite often form a treelike structure and the *axon* that, eventually, branches (Fig. 1.10). Information produced in other neurons is transferred to the neuron under consideration by means of localized contacts, the *synapses*, that are located on the dendrites and also on the cell body. Electrical charges produced at the synapses propagate to the soma and produce a *net postsynaptic potential.* If the postsynaptic potential at the soma is sufficiently large to exceed a threshold value, typically a depolarization of $10 - 15$ mV, the neuron generates a brief electrical pulse, called a *spike or action potential*, at its axon hillock, that is the point of connection between the soma and the axon. The spikes run down the axon, finally reach the synapses that, in a way to be discussed below, transfer the information to another neuron. Using the terminology of electronic circuits we can therefore say that the dendrites represent the input device

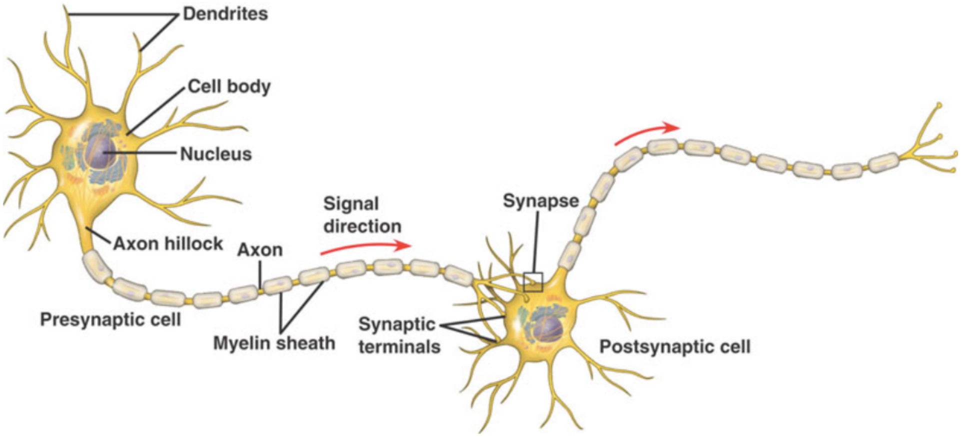


Figure 1.10: Sketch of two neurons in synaptic contact.

("input"), the axons the output device ("output") and the soma the information processing unit.

In order to be able to model the functioning of a neuron we have to deal with these processes in more detail. In this chapter we will be satisfied by a qualitative discussion, whereas a mathematical discussion will be developed in the next chapters.

1.3.1 The action potential

The *membrane potential* is defined as the potential difference measured at the ends of two electrodes, one placed inside the neuronal cell and one placed in the surrounding extracellular fluid. When we talk about the *neuronal signal* we refer to the temporal and spatial variation of the membrane potential. When the neuron is at rest, that is, it is not in some way excited from the outside, the membrane potential assumes a characteristic value called *resting potential*, typically in the order of -65 mV , i.e. the inside of the cell is at a lower potential than outside.

Action potentials are typical voltage pulses[7] generated during neuronal dynamics; they have a nearly stereotypical form and are not subject to attention or distortion during propagation along the axon. Fig. 1.11 shows the typical form of an action potential. Note the following characteristics:

- The voltage pulse has a duration of about $1-2$ ms and an amplitude measured between the minimum and the maximum of about $100-120$ mV;
- In the first phase of the impulse there is a vigorous growth of the membrane potential up to the so called *depolarization phase* where the membrane potential becomes positive, i.e. the inside of the cell is at a higher potential than the external;
- During the descent phase, the impulse before returning to the resting value goes through a phase called *hyperpolarization*, typically lasting about 10 ms (and therefore much slower than depolarization), in which the cell has a membrane potential lower than that at rest.

[7] The voltage pulse represents the cycle-time of a biological neuron, i.e. the time the spike travels the full length of the pre-synaptic axon, the neurotransmitter crosses the synaptic gap and the post-synaptic potential diffuses the soma.

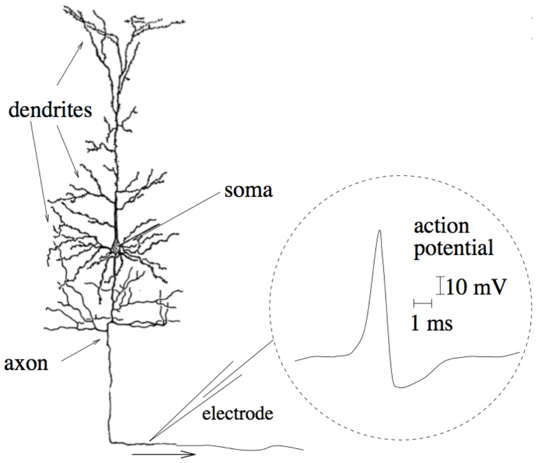


Figure 1.11: Single neuron in a drawing by Ramoń y Cajal, where the inset shows an example of a neuronal action potential. The action potential is a short voltage pulse of $1-2$ ms duration and an amplitude of about 100 mV. The figure is taken from Gersten and Kistler's book, [21].

The action potential, once it is generated, travels along the axon and is transmitted to the other neurons. Thus it constitutes the elementary unit associated with the transmission of neuronal signals. Typically therefore when referring to the signal emitted by a neuron we mean the temporal sequence of these action potentials, also called *spike train.*

1.3.2 Synapses

Synapses, or *nerve endings*, are the elementary structural and functional units that mediate the interactions between neurons. In this context, the *presynaptic neuron* is defined as the neuron "transmitting" the action potentials upstream of the synapse and the *postsynaptic neuron* as the neuron "receiving" the action potentials downstream of the synapse. With this terminology the synapse is thus the region in which the axon of the presynaptic neuron "interacts" with the postsynaptic neuron dendrite and the *post-synaptic potential* (which has the acronym PPS) is defined as the voltage response of the postsynaptic neuron following the arrival of the action potential coming from the presynaptic neuron.
There are essentially two types of synapses: the chemical synapse and the electrical synapse.

- The *chemical synapse*, which is shown in the diagram in Fig. 1.12, is the most common in the vertebrate brain and is based on the mechanism that we are now going to describe. The presynaptic neuron generates an action potential, that when it reaches the end of the axon, locally depolarizes the cell membrane causing the release within the synaptic fissure (i.e. the small gap between the two presynaptic and postsynaptic cell membranes) of particular chemical substances called *neurotransmitters.* The neurotransmitter, as soon as it reaches the postsynaptic side of the synapse, is revealed by special molecules (*chemoreceptors*) placed on the postsynaptic membrane that cause the opening (or directly or through a chain of biochemical signals) of specific channels through which an ionic current flows from the extracellular fluid to the cell. The entry of these ions in turn leads to a change in the value of the postsynaptic membrane potential. So in a chemical synapse there is first the transformation of an electrical signal into a chemical signal on the presynaptic membrane and then the subsequent transformation on the postsynaptic membrane of a chemical signal into an electrical signal.

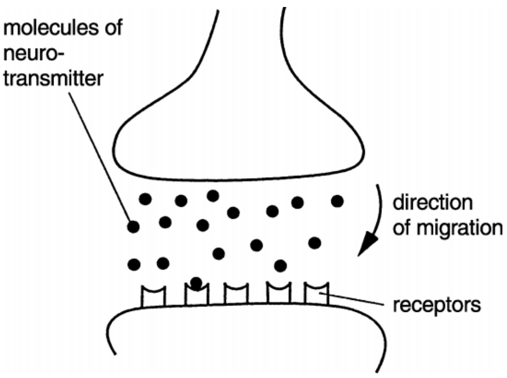


Figure 1.12: Scheme of a synapse. Figure taken from Haken's book, [24]

- The electric synapse instead realizes an electrical coupling between two neurons through highly specialized ion channels (called *gap-junctions*) that connect the presynaptic and postsynaptic membrane. The electric synapse allows a direct current flow between adjacent neurons.

Excitatory and inhibitory synapses

We have already mentioned that the arrival of an action potential from the presynaptic neuron causes a voltage response (the postsynaptic potential) in the membrane potential of the receiving neuron. In this regard, we distinguish between *excitatory postsynaptic potential* (which has the acronym PPSE) and *inhibitory postsynaptic potential* (which has the acronym PPSI) depending on whether the effect is to increase or decrease the value of the membrane potential. A similar meaning has the distinction between *excitatory synapses* and *inhibitory synapses*, or between a depolarizing stimulus and a hyperpolarizing stimulus.

From biological studies it is known that each neuron can have only one type of synapse: either all excitatory or all inhibitory, while it can receive action potentials from synapses of any kind. The difference between the two types of synapses resides in the post-synaptic receptors, which bind different neurotransmitters and allow different ions to pass through their channels: sodium Na^+ and potassium K^+ for the excitatory synapses, chlorine Cl^+ for the inhibitory ones. Therefore the opening of the channels of a determined ion allows to bring the membrane potential towards the equilibrium potential of that specific ion, that is -65 mV for Cl^+ and 0 mV for Na^+ and K^+.

CHAPTER 2 Hopfield Model

This chapter represents the main part of the thesis, where we will present the model of Hopfield and the various strategies that can be implemented to make it more biologically plausible and to study some of its salient characteristics.

2.1 Towards the Hopfield model

The modern era of neural networks began with the pioneering work of McCullogh and Pitts published in 1943 [38]. McCulloch was a psychiatric and neuroanatomist by training, while Pitts was a mathematical prodigy. In their paper they have described a logical calculus of neural networks that united the studies of neurophysiology and mathematical logic. Their formal model of a neuron was assumed to follow an "all-or-none" law. With a sufficient number of such simple units and synaptic connections set properly and operating synchronously, McCulloch and Pitts showed that a network so constituted would compute any computable function. This was a very significant result and it is generally agreed that the disciplines of neural networks and of artificial intelligence were born with it.
The next major development in neural networks came in 1949 with the publication of Hebb's book *"The Organization of Behavior"* [26], in which an explicit statement of a physiological learning rule for synaptic modification was presented for the first time. Specifically, Hebb proposed that the connectivity of the brain is continually changing as an organism learns different functional tasks and that neural assemblies are created by such changes. He followed up an early suggestions by Ramón y Cajál and introduced his now famous *postulate of learning*, which states that the effectiveness of a variable synapse between two neurons is increased by the repeated activation of one neuron by the other across that synapse. For this reason, Hebb's book has been a source of inspiration for the development of computational models of *learning and adaptive systems.* The paper by Rochester, Holland, Habit and Duda (1956) is the first attempt to use computer simulation to test a well-formulated neural theory based on Hebb's postulate of learning. In that same year, Uttley demonstrated that a neural network with modifiable synapses may learn to classify simple sets of binary patterns into corresponding classes, introducing the so-called *leaky integrate and fire neuron* model.
In the 1950s a work on *associative memory* was initiated by Taylor. Then, this was followed by the introducing of the *learning matrix* by Steinbuch (1961); this matrix consists of a planar network of switches interposed between arrays of sensory receptors and motor effectors. Other significant contributions to the early development of associative memory include papers by Anderson (1972), Kohonen (1972) and Nakano (1972), who independently and in the same year introduced the idea of a *correlation matrix memory* based on the outer product learning rule, of which Hebb's rule is an example.
Some 15 years after the publication of McCulloch and Pitt's classic paper, a new approach to the pattern recognition problem was introduced by Rosenblatt (1958) in his work on *perceptron*, a novel method of supervised learning. In the 1960s it seemed as if neural networks could do anything, thanks also to the crowning achievement of Rosemblatt, the *perceptron convergence theorem.* But then came the book of Minsky and Papert (1969), who used mathematics to demonstrate that there are fundamental limits on what single-layer

perceptrons can compute. They even stated that there was no reason to assume that any of the limitations of single-layer perceptrons could overcome in the multilayer version.

All these problems that emerged with perceptrons contributed to the dampening of continued interest in neural networks in the 1970s. Many of the researchers, except for those in psychology and the neurosciences, deserted the field during that decade.

In the 1980s major contributions to the theory and design of neural networks were made on several fronts and with it there was a resurgence of interest in neural networks.

In 1982 Hopfield used the idea of an energy function to formulate a new way of understanding the computation performed by recurrent networks with symmetric synaptic connections. Moreover, he established the isomorphism between such a recurrent network and the *Ising model* used in statistical physics. This analogy paved the way for a deluge of physical theory to enter neural modeling, thereby transforming the field of neural networks. This particular class of neural networks with feedback attracted a great deal of attention in the 1980s and in the course of time it has come to be known as *Hopfield neworks*. Although Hopfield networks may not be realistic model for neurobiological system, the principle they embody, namely that of storing information in dynamically stable networks, is profound.

In 1983 Cohen and Grossberg established a general principle for assessing the stability of a *content addressable memory* that includes the continuous-time version of the Hopfield network as a special case. A distinctive feature of an attractor neural network is the natural way in which time, an essential dimension of learning, manifests itself in the nonlinear dynamics of the network.

In the same year Kirkpatrick, Gelatt and Vecchi described a new procedure called *simulated annealing* for solving combinatorial optimization problems. This idea was later developed by Ackley, Hinton and Sejnowski (1985) in the development of a stochastic machine known as the *Boltzmann machine*, which was the first successful realization of a multilayer neural network.

In 1986 the development of the *back-propagation algorithm* was reported by Rumelhart, Hinton and Williams. In that same year, the celebrated two-volume book, *"Parallel Distributed Processing: Explorations in the Microstructures of Cognition"* edited by Rumelhart and McLelland, was published. This latter book has been a major influence in the use of back-propagation learning, which has emerged as the most popular learning algorithm for the training of multilayer perceptrons.

More than any other publications, the 1982 paper by Hopfield and the 1986 two-volume book by Rumelhart and McLelland were the most influential publications responsible for the resurgence of interest in neural networks in the 1980s. Neural networks have certainly come a long way from the early days of McCulloch and Pitts. Indeed, they have stablished themselves as an interdisciplinary subject with deep roots in neuroscience, psychology, mathematics, physical sciences and engineering.

We will now focus our attention on the study of the Hopfield Model, one of the most important examples of *recurrent neural network* (RNN).

2.2 Hopfield model: an example of Content Addressable Memory

The Hopfield model describes a system of a large number[8] of highly interconnected neurons, that have useful collective computational properties. The auto-associative memory model proposed by Hopfield, [28], has attracted considerable interest both as a content addressable memory and as a method of solving difficult optimization problems. *Content addressable memory* (CAM) is indeed one of the simplest collective properties of such a system and describes the capability of retrieving the entire memory (the memorized pattern) on the basis of partial information (a stimulus).
The model consists of a certain number, N, of neurons, which are the processing elements, and a large number of connections having fixed weights between these neurons. In general, the i-th processing element or neuron is described by two variables: its "mean soma potential" denoted by u_i and its output denoted by V_i. The output is usually related to the variable u_i by a simple nondecreasing monotonic output function g. This function is normally designed to limit the possible values of V_i to the range -1 to $+1$: hence the function will be nonlinear. For simplicity, g is frequently a step function or a hyperbolic tangent.
Each neuron has feedback connections to the output of all the other neurons as well as itself and its current state is a weighted sum of all the outputs of the previous instant. Thus the output of the i-th neuron is fed to the input of the j-th neuron by a connection of strength T_{ij}. In addition, each neuron has an offset bias I fed to its input.
There are two variants of the Hopfield model (Hopfield, [29]):

- *Discrete time version*: The dynamics of the network is described by the following difference equation
$$V(t+1) = f(TV(t) + I) \tag{2.1}$$
where $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is a sigmoid type nonlinearity or *sign* function. We note that f acts on every component of its argument.

- *Continuous time version*: The dynamics of the network is described by the following differential equation
$$\dot{u} = -\frac{u}{\tau} + Tg(u) + I \tag{2.2}$$
where $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a sigmoid type nonlinearity or sign function and $\tau = CR$.

We will start by describing the discrete version of the model, following Hopfield article of 1982 [28], and then we will see how the relevant properties are also found in the continuous case, as done by Hopfield in [29].

2.2.1 The discrete Hopfield model

Consider a set of N pairwise connected neurons described by their states V_i, $i = 1, \ldots, N$, where any V_i can take two values[9]: $V_i = -1$ when it is "not firing" and $V_i = +1$ if it is "firing

[8] In the following analysis this number will be infinitely large. The results approximate well even networks with a couple of hundred neurons, that are indistinguishable from simulations on networks with a couple of thousand neurons (Amit, [4]).

[9] In the original model of Hopfield the possible states of the neurons are 0 and 1, but we have decided to use as states -1 and $+1$ for simplicity and because this choice is widely adopted. It is easy to prove that the two approaches are equivalent and linked trough the formula $V_i = 2\sigma_i - 1$, where σ_i are the components of the vector in the binary notation $\{0, 1\}$ and V_i in the ± 1 notation.

at maximum rate". Thus the term *discrete* refers to the fact that neurons take on either binary or bipolar values. The instantaneous state of the system is specified by listing the N values of V_i, $i = 1, \ldots, N$, so it is represented by a bipolar word of N bits.

We assume the network to be fully connected, i.e. any two neurons are equally likely to be connected by a synapse. As outlined in section 1.3, the fundamental dynamical process of neuronal communication is based upon the following steps:

1. The neuronal axon is in an all-or-none state. In the first state it propagates a signal, *spike,* or *action potential (A.P.)*, based on the result of the summation performed in the soma. The shape and amplitude of the propagation signal (the potential difference across the cell membrane) is very stable and is replicated at branching points in the axon. In the none state there is no signal traveling in the axon, rather there is a resting potential. It is important to highlight that the presence of a traveling impulse in the axon blocks the possibility of a second impulse transmission.

2. When the traveling signal arrives at the ending of the axon it causes the secretion of neurotransmitters into the synaptic cleft.

3. The neurotransmitters arrive, across the synapse, at the membrane of the postsynaptic neuron. On the postsynaptic side these neurotransmitters bind to the receptors, thus causing the latter to open up and allow for the penetration of ionic current into the postsynaptic neuron. The amount of penetrating current per pre-synaptic spike is a parameter which specifies the *efficacy* of the synapse.

4. The postsynaptic potential (PSP) diffuses in a graded manner (unlike the spike in the axon) toward the soma where the inputs from all the pre-synaptic neurons connected to the postsynaptic one are summed. The individual PSP's are about one millivolt in amplitude. These inputs may be *excitatory* and in this case they depolarize the membrane of the postsynaptic neuron, increasing the likelihood of the appearance of a spike (firing), or they may be *inhibitory* and then they hyperpolarize the postsynaptic membrane, reducing the likelihood of firing.

5. If the total sum of the PSP's arriving within a short period surpasses a certain threshold, which is the level at which the postsynaptic membrane becomes unstable against depolarizing ionic current flows, the probability for the emission of a spike, which is a manifestation of this instability, becomes significant. This threshold is again of millivolts and hence quite a number of inputs are required in order to produce a spike.

6. After the emission of a spike, the neuron needs time to recover. There is a period of $1-2$ milliseconds in which the neuron cannot emit a second spike, no matter how large the depolarizing potential may be. This time interval is called the *absolute refractory period* of the neuron.

Now, going back to the communication process, we have that in general the *total input* h_i of each neuron i comes from two sources: external inputs I_i and inputs from other neurons; so is given by

$$h_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} T_{ij} V_j + I_i \tag{2.3}$$

where

- The first term represents the postsynaptic currents induced in i by the presynaptic activity in neuron j.
- The second term is an offset bias. As highlighted by Amit in [4], the Hopfield model is isomorphic to the *Ising model of magnetism* at temperature zero. Thus, in analogy with that model, I_i expresses the external magnetic field h_i^e, independent of the dynamics of the system, as indicated in [4]. For simplicity we will assume from now on that $I_i = 0$ for every i.
- The matrix T_{ij} describes the *synaptic interconnection strength* from neuron j to i, so in the case of non-connected neurons $T_{ij} \equiv 0$. A first model simplification can be introduced by imposing the *symmetry* of the matrix T, which implies that the efficacy of a synapse communicating output from neuron j to neuron i equals the efficacy of the synapse communicating the output of neuron i as input to neuron j (i.e. $T_{ij} = T_{ji}$). This strong assumption has no biological justification, but it has made the model completely tractable. For this reason, as it will be stressed in section 2.4, various attempts go in the direction of the introduction of the asymmetry in the model and of other features that make it closer to the biological one.
  We now assume the symmetry of the interconnection matrix T.

The updating algorithm

Every neuron i receives inputs from connected neurons and changes its state at each discrete time step t. This dynamic process can be described by the following equation[10] :

$$V_i(t+1) = \text{sgn}\Big(h_i(t+1) - U_i\Big) \tag{2.4}$$

where $h_i(t+1) = \sum_{i=1}^{N} T_{ij}V_j(t)$ and U_i represents a fixed threshold. Thus a neuron i changes the value of its output or leaves it fixed according to a fixed threshold U_i and to the following *updating algorithm*:

$$\begin{aligned} V_i(t+1) &\to 1 \quad \text{if} \quad \sum_{j=1}^{N} T_{ij}V_j(t) > U_i \\ V_i(t+1) &\to -1 \quad \text{if} \quad \sum_{j=1}^{N} T_{ij}V_j(t) < U_i \end{aligned} \tag{2.5}$$

From now on we will set the value of U_i to zero for every neuron i and so we can rewrite the previous algorithm in this way

$$V_i(t+1) = \text{sgn}\Big(\sum_{j=1}^{N} T_{ij}V_j(t)\Big) = \begin{cases} +1, & \text{if } \sum_j T_{ij}V_j(t) > 0 \\ -1, & \text{if } \sum_j T_{ij}V_j(t) < 0 \end{cases} \tag{2.6}$$

where we indicate with $V_i(t+1)$ the updated state.
What if the argument of the sign function is zero? The action taken here can be quite arbitrary. We will use the same convention of Baram [9]: if the argument is zero, neuron i remains in its previous state, regardless of whether it is on or off.

[10] The terms in brackets $t+1$ and t are used to indicate the time step we are referring to.

Synchronous and asynchronous update

The neurons can be updated in two different ways, synchronously or asynchronously, as reported by Amit [4], p. 70:

- In *synchronously or parallel* updating all neurons of the network update their activity states simultaneously according to rule (2.6) at discrete time steps $k = 1, 2, \dots,$ so they can fire only at integral multiples of a time period. The inputs of every neuron in the network are determined by the same activity state of the network in the time interval $(k-1) < t < k$ and at the beginning of each period the neurons start from a zero PSP, carrying no trace of previously accumulated inputs. In other words, after every time unit they all return to their resting membrane potentials. This type of dynamics was introduced in earlier studies of neural networks and is still a favorite among investigators with a foothold in the culture of cellular automata. Given the inherent stochastic nature of neural communication, the strong synchronicity makes this type of dynamics even more unrealistic than would be implied by all the other simplification we introduce. It does in fact have some special features which are not robust, such as two-cycles, i.e. a kind of "anti-ferromagnetic" order, characterized by alternating sites.
- In the *asynchronous or sequential* dynamics the neurons are updated one by one in some prescribed sequence or in a random order. In this mode every neuron coming up for a decision has full information about all the decisions of the individual neurons that have been updated before it. Also in this case the components of the current state vector are updated according to rule (2.6), where the one component i, chosen to be updated, is selected among the N indices with equal probability $1/N$, independently of which components were updated previously and of what the values of the probe were before and after update. This alternative class of dynamical process for networks of discrete neurons is still an idealization of the real picture but comes closer to capturing the asynchronous and stochastic nature of the operation of a neural network. For this reason this type of updating is preferred.

Associative memory

Another feature of this model is that it behaves as an *associative memory* when the state of flow generated by the previous algorithm (2.6) is characterized by stable fixed points. If these stable points describe a simple flow in which nearby points in state space tend to remain close during the flow, then initial states that are close (in Hamming distance[11]) to a particular stable state and far from all others will tend to terminate in that nearby stable state. Hence, the location of a particular stable point in the state space represents the information of a precise memory of the system (i.e. a stored pattern) and states nearby that stable point contain partial information about that memory. From an initial state with partial information a memory is reached supplying in the initial state some subpart of the memory. This highlights how this memory is truly addressable by content and not by location.

We need now to prove and understand better this last topic, starting from the definition of the interconnection matrix T and the study of its properties. Before going deeply onto this discussion, we introduce two different measures to quantify the distance between stored patterns and a probe vector.

[11] The Hamming distance is used to recognize the stored pattern closer to a given one and will be defined in section 2.2.2.

2.2.2 Hamming distance and overlap measure

In this space of network states we will need a measure for a distance between states, each of which is an N-bit word. A natural one is given by the *Hamming distance* d_H, that points out how many components are different between the two vectors.

Definition 2.2.1 (Hamming distance)**.** *The Hamming distance d_H between two vectors V and W is defined as*

$$d_H(V,W) := \frac{1}{2}\sum_{j=1}^{N}|V_j - W_j| \tag{2.7}$$

Remark. Notice that $0 \leq d_H \leq N$, where 0 is attained when all the components of the two vectors are equal and N when they differ.

In our context, we will use it to judge how far is a suitable stored pattern ξ^μ from the initial state of the network V and we will stress the pattern we are referring to with an apex:

$$d_H^\mu = \frac{1}{2}\sum_{j=1}^{N}|\xi_j^\mu - V_j| \tag{2.8}$$

An alternative measure is the *overlap*[12], which measures the degree of similarity between vectors.

Definition 2.2.2 (Overlap measure)**.** *The overlap measure between two vectors V and W is defined by setting*

$$m(V,W) := \frac{1}{N}\sum_{j=1}^{N} V_j W_j \tag{2.9}$$

Remark. Notice that $-1 \leq m \leq 1$, where -1 is reached if the two vectors differ in all the components and 1 if they are the same vector.

In what follows, we will use the overlap m to describe how similar the probe vector V is to a memorized pattern ξ^μ, putting an apex to underline which memory we are considering

$$m^\mu = \frac{1}{N}\sum_{j=1}^{N} \xi_j^\mu V_j \tag{2.10}$$

These two measures are completely equivalent and are linearly related, as we proceed to show following Amit [4], pp. 33-35.

Let N_R be the number of bits which are identical between a vector V and vector W and N the number of total bits in each word. We have that

$$N = N_R + d_H \quad \Longrightarrow \quad N_R = N - d_H \tag{2.11}$$

Then defining M as the difference between the number of agreeing bits and the number of differing bits $N_R - d_H$ and using the previous relation, we gain

$$M = N - 2d_H \quad \Longrightarrow \quad d_H = \frac{1}{2}N(1 - \frac{M}{N}) = \frac{1}{2}N(1 - m) \tag{2.12}$$

where the quantity M/N corresponds to the overlap m defined above. Hence

$$m = 1 - \frac{2}{N}d_H \tag{2.13}$$

[12] This is the kind of quantity which naturally maps on such familiar theoretical constructs as magnetization and for this reason is appreciated by physicists.

2.2.3 The interconnection matrix and Hebb's rule

In 1949 Hebb in [26] has formulated a postulate that has inspired a large body of theoretical and experimental work on learning in neural networks: "When an axon of cell A is near enough to excite cell B or repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." This famous postulate is often rephrased in the sense that modifications in the synaptic transmission efficacy are driven by the correlations in the firing activity of the pre- and post-synaptic neuron. The translation of this wise speculation in mathematical form gives *Hebb's rule* (2.16), that represents this *synaptic plasticity*, i.e. the ability of the brain to change and adapt to new information by strengthening or weakening the synapses over time.
Let $\bar{\xi} = \{\xi^{(1)}, \dots, \xi^{(p)}\}$ be a p-set of N-dimensional bipolar (± 1) column vectors, which are to be stored. These vectors will be called *fundamental memories* or *stored patterns.* We assume that the memories to be stored are random and the number of memories intentionally stored in the network is small compared to the number of neurons in the network, i.e $p \ll N$. For each memory ξ^μ we form the $N \times N$ matrix T^μ as

$$T^\mu = \frac{1}{N}\Big(\xi^\mu(\xi^\mu)^T - \mathbb{I}_N\Big) \tag{2.14}$$

where $\mathbb{I}$ denotes the $N \times N$ identity matrix and the coefficient in front of the sum is there for technical reasons[13]. To be more precise we are requiring that $T^\mu = \frac{1}{N}(\xi^\mu(\xi^\mu)^T)$ and that the diagonal elements of the matrix are zero.

Definition 2.2.3. *The* Hopfield interconnection matrix T *for the set of* p *memories* $\bar{\xi} = \{\xi^{(1)}, \dots, \xi^{(p)}\}$ *is defined by setting*

$$T = \sum_{\mu=1}^{p} T^\mu, \quad with\ T^\mu\ defined\ in\ (2.14), \tag{2.15}$$

or in components can be written as

$$T_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}(\xi_i^\mu \xi_j^\mu - \delta_{ij}) \tag{2.16}$$

where T_{ij} *is often called* synaptic weight. *This last formula represents an outer product learning rule, called* Hebbian learning rule.

The main features of this rule are the following:

- It is *local.* Each pattern contributes to synapse ij with a term which is the product of the corresponding ξ_i and ξ_j. These are exactly the activities of the neurons i and j when the network is in a state identical to the pattern.
- It is *additive*, which is reminiscent of learning.
- It is *generic*, in the sense that it involves no knowledge about the memories.

The main limitation is perhaps the equal intensity with which all the patterns are imprinted, that leads to excessive symmetry between all the stored memories.
Then it is assumed that the performance of the network can be analyzed keeping the synaptic values fixed, or *quenched.* This implies that during a typical retrieval time the changes that may occur in synaptic values are very small.

[13] The explanation of this fact is connected to the analogy with the Ising model. The interested reader can refer to chapter 3 of Amit's book [4].

2.2.4 Fixed stable points

We wish to prove that the stored patterns ξ^μ are stable points for the dynamics we have described. There are two different concepts of stability that can be investigated: the network stability or the condition to be fixed points, that we will describe in this section, and the Lyapunov stability, that will be the main topic of section 2.2.5.

Definition 2.2.4 (Network stability). *A vector V is said to be* network stable *or a* fixed point *of the system if*

$$\mathrm{sgn}((TV)_i) = V_i, \qquad \forall\ i \tag{2.17}$$

or equivalently if applying the updating algorithm (2.6), it does not change

$$V(t+1) = V(t), \quad \forall\ t, \quad i.e. \quad V_i(t+1) = V_i(t), \quad \forall\ t,\ \forall\ i \tag{2.18}$$

We make it plausible for the memories to be fixed points by proving they are eigenvectors for T. Assume that a memory ξ^μ is the initial state of the system. For each $i = 1, \dots, N$ we have

$$[T\xi^\mu]_i = \sum_{j=1}^{N} T_{ij}\xi_j^\mu = \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\beta=1}^{p}\xi_i^\beta\xi_j^\beta\xi_j^\mu = \frac{N-1}{N}\xi_i^\mu + \frac{1}{N}\sum_{\beta\neq\mu}\sum_{j\neq i}\xi_i^\beta\xi_j^\beta\xi_j^\mu \tag{2.19}$$

where the first term can be seen as the signal component and the second addend in the last equality, called *crosstalk term*, as the noise component. To obtain an exact result, we need to require that the stored patterns form an orthogonal basis. Unfortunately this is not always true, so for this reason we need to take into account two different situations and study them separately.

Orthogonal case

Following the contribution by McEliece et al. in [39], if the stored patterns form an orthogonal basis (i.e. they are uncorrelated), the second term in the last equality of (2.19) will simplify, since the dot product between two different memories vanishes: $0 = \xi^\beta \cdot \xi^\mu = \sum_j \xi_j^\beta \xi_j^\mu$. Now starting from the following relation

$$0 = \sum_j \xi_j^\beta\xi_j^\mu = \sum_{j\neq i}\xi_j^\beta\xi_j^\mu + \xi_i^\beta\xi_i^\mu \quad \Longrightarrow \quad \sum_{j\neq i}\xi_j^\beta\xi_j^\mu = -\xi_i^\beta\xi_i^\mu \tag{2.20}$$

and substituting it in (2.19) we have:

$$\begin{aligned}[T\xi^\mu]_i &= \frac{N-1}{N}\xi_i^\mu - \frac{1}{N}\sum_{\beta\neq\mu}\xi_i^\beta\xi_i^\beta\xi_i^\mu = \frac{N-1}{N}\xi_i^\mu - \frac{1}{N}\sum_{\beta\neq\mu}\xi_i^\mu = \\ &= \frac{N-1}{N}\xi_i^\mu - \frac{p-1}{N}\xi_i^\mu = \frac{N-p}{N}\xi_i^\mu\end{aligned} \tag{2.21}$$

Thus we gain that

→ The stored patterns ξ^μ are eigenvectors for T with positive eigenvalues

$$\lambda^\mu = \frac{N-p}{N}, \quad \mu = 1, \dots, p \tag{2.22}$$

→ The memories ξ^μ are fixed point of the dynamics ($y = f(y) = \mathrm{sgn}(Ty)$):

$$\mathrm{sgn}((T\xi^\mu)_i) = \mathrm{sgn}(\lambda^\mu \xi_i^\mu) = \xi_i^\mu \qquad \forall\ i \tag{2.23}$$

since ξ^μ is an eigenvector of T with positive eigenvalue λ^μ. Therefore if we apply the updating algorithm to these vectors, they will not change: $\xi_i^\mu(t+1) = \xi_i^\mu(t)$.

To gain the fact that our stored patterns are fixed points, we have required that they form an orthogonal basis. It is important to stress that they are orthogonal from the beginning, we do not orthogonalize them later and they have not unit norm: $\|\xi^\mu\| = N,\ \ \forall\ \mu$.
The next step will be to guarantee that the orthogonality condition is maintained by the neural network, as it is presented by Baram in [9].
Denoting by $T_{(k)}$ the matrix of synaptic connections corresponding to k stored patterns, we modify the previous storage rule as

$$T_{(k+1)} = \begin{cases} T_{(k)} + \dfrac{1}{N} V^{k+1}(V^{k+1})^T & \text{if } T_{(k)}V^{k+1} = 0 \\ T_{(k)} & \text{otherwise} \end{cases} \tag{2.24}$$

always requiring that the diagonal terms in the matrix $T_{(k+1)}$ are equal to zero. The condition $T_{(k)}V^{k+1} = 0$ is equivalent to the request of orthogonality of the new pattern respect all the stored memories. We have indeed that

$$0 = T_{(k)}V^{k+1} = \frac{1}{N}\sum_\beta \xi^\beta (\xi^\beta)^T V^{k+1} = \sum_\beta c_\beta \xi^\beta \tag{2.25}$$

where we used that $(\xi^\beta)^T \xi^{k+1} = c_\beta \in \mathbb{R}$ from the property of the scalar product. Now there are two possibilities: or the constants c_β are all zero or some constants are zero and other not. If $c_\beta = 0$ for every β we gain the orthogonality condition between V^{k+1} and all the stored patterns. If not all the c_β are zero, we have a linear combination of independent vectors that is equal zero with not all the coefficient zero and so we have an absurd.

General case

As discussed in [39], if we are not in the case of having an orthogonal set of memories, we cannot in general neglect the crosstalk term, that represents a sort of noise. From Eq. (2.19) it follows that the p fundamental memories are approximately eigenvectors of the linear transformation T, with approximate positive eigenvalues 1. We also deduce a probabilistic pseudoorthogonality by considering the mean value of the product

$$[T\xi^\mu]_i = \sum_{j=1}^N T_{ij}\xi_j^\mu = \frac{1}{N}\sum_\beta \xi_i^\beta \Big[\sum_{j\neq i}(\xi_j^\beta \xi_j^\mu)\Big] \equiv H_i^\mu \tag{2.26}$$

It is easy to prove that the mean value of the bracketed term in the previous equation is 0 unless $\beta = \mu$ for which is $N-1$.
First, we recall the definition of mean value for a variable X

$$\langle X\rangle = \sum_j X_j \mathbb{P}(X_j)$$

where X_j are the values that the variable can assume and $\mathbb{P}(X_j)$ the associated probability. In our context the variables in exam are the component of the stored pattern ξ_j^μ, which are independent random variables with values ±1, each with probability 1/2, and zero mean.

→ In the case of $\beta = \mu$ we have

$$\langle \sum_{j\neq i} (\xi_j^\mu \xi_j^\mu) \rangle = \sum_{j\neq i} \langle (\xi_j^\mu)^2 \rangle = \sum_{j\neq i} \langle 1 \rangle = N-1 \quad \Longrightarrow \quad \langle H_i^\mu \rangle = \frac{N-1}{N} \xi_i^\mu$$

where we have used the linearity of the mean value.

→ When $\beta \neq \mu$ we obtain:

$$\langle \sum_{j\neq i} (\xi_j^\beta \xi_j^\mu) \rangle = \sum_{j\neq i} \langle \xi_j^\beta \xi_j^\mu \rangle = \sum_{j\neq i} \Big(\langle \xi_j^\beta \rangle \, \langle \xi_j^\mu \rangle \Big) = 0$$

since ξ_j^β and ξ_j^μ are independent variables with zero mean.

In this situation the crosstalk term will also determine the network stability of our stored patterns, as explained by Hertz, Palmer and Krogh in [27]. We start by defining the quantity C_i^μ:

$$C_i^\mu = -\xi_i^\mu \frac{1}{N} \sum_j \sum_{\beta\neq\mu} \xi_i^\beta \xi_j^\beta \xi_j^\mu \tag{2.27}$$

- If $C_i^\mu < 0$, the crosstalk term has the same sign of ξ_i^μ and so this term does not affect the network stability of the pattern.
- If $C_i^\mu > 0$, they have different sign and there are two possible situations that can happen:
  - if $0 < C_i^\mu < (N-1)/N$, with the absolute value of the crosstalk term less than $(N-1)/N$, it does not change the sign of $[T\xi^\mu]_i$ and we have that the stability condition $\text{sgn}([TV^\mu]_i) = V_i^\mu$ is satisfied. In this case there can be some wrong bits but they will be correct during the evolution.
  - If $C_i^\mu > (N-1)/N$ the crosstalk term modifies the sign of our product and the pattern is unstable. Hence, the dynamics of the network tends to depart from the fundamental memory.

Storage Capacity and network stability

The concept of network stability is linked to the maximum capacity of the network. An estimate of the number of state vectors that can be made fixed points in a Hopfield network can serve as common criterion for evaluation of Hopfield networks functioning as associative memory. This measure is known as *storage capacity* or *capacity of Hopfield network*. To study the storage capacity we need to define some new tools, as stated in [25], pp. 715-718. We assume for simplicity that $N, p \gg 1$ and we note that C_i^μ, defined in Eq. (2.27), is the sum of (roughly) $N(p-1)$ independent and identically distributed (i.i.d.) random variables, say y_m for $1 \leq m \leq Np$, that are equally likely to be $+1$ or -1. We then observe that every y_m has zero mean and variance $\sigma^2 = 1$. Thus, using *the Central Limit Theorem*[14] we obtain that C_i^μ has approximately a Gaussian distribution with zero mean and variance $\sigma^2 = \frac{1}{N^2}(N(p-1)) = \frac{p-1}{N}$.

[14] *Central Limit Theorem*: If z_m is a sequence of i.i.d. random variables each with mean μ and variance σ^2 then for large n, $X_n = \frac{1}{n}\sum_{m=1}^n z_m$ has approximately a normal distribution with mean μ and variance σ^2/n.

Definition 2.2.5. *We introduce the following notions, which we will need in the study of network capacity, as presented in [25], p.716:*

1. *The* signal-to-noise ratio *is defined by setting*

$$\rho = \frac{\textit{variance of signal}}{\textit{variance of noise}} = \frac{1}{(p-1)/N} \simeq \frac{N}{p} \quad \textit{for large } p \tag{2.28}$$

2. *The reciprocal of the signal-to-noise ratio, that is,*

$$\alpha = \frac{p}{N} \tag{2.29}$$

is called the load parameter *or* storage efficiency *and is the ratio of number of candidate state vectors p that are made stored stable states to the number of neurons N in the network. Thus α is the number of candidate state vectors made stable state per neuron. From Eq. (2.29) we define the* storage capacity *of a Hopfield network as*

$$p = \alpha N \tag{2.30}$$

3. *The* critical storage capacity *p_c is defined as that storage capacity of Hopfield network beyond which it is not possible to store candidate state vectors without affecting the network stability of already stored state vectors.*

Statistical physics considerations reveal that the quality of memory recall of the Hopfield network deteriorates with increasing load parameter α and breaks down at the critical value $\alpha_c = 0.14$ (see for example Amit, 1989 [4]). This critical value is in agreement with the estimate of Hopfield (1982, [28]), where it is reported that as a result of computer simulation $0.15N$ states can be recalled simultaneously before errors become severe.

→ With $\alpha_c = 0.14$, we find from Eq. (2.28) that the critical value of the signal-to-noise ratio is $\rho_c \simeq 7$ and below this critical value memory recall breaks down. Therefore, it follows from Eq. (2.28) that as long as the storage capacity of the network is not overloaded (i.e. is the number p of fundamental memories is small compared to the number N of neurons in the network) the fundamental memories are network stable. Then, the critical value

$$p_c = \alpha_c N = 0.14N \tag{2.31}$$

defines the *storage capacity with errors* on recall.

To determine the storage capacity without errors we must use a more stringent criterion defined in terms of probability of errors as described by Hertz et al. in [27] and Orhan [43]. Coming back to what we have seen before for C_i^μ, now we want to evaluate the probability to have an unstable condition:

$$P_{err} = P\Big(C_i^\mu > \frac{N-1}{N}\Big) \tag{2.32}$$

If we choose an acceptable performance criteria with the error probability fixed, we can determine the maximum storage capacity, i.e. the maximum value of p that satisfies this criteria.

1. If we store p patterns in a Hopfield network with a large number of N neurons, then the probability of error, i.e. the probability that $C_i^\mu > (N-1)/N$, is:

$$P_{err} = \frac{1}{\sqrt{2\pi}\sigma}\int_{\frac{N-1}{N}}^{\infty} e^{-\frac{x^2}{2\sigma^2}}dx = \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\infty} e^{-\frac{x^2}{2\sigma^2}}dx - \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\frac{N-1}{N}} e^{-\frac{x^2}{2\sigma^2}}dx \quad (2.33)$$

because C_i^μ has a Gaussian distribution.

2. Using in Eq. (2.33) the substitution $u = \frac{x}{\sqrt{2\pi}\sigma}$ we gain

$$\begin{aligned} P_{err} &= \frac{1}{2} - \int_0^{\frac{1}{\sqrt{2\sigma^2}}\frac{N-1}{N}} e^{-u^2}du = \frac{1}{2}\Big(1 - erf\Big(\frac{1}{\sqrt{2\sigma^2}}\frac{N-1}{N}\Big)\Big) = \\ &= \frac{1}{2}\Big(1 - \mathrm{erf}\Big(\frac{N-1}{N}\sqrt{\frac{N}{2p}}\Big)\Big) \end{aligned} \quad (2.34)$$

where

$$\mathrm{erf}\Big(\frac{1}{\sqrt{2\sigma^2}}\Big) := \frac{2}{\sqrt{\pi}}\int_0^{\frac{1}{\sqrt{2\sigma^2}}} e^{-u^2}du \quad (2.35)$$

3. Fixing an acceptable value for the error probability, e.g. $P_{err} < 0.001$, we find the maximum number of memories p_{max} that can be stored using the chosen criteria. A long and sophisticated analysis of the stochastic Hopfield network shows that p cannot be larger than about $0.185N$, otherwise small errors can pile up in updating and the memory becomes useless.

4. The above result gives the single-bit error probability. Since there are N bits in a stored pattern, the probability of error-free recall of a stored pattern is given by $(1 - P_{err})^N$. We require that this probability be grater than some set value, say 0.99, i.e.

$$(1 - P_{err})^N > 0.99$$

Using the binomial expansion of the left-hand side and keeping the two lowest-order terms with respect to P_{err} (because P_{err} is small), we get

$$P_{err} < 0.01/N$$

Thus, $P_{err} \to 0$ as $N \to \infty$.

5. From equation (2.34), this implies

$$\mathrm{erf}\Big(\frac{N-1}{N}\sqrt{\frac{N}{2p}}\Big) \to 1 \quad \Longrightarrow \quad \frac{N-1}{N}\sqrt{\frac{N}{2p}} \to \infty$$

which yields $p/N \to 0$ as $N \to \infty$. Therefore, using the following asymptotic approximation for the *erf* function in Eq. (2.34):

$$1 - \mathrm{erf}(x) \to \exp(-x^2)/\sqrt{\pi}x \quad \text{as } x \to \infty \quad (2.36)$$

we gain

$$\log(P_{err}) \approx -\log 2 - \Big(\frac{N-1}{N}\Big)^2\frac{N}{2p} - \frac{1}{2}\log\pi - \log\Big(\frac{N-1}{N}\Big) - \frac{1}{2}\log\Big(\frac{N}{2p}\Big) \quad (2.37)$$

6. Requiring $P_{err} < 0.01/N$ in Eq. (2.37), we have

$$-\log 2 - \Big(\frac{N-1}{N}\Big)^2 \frac{N}{2p} - \frac{1}{2}\log\pi - \log\Big(\frac{N-1}{N}\Big) - \frac{1}{2}\log\Big(\frac{N}{2p}\Big) < \log 0.01 - \log N \quad (2.38)$$

   Keeping only terms of leading order in N, we find

$$\frac{(N-1)^2}{2Np} > \log N \quad \Longrightarrow \quad p < \frac{(N-1)^2}{2N\log N} \quad (2.39)$$

   Thus asymptotically, for $N \to \infty$, this behaves as

$$p < \frac{N}{2\log N} \quad (2.40)$$

7. We have found an upper bound for the storing capacity and we have proved that as long as the number of the stored patterns p satisfies Eq. (2.40), the absolute value of the crosstalk term will be highly unlikely to be larger than $(N-1)/N$. Thus in these hypothesis the patterns ξ^μ are fixed points of the network dynamics.

→ The *storage capacity almost without errors* $p_{\max}$ is defined as the largest number of fundamental memories that can be stored in the network and most of them be recalled correctly. From Eq. (2.40) we gain the formula

$$p_{\max} = \frac{N}{2\log N} \quad (2.41)$$

2.2.5 The energy function

The main feature of the Hopfield model is the converge of the state space flow algorithm to stable states. Any symmetric matrix T with zero diagonal elements (i.e $T_{ij} = T_{ji}$ and $T_{ii} = 0$) will produce such a flow. The proof of this property is based on the construction of an appropriate energy function, that is monotonically decreasing and represents a candidate Lyapunov function. If this is true, we obtain that the Hopfield net is Lyapunov stable[15], a condition stronger than the simple network stability.

Definition 2.2.6 (Lyapunov stability). *Given a discrete in time system*

$$x(t+1) = f(x(t))$$

where $f : D \to \mathbb{R}^N$ continuous on D, with $D \subset \mathbb{R}^N$ an open set containing the origin and $x(t) \in D \subset \mathbb{R}^N$. Suppose f has an equilibrium point x^* so that $f(x^*) = 0$, then this equilibrium is said to be

- Lyapunov stable *if, for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that*

$$\|x(0) - x^*\| < \delta \quad \Longrightarrow \quad \|x(t) - x^*\| < \epsilon \quad \forall\ t \geq 0$$

- asymptotically Lyapunov stable *if it is Lyapunov stable and there exists $\delta > 0$ such that*

$$\|x(0) - x^*\| < \delta \quad \Longrightarrow \quad \lim_{t\to\infty} \|x(t) - x^*\| = 0$$

[15] A neural network is said to be stable, if all its neurons are stable.

We assume asynchronous updating. Then the state of an asynchronous Hopfield network can be characterized by an energy function, defined as follows

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}V_iV_j - \sum_{i=1}^{N} I_iV_i + \sum_{i=1}^{N} U_iV_i \tag{2.42}$$

We assume $U_i = 0$ and $I_i = 0$, thus E is simplified into

$$E = -\frac{1}{2}\sum_{i}\sum_{j} T_{ij}V_iV_j \tag{2.43}$$

We also assume that the stored patterns form an orthogonal basis and are not correlated to simplify the notations, but the result we will present (Prop. 2.2.9) holds also in the general case.

We recall the classical definition of Lyapunov function and introduce the definition of Lyapunov function in the abstract theory of dynamical systems, as reported in [1].

Definition 2.2.7 (Lyapunov function). *A function $f(x)$, $x \in \mathbb{R}^N$ is said to be a* Lyapunov function *if there exists an equilibrium state x^*, such that the following three conditions are satisfied:*

1. *$f(x)$ is continuous with respect to all components of x.*
2. *$f(x)$ is positive definite, that is $f(x^*) = 0$ and $f(x) > 0$ for $x \neq x^*$.*
3. *The time derivative $\dot{f}(x)$ is negative semidefinite, i.e. the function decreases in time.*

If it is defined only on a open neighborhood of the equilibrium point, f is said to be a local Lyapunov function *([23]).*

Definition 2.2.8 (Lyapunov function-2). *Let (X, d) be a metric space. A function $h : X \to \mathbb{R}$ is said to be a* Lyapunov function *for the flow $\psi = \{\psi_t\}_{t\in\mathbb{R}}$ on X if*

$$h \circ \psi_t \leq h, \quad \forall\ t \geq 0 \tag{2.44}$$

and is said to be a first integral *if*

$$h \circ \psi_t = h, \quad \forall\ t \in \mathbb{R} \tag{2.45}$$

Proposition 2.2.9. *The energy function (2.43) is a Lyapunov function*[16].

Proof. The function (2.43) satisfies the tree conditions of the definition of Lyapunov function (2.2.7):

1. $\frac{\partial}{\partial V_i}E(V) = -\sum_j T_{ij}V_j, \quad \forall\ V_i$, implying[17] $E(V)$ is continuous with respect to all components of V.

[16] In this proof we will follow [6], pp. 30-31, except for the second point where we will prove that E is also positive definite and not only bounded.

[17] If $f : A \to \mathbb{R}^N$ is differentiable at $c \in A$, then there exists strictly positive numbers δ, K such that $\|x - c\| < \delta$, then $\|f(x) - f(c)\| \leq K\,\|x - c\|$. In particular, it follows that f is continuous.

2. Equation (2.43) does not satisfy Condition (2) of the Lyapunov function. In general this condition is substitute with the boundedness of the function, that can be easily proved and that highlights the existence of stopping points for the dynamics of the network, at which the function attains its lower bound. This represents the definition of Lyapunov function in the abstract theory of dynamical systems, as defined in Def. 2.2.8. As presented in [6], a rough bound[18] for E can be found in this way:

$$E \geq -\frac{1}{2}\sum_i \sum_j |T_{ij}V_iV_j| = -\frac{1}{2}\sum_i \sum_j |T_{ij}| \geq -\frac{1}{2}N^2\frac{p}{N} = -\frac{1}{2}pN \tag{2.46}$$

where in the last inequality we used that $|T_{ij}| \leq \frac{1}{N}\sum_\beta |\xi_i^\beta \xi_j^\beta| = \frac{p}{N}$ and that the absolute value of all the components of the vector V_i are equal to one.

Statement. *The bound given in Eq. (2.46) is never realized and so the inequality is strict.*

Proof. For absurd we suppose that for a pattern V the inequality (2.46) is satisfied exactly:

$$E(V) = -\frac{1}{2}\sum_i \sum_j T_{ij}V_iV_j = -\frac{1}{2}pN$$

We can find a bound for the argument of this double sum $T_{ij}V_iV_j$ by considering the best case that can occur.
First, we have that

$$T_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \leq \frac{p}{N}$$

because $\xi_i^\mu \xi_j^\mu \leq 1$ for every μ and i, j. The equality is obtained when for every μ and i, j fixed

$$\xi_i^\mu \xi_j^\mu = 1 \quad \Longrightarrow \quad \xi_i^\mu = 1 = \xi_j^\mu \ \text{ or } \ \xi_i^\mu = -1 = \xi_j^\mu$$

Now we consider $T_{ij}V_iV_j$ and we have

$$T_{ij}V_iV_j \leq \frac{p}{N}$$

where we used $T_{ij} \leq \frac{p}{N}$ and $V_iV_j \leq 1$ for every i, j. From these considerations we have that for i, j fixed

$$T_{ij}V_iV_j = \frac{p}{N} \quad \Longleftrightarrow \quad T_{ij} = \frac{p}{N} \ \text{ and } \ V_iV_j = 1$$

Now we take the sum over i and over j of these quantities and to maximize this double sum the conditions seen before have to be true for every i and j:

(a) $V_i = 1 = V_j$ or $V_i = -1 = V_j$ for every i, j.

(b) $\xi_i^\mu = 1 = \xi_j^\mu$ or $\xi_i^\mu = -1 = \xi_j^\mu$ for every μ, i, j.

and in this way we obtain

$$-\frac{1}{2}\sum_i \sum_j T_{ij}V_iV_j = -\frac{1}{2}pN$$

[18] The boundedness of E is very important, because is telling us that E cannot converge to $-\infty$, but is bounded from below by its value in the minimum points.

- The first condition (a) implies that V is equal to one of these two $N-$vectors $V_1 = (1, \dots, 1)^T$ or $V_2 = (-1, \dots, -1)^T$.
- The second condition (b) means that the unique stored patterns we can have, are represented by V_1 and V_2! This is absurd for the model we are describing, where the number of memories is quite high.

□

We have thus obtain that (2.46) can be rewritten as

$$E > -\frac{1}{2}pN \tag{2.47}$$

We now can derive a better lower bound identifying the minimum points of the energy function.
First, we compute the time derivative[19], that in this case corresponds to the change of energy ΔE due to the change in the state of the Hopfield network ΔV_i,

$$\Delta E = -\sum_i \sum_j T_{ij} V_j \Delta V_i = -\sum_i \Delta V_i \sum_j T_{ij} V_j = -\Delta V^T (TV) \tag{2.48}$$

where

$$\Delta E = E(t+1) - E(t), \qquad \Delta V_i = V_i(t+1) - V_i(t)$$

Then we identify the points at which this derivative becomes zero and thus $\Delta E = 0$ if and only if one of this condition is satisfied:

(a) $\Delta V_i = 0 \ \forall \ i$: The vectors V satisfying this property correspond to fixed points of the net, i.e. points such that their value does not change in further updating, $V_i(t+1) = V_i(t)$, thus the stored patterns. Using (2.21), the value of E in such points is equal to

$$\begin{aligned} E(\xi^\mu) &= -\frac{1}{2}\sum_i \sum_j T_{ij} \xi_j^\mu \xi_i^\mu = -\frac{1}{2}\sum_i \Big[\frac{N-p}{N}(\xi_j^\mu)^2\Big] = \\ &= -\frac{N(N-p)}{2N} = -\frac{N-p}{2} < 0 \end{aligned} \tag{2.49}$$

(b) $\sum_j T_{ij} V_j = 0, \ \forall \ i$: This condition implies that V is in the kernel of the connection matrix T

$$ker(T) = \{V \mid (TV)_i = \sum_j T_{ij} V_j = 0, \ \forall \ i\}$$

and evaluating E in these patterns we obtain

$$E(V^{ker(T)}) = -\frac{1}{2}\sum_i \sum_j T_{ij} V_j^{ker(T)} V_i^{ker(T)} = 0 \tag{2.50}$$

(c) $\sum_j \sum_i T_{ij} V_j \Delta V_i = 0$: In this case we are requiring that $TV \perp \Delta V$ and thus we obtain a null value for E also for these points.

Statement. *The first vectors analyzed represent minimum points for E and thus the only minimum points for E are represented by the stored patterns.*

[19] For the explicit calculation of the time derivative see Appendix A.

Proof. We consider the last two types of points that cancel the derivative of E.

- The second type of vectors are in $ker(T)$, so they are such that $\sum_j T_{ij}V_j = 0$. Since we have assumed that the stored patterns form an orthogonal basis and we have seen previously in (2.24) how to modify the storage rule in order to memorize a new vector as a fundamental memory, we have that a vector of this type is exactly a new stored pattern. Therefore these are also minimum points for the energy function[20].

- In the last situation we have that $TV \perp \Delta V$ with $\Delta V \neq 0$ and $TV \neq 0$, otherwise we are in one of the previous cases. For this type of vectors we have that $E(V) = 0$, which is different from the value attained in the stored patterns, and they are not fixed points because $\Delta V \neq 0$. So they are only critical points and not minimum points.

Thus, we have proved that the only minimum points[21] for E are represented by the stored patterns ξ^μ. □

Remark. We note that the minimum points of E are global minimum points and thus we have obtained that the value of E in all the minimum points is the same.

Therefore, we can improve the lower bound of E:

$$E(V) > E(\xi^\mu) = -\frac{1}{2}(N-p), \quad \forall\, V \neq \xi^\mu. \tag{2.51}$$

From Eq. (2.51), we can immediately see how the energy function (2.43) is not the correct candidate to be a Lyapunov function, because from (2.49) if we evaluate it in a fundamental memory ξ^μ, $E(\xi^\mu)$ is not equal zero. In order to gain the Lyapunov characterization we have to substitute (2.43) with

$$E = -\frac{1}{2}\sum_i \sum_j T_{ij}V_iV_j + \frac{N-p}{2} \tag{2.52}$$

and for this function we have that $E(\xi^\mu) = 0$ and the property of positive definiteness is achieved:

$$E(V) > 0, \quad \forall\, V \neq \xi^\mu. \tag{2.53}$$

3. Condition (3) of the Lyapunov function is satisfied if we can prove that there is a decreasing in the energy when we update one or more components, thus any change in V_i results in a decrease of E. As seen before, we have that

$$\begin{aligned}\Delta E &= -\sum_{ij} T_{ij}V_i(t)(V_j(t+1) - V_j(t)) \\ &= -\sum_j \Big(\sum_i T_{ji}V_i(t)\Big)(V_j(t+1) - V_j(t))\end{aligned} \tag{2.54}$$

[20] In the general case, this deduction is not correct. For these patterns we have that the network stability condition is satisfied if and only if $V_i = sgn((TV)_i)\ \forall\ i$ and in our assumptions this is true. Furthermore they are also minimum points, because using the updating algorithm they change in this way: $V_i(t+1) = sgn((TV)_i) = V_i(t)\ \forall\ i$. Thus in this case we have to take into account also these points with null value of energy and null eigenvalue, that in section 2.2.6 we will call *spurious states.*

[21] In the general case we have that also the points with null eigenvalue are minimum points for E, but this is not a problem. For the deduction of the better bound (2.51) the presence of these points does not affect the result, since the value of E in these points is larger than that in the stored patterns, where $E(\xi^\mu) = -\frac{(N-1)}{2} - \frac{1}{2N}\sum_{\beta\neq\mu}\Big[\sum_i \xi_i^\beta \xi_i^\mu\Big]^2 < 0$.

where $E(t+1)$ is the energy associated with the updated pattern $V(t+1)$. By using rule (2.6) we have

$$\begin{cases} \sum_i T_{ji}V_i(t) \geq 0 & \Longrightarrow \quad 1 - V_j(t+1) \geq 0 \\ \sum_i T_{ji}V_i(t) < 0 & \Longrightarrow \quad -1 - V_j(t+1) \leq 0 \end{cases} \tag{2.55}$$

From this consideration we can conclude that $E(t+1) - E(t) \leq 0$ and it reaches the value zero when we are considering a stored pattern ξ^μ, that is a fixed point.

Therefore we have proved that the energy[22] (2.52) is a Lyapunov function characterized by a matrix of weights with zero diagonal. □

We now recall the *Lyapunov stability criterion*:

Theorem 2.2.10 (Lyapunov stability criterion)**.** *The equilibrium state*[23] x^* *is* Lyapunov stable *if in a neighborhood of* x^* *there exists a positive definite function* $L(x)$*, whose time derivative is negative semidefinite in that neighborhood and such that* $L(x^*) = 0$*. Thus* L *is a Lyapunov function. If the time derivative of* L *is negative definitive,* x^* *is said to be* asymptotically Lyapunov stable*.*

We have seen[24] that in the Hopfield net the equilibrium points are the stored patterns and thus from the Lyapunov theorem we have that they are asymptotically Lyapunov stable. For this reason starting from a probe pattern we will converge through the iterations of the updating algorithm to a Lyapunov stable state, that do not further change with time.

2.2.6 Spurious states and eigenanalysis

The weight matrix T of a discrete Hopfield network is symmetric. Therefore, the eigenvalues of T are all real and T can be completely characterized by its eigenvalues and corresponding orthogonal eigenvectors. Let these be denoted

$$\lambda_1, \ldots, \lambda_N \qquad e^1, \ldots, e^N$$

→ For large p, the eigenvalues are ordinarily *degenerate*, which means that there are several eigenvectors with the same eigenvalue. The eigenvectors associated with a degenerate eigenvalue form a subspace.

→ The weight matrix T has a degenerate eigenvalue with a value of zero, whose relative subspace is called the *null space*. The null space exists by virtue of the fact that the number of fundamental memories, p, is smaller than the number of neurons N in the network. The presence of this null space is an intrinsic characteristics of the Hopfield network.

[22] If in Eq. (2.42) we assume $I_i \neq 0$ and $U_i \neq 0$ this result is not true and we will have only local Lyapunov functions, one for each ξ^μ. A detailed proof for a case similar to this has been done for the continuous time energy and can be found in Prop. 2.3.1.

[23] An equilibrium point for a differential equation $\dot{x} = f(t,x)$ is a point x^* such that $f(t,x^*) = 0, \ \forall \ t$ or equivalent is an equilibrium point for $x(t+1) = f(t,x(t))$ if $f(t,x^*) = x^*$ for $t = 0, 1, 2, \ldots$

[24] As discussed above the same result holds also in the general case with an analogous proof. Indeed, we have used the hypothesis of orthogonality only in the analysis of the vector in $ker(T)$ (we have already seen how we can overcome this) and in the evaluation of the energy function in the stored pattern. In the general case we don't know these values exactly, because they depend on the scalar product $\xi^\mu \cdot \xi^\beta$ and so for this reason we have preferred to present the proof with the simplest case.

To understand better, we make an eigenanalysis of T, as done in Aiyer et al. [3].
Let $e^1, \ldots, e^{p_1}$ be the eigenvectors with non zero eigenvalue $\lambda_1, \ldots, \lambda_{p_1}$ and $e^{p_1+1}, \ldots, e^N$ those with null eigenvalue. A vector V can be written in terms of its component V^k in the direction of the eigenvector e^k, $k = 1, \ldots, p_1$, plus its component q in the null subspace as follows:

$$V = \sum_{k=1}^{p_1} V^k + q \tag{2.56}$$

where

$$V \cdot e^k = V^k = \gamma_k e^k$$

Similarly, the connection matrix, energy function and dynamic update equation can be expressed as follows:

$$\begin{aligned} T &= \sum_{k=1}^{p_1} \lambda_k e^k (e^k)^T \\ E &= -\frac{1}{2}\sum_{k=1}^{p_1} \lambda_k |V^k|^2 = -\frac{1}{2}\sum_{k=1}^{p_1} \lambda_k \gamma_k^2 \\ TV &= \sum_{k=1}^{p_1} \lambda_k v^k \end{aligned} \tag{2.57}$$

It can be seen that to minimize E the network must move V so as to:

- reduce to zero magnitude all V^k's where $\lambda_k < 0$
- increase the magnitude of all V^k's where $\lambda_k > 0$

Let V be a memory vector and thus a network stable state. The condition of network stability requires

$$\begin{aligned} V_i &= sgn((TV)_i) \qquad \text{for } i = 1, \ldots, N \\ &= sgn\Big(\sum_{k=1}^{p_1} \lambda_k V_i^k\Big) \\ \text{using (2.56)} \implies \sum_{k=1}^{p_1} V_i^k + q_i &= sgn\Big(\sum_{k=1}^{p_1} \lambda_k V_i^k\Big) \end{aligned} \tag{2.58}$$

The only way that (2.58) can be guaranteed for any set of memory vectors is if

$$q_i = 0 \quad \text{for } i = 1, \ldots, N \quad \text{and} \quad \lambda_k = \lambda \qquad \text{for } k = 1, \ldots, p_1 \ \text{ with } \lambda > 0$$

In other words:

1. To ensure $q = 0$, the null subspace must be orthogonal to all the memory vectors.
2. As a result, all the memory patterns must be completely specified by $\sum_{k=1}^{p_1} V^k$, hence the eigenvectors of T must at least span the subspace formed by the memory vectors. This will automatically ensure that the previous point is true.
3. So that $\lambda_k = \lambda$, $\forall\ k$ with $k = 1, \ldots, p_1$, the connection matrix must have a single positive degenerate eigenvalue corresponding to the memory vector subspace.

Let $\mathcal{M}$ denote the subspace spanned by the memory vectors. From 1. the complementary subspace to this will be the null subspace of T, denoted with $\mathcal{Q}$. Thus a vector V can be written in terms of its component in $\mathcal{M}$, denoted by $\tilde{m}$, and its component in $\mathcal{Q}$, denoted by q

$$V = \tilde{m} + q, \qquad \text{where } \tilde{m} \cdot q = 0 \tag{2.59}$$

Remark. We note that if the memorized patterns formed an orthogonal basis, they correspond to the eigenvectors e^k of T, as seen in section 2.2.4, and thus $p_1 = p$.

This eigenanalysis of the weight matrix T leads us to take the following viewpoint of the discrete Hopfield network used as a CAM:

→ The discrete Hopfield network acts as a *vector projector* in the sense that it projects a probe vector onto a subspace $\mathcal{M}$ spanned by the fundamental memory vectors. Then the underlying dynamics of the network drives the resulting projected vector to one of the corners of a unit hypercube where the energy function is minimized.

→ The p fundamental memories, spanning the subspace $\mathcal{M}$, constitute a set of fixed points represented by certain corners of the unit hypercube. The other corners of this hypercube that lie in or near $\mathcal{M}$ are potential locations for *spurious states*, also referred to as *spurious attractors*. Spurious states represent network stable states of the Hopfield network that are different from the fundamental memories of the network and are characterized by a higher energy. Furthermore, their number increases very rapidly with the number of stored memory vectors p (Ayer et al. [3]).

Following Amit [4], the spurious states may be grouped as follows:

- *Reversed fundamental memories.* The spurious states are reversed (i.e. negative) versions of the fundamental memories of the network. To explain this kind of spurious state, we note that the energy function E is symmetric in the sense that its value remains unchanged if the states of the neurons are reversed (i.e. the state V_i is replaced by $-V_i$ for all i). Accordingly, if the fundamental memory ξ^μ corresponds to a particular local minimum of the energy landscape, that same local minimum also corresponds to $-\xi^\mu$. This sign reversal does not pose a problem in the retrieval of information if it is agreed to reverse all the information bits of a retrieved pattern if it is found that a particular bit designed as the "sign" bits is -1 instead of $+1$.

- *Mixture states* A mixture spurious state is a linear combination of an *odd* number of stored patterns. For example, consider the state

  $$V_i = sgn(\xi_i^{\mu_1} + \xi_i^{\mu_2} + \xi_i^{\mu_3})$$

  which is a three-mixture spurious state. It is a state formed out of three fundamental memories $\xi_i^{\mu_1}, \xi_i^{\mu_2}, \xi_i^{\mu_3}$ by a majority rule. The network stability condition $y = sgn(Ty)$ is satisfied by such a state for a large network.

- *Spin-glass states.* This kind of spurious state is so named by analogy with spin-glass models of statistical mechanics. Spin-glass states are defined by local minima of the energy landscape that are not correlated with any of the fundamental memories of the network.

In the design of a Hopfield network as a content-addressable memory we are therefore faced with a tradeoff between two conflicting requirements:

1. the need to preserve the fundamental memory vectors as fixed points in the state space
2. the desire to have few spurious states.

2.2.7 The Attraction Basin

From what has been said up to now, it is obvious that stored memories are attractors for the dynamics, i.e. they have a region of influence around them so that states which are sufficiently similar to a stored pattern are mapped into it by repeated iterations of the system operator. In Hopfield neural network this region is called *basin of attraction*[25].

Definition 2.2.11 (Basin of Attraction)**.** *The* basin of attraction *is the set of states in the system within which almost all states flow to one attractor and it is usually measured by Hamming distance.*

Using the "landscape description" of the dynamical process, we have that the basins of attraction correspond to the valleys around each minimum. Thus, for any starting point the state slides down hill until it comes to rest (a stable state) at one of these minima (the attractors). The flow of a state to a pattern is not determined solely by the Hamming distance, but depends also on the proximity of the state to the other stable states and to the spurious states and secondly on the path that is taken. This means that the basins of attraction are sensitive to the details of the dynamics. For instance, serial and parallel dynamics may define different basins of attraction, that may be affected also by order of updating in the serial dynamics.
The attraction size of the basin of attraction is represented by the *hamming radius*, that is defined by Kanter and Sompolinsky in [31] and by Storkey and Valabregue in [57] as follows:

Definition 2.2.12 (Radius of attraction)**.** *Given specific dynamic rules, the* radius of attraction R^μ *of a stored pattern* ξ^μ *is defined as the largest Hamming distance within which almost all of the states (but not necessarily all of them) flow to the pattern.*

$$R(\xi^\mu) = R^\mu = \max\{d_H^\mu(\xi^\mu, V) : V \in Basin(\xi^\mu)\} \tag{2.60}$$

where $Basin(\xi^\mu)$ *is the basin of attraction of* ξ^μ*, i.e. the set of states that are attracted to* ξ^μ*. It is common for* R *to be normalized with respect to the size of the network, so that it lies between zero and one:* $0 \leq R^\mu \leq 1$*.*
The average size of attraction basin over all the stored patterns is then defined as

$$R = \left\langle R(\xi^\beta) \right\rangle_{\xi^\beta} = \frac{1}{p} \sum_{\xi^\beta \in \bar{\xi}} R(\xi^\beta) \tag{2.61}$$

Remark. Recalling the relation between Hamming distance and overlap measure given in Eq. (2.12) and using it in Eq. (2.60), we have that

$$R(\xi^\mu) = \frac{1}{2}(1 - m^\mu) \tag{2.62}$$

where we omit the value N at the denominator because $R(\xi^\mu)$ has been normalized.

The mean radius of attraction over the patterns R can act as a measure of the quality of a particular associative memory and clearly, the larger the sample size the higher the quality of the estimate. As done in [31], the radii of attraction R of the embedded memories can be measured by following the time evolution of states with varying initial overlaps.
Let $\mathcal{V} = \{V^1, \ldots, V^k\}$ be a set of probe vectors. Consider one of the sample state $V^i \in \mathcal{V}$,

[25] All the definitions and the results we will present in this section are true both for uncorrelated and for correlated pattern, that willl be presented in section 2.4.2.

that has an overlap $m_i^\mu = m(V^i, \xi^\mu)$ with the pattern ξ^μ, i.e. a fixed fraction $m_i^\mu N$ of the state identical to the corresponding one of the stored pattern ξ^μ and the rest of the state is random. From simulations, it can be seen that at high value of m_i^μ, states always flow to the pattern ξ^μ, whereas as it is reduced, a substantial number of states flow to different fixed points. From Eq. (2.62) a working definition[26] of $R(\xi^\mu)$ is given by

$$R^\mu = \frac{1}{2}(1 - m_c^\mu) \tag{2.63}$$

where m_c^μ is the smallest value of the overlap m_i^μ such that as $N \to \infty$ almost all of the states having $m_i^\mu > m_c^\mu$ will evolve toward ξ^μ.
Averaging the final value of m_c^μ over different sets of probe patterns $\mathcal{W}$ and all the stored patterns yields

$$R = \frac{1}{2}\Big(1 - \langle\langle m_c^\mu \rangle_{\mathcal{W}}\rangle_{\xi^\mu}\Big) \tag{2.64}$$

As pointed out in [31], for finite size associative memories another factor needs to be considered. A probe vector V^i may overlap also other stored patterns ξ^β, with $\beta \neq \mu$, and to compensate for this the definition of R is modified to:

$$R = \frac{1}{2}\left\langle\left\langle \frac{1 - m_c^\mu}{1 - m_{av}} \right\rangle_{\mathcal{W}}\right\rangle_{\xi^\mu} \tag{2.65}$$

where m_{av} is the average overlap of the given state with all of the other stored states, e.g. the largest overlap with the rest of the patterns $m_{av} = \max\{m^\beta\}$, $\beta \neq \mu$, and the brackets indicate an average over all stored states and all starting configurations. Note that as $N \to \infty$, $m_{av} \to 0$ and Eq. (2.65) reduces to Eq. (2.64).
In [31], Kanter and Sompolinsky have also shown the dependance of R on the load parameter α, because as α increases, the number of spurious states becomes bigger and affects the radius of attraction of the memories. Trough measures of R for asynchronous and synchronous dynamics, they have found that R follows approximately a straight line as a function of α, i.e.

$$R \simeq 1 - \alpha, \qquad \text{with } 0 \leq \alpha \leq 1. \tag{2.66}$$

Thus, the radius of attraction $R(\alpha)$ is unit as $\alpha \to 0$ and decreases monotonically to zero as $\alpha \to \alpha_c$.

2.3 A continuous deterministic model

From the original Hopfield model [28] we can construct a model based on continuous variables and responses with all the significant behavior of the previous one, as described by Hopfield in [29]. The extension of the analysis into this domain is very important, because real neurons have input-output relations which may be more reasonably modeled on a continuous basis than on an all-or-none basis.

[26] In [31] this definition is given without the factor 1/2, because as discussed in Personnaz et al. [47], they exemplified the model by considering stored patterns with equal distance $N/2$ between them. This implies that the minimum overlap a probe vector can have with ξ^μ to be attracted to it is 0, i.e. at most $N/2$ different components.

2.3.1 Additive model

Following Chapter 17 of Gerstner et al. [22] and Chapter 14 of Haykin [25], in particular pp. 698-706, we derive an equation for the dynamics of the network in the continuous case. A neuron is surrounded by a cell membrane, which is a rather good insulator. If a short current pulse $I(t)$ is injected into the neuron, the additional electrical charge $q = \int I(t')dt'$ has to go somewhere: it will charge the cell membrane. The cell membrane therefore acts like a capacitor of capacity C. Because the insulator is not perfect, the charge will, over time, slowly leak through the cell membrane, which therefore can be characterized by a finite leak resistance R.

Let consider neuron j and let the output variable V_j for neuron j have the range $-1 \leq V_j \leq +1$ and suppose that V_j is a continuous and monotone-increasing function of the instantaneous input u_j to neuron j. Typically the input-output relation between u_j and V_j is described by a sigmoid function g with asymptotes -1 and $+1$, i.e.

$$V_j = g(u_j) \implies u_j = g^{-1}(V_j) \tag{2.67}$$

where we have assumed that the inverse of the nonlinear activation function exists. Figure 2.1 shows a plot of a standard nonlinear sigmoidal function $g(u)$ and of its inverse $g^{-1}(V)$.

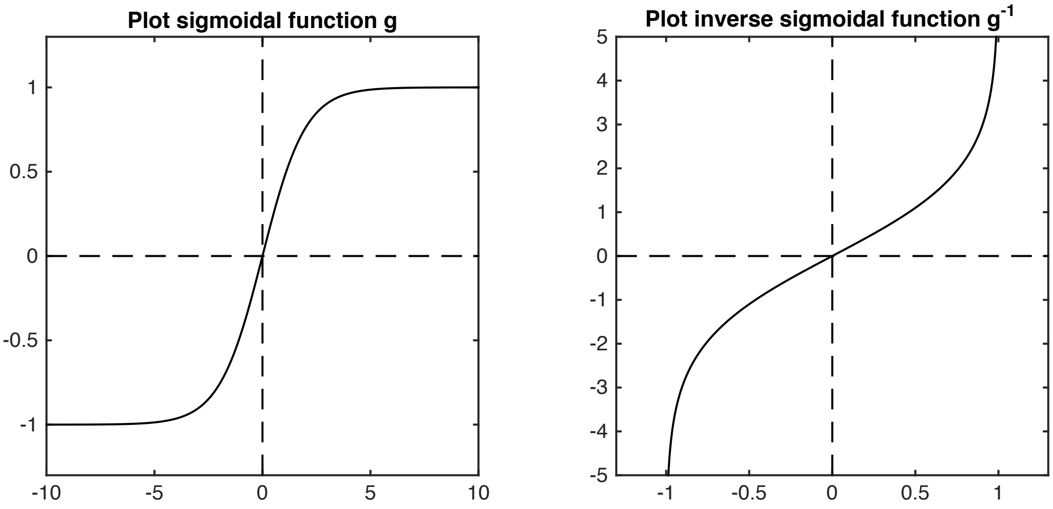


Figure 2.1: Plot of a standard sigmoidal function $g(u) = \tanh(u/2)$ and its inverse $g^{-1}(V) = -\log(\frac{1-V}{1+V})$.

From the biological point of view we have that:

- u_j can be interpreted as the mean soma potential of neuron j from the total effect of its excitatory and inhibitory inputs or the membrane potential (that corresponds to h_j defined in Eq. (2.3) with $I_j = 0$)
- V_j can be seen as the output potential, i.e. the short-term average of the firing rate of neuron j.

In physical terms, the synaptic weights $T_{j1}, \ldots, T_{jN}$ represents *conductances* and the respective inputs $V_1(t), \ldots, V_N(t)$ represents *potentials.* These inputs are applied to a current-summing junction, that acts as a summing node for the input currents and is characterized as follows:

- Low input resistance
- Unity current gain
- High output resistance

The total current flowing towards the input node of the nonlinear element (activation function) in Fig. 2.2 is therefore

$$\sum_{i=1}^{N} T_{ji} V_i(t) + I_j$$

where the first (summation) term is due to the stimuli $V_1(t), V_2(t), \ldots, V_N(t)$ acting on the synaptic weights (conductances) $T_{j1}, \ldots, T_{jN}$ respectively and the second term is due to the current source I_j representing an externally applied bias.

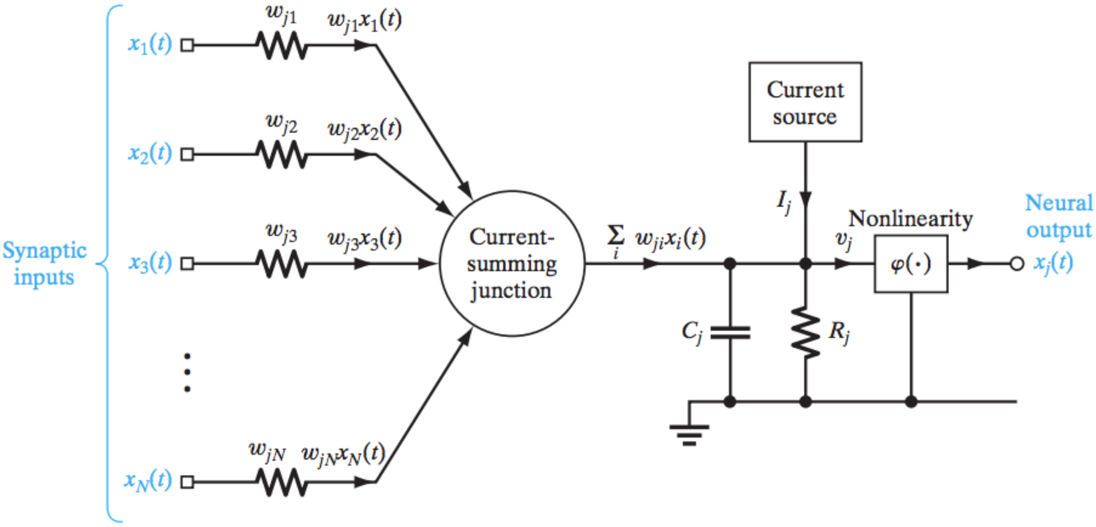


Figure 2.2: Additive model of a neuron. This figure is taken from Haykin's book [25], that uses a different notation from the one used here: w_{ji}, for $i = 1, \ldots, N$, corresponds to the synaptic weights T_{ji}, $x_j(t)$, with $j = 1, \ldots, N$, to the inputs $V_j(t)$, φ to the activation function g and v_j the local field u_j.

Let $u_j(t)$ denote the induced local field at the input of the nonlinear activation function $g(\cdot)$. We may then express the total current flowing away from the input node of the nonlinear element as the sum of two terms:

$$\frac{u_j(t)}{R_j} + C_j \frac{du_j(t)}{dt}$$

where the first term is due to leakage resistance R_j and the second term is due to leakage capacitance C_j. From *Kirchoff's current law*, we know that the total current flowing towards any node of an electrical circuit is zero. By applying Kirchoff's current law to the input node of the nonlinearity in Fig. 2.2, we get

$$C_j \frac{du_j(t)}{dt} + \frac{u_j(t)}{R_j} = \sum_{i=1}^{N} T_{ji} V_i(t) + I_j \tag{2.68}$$

Therefore, the dynamics of the system is given by a resistance-capacitance (RC) charging equation[27] that determines the rate of change of u_j

$$C_j\Big(\frac{du_j}{dt}\Big)= \sum_i T_{ji}V_i - \frac{u_j}{R_j} + I_j \tag{2.69}$$

where $T_{ji}V_i$ represents the electrical current input to cell j due to the present potential of cell i and T_{ji} is the synapse efficacy. Then C_j is the input capacitance of the cell membranes and R_j the transmembrane resistance.
We assume the symmetry of T, but we do not require that T has a zero diagonal and thus do not have self-loops.

2.3.2 The energy function

The energy associated to this dynamical system is given by (Hopfield, [29])

$$E = -\frac{1}{2}\sum_i\sum_j T_{ij}V_iV_j + \sum_i \frac{1}{R_i}\int_0^{V_i} g_i^{-1}(V)dV - \sum_i I_iV_i \tag{2.70}$$

As in the discrete case, it can be proved that E is a Lyapunov function:

Proposition 2.3.1. *The energy function (2.70) is a Lyapunov function in the sense of Def. (2.2.8).*

Proof. We want to prove that the Lyapunov conditions of Def. (2.2.8) holds. In reality, in a sense to be specified, the three conditions of Def. (2.2.7) are satisfied:

1. $\frac{\partial}{\partial V_k}E(V) = -\sum_j T_{kj}V_j + \frac{u_k}{R_k} - I_k, \quad \forall\ V_k$, implying[28] $E(V)$ is continuous in its interval of definition with respect to all components of V.

2. We have that $V_i \in [-1,1]$ and g^{-1} in such interval are inferior bounded, thus the energy function is bounded and E admits at least a minimum.
   To prove that E is positive definite we can note that the second factor in (2.70) is always positive, thus
   $$E \geq -\frac{1}{2}\sum_i\sum_j T_{ij}V_iV_j - \sum_i I_iV_i \tag{2.71}$$
   Now the right-hand side coincides with the energy function of the discrete case (2.43) with $I_i \neq 0$. Thus arguing in the same way we gain the result $E(V) \geq c$, with $c = \min_{\xi^\mu} E(\xi^\mu)$, since now the values of E in the stored patterns are not equal.
   As in the discrete case, the correct energy function has to be shifted in order to have $E(\xi^\mu) = 0,\ \forall\ \mu$. Here the values $c_\mu = E(\xi^\mu)$ are different for the various stored patterns because of the presence of the second and third term on the right-hand side of Eq. (2.70). Thus, we will obtain only local Lyapunov functions, one for each ξ^μ. In the basin of attraction $Basin(\xi^\mu)$ of a stored pattern ξ^μ the energy (2.70) is replaced with $E^\mu = E - c_\mu$, where
   $$c_\mu = E(\xi^\mu) = -\frac{N-p}{2} - \sum_i I_i\xi_i^\mu + \sum_i \frac{1}{R_i}\int_0^{\xi_i^\mu} g_i^{-1}(V)dV$$

[27] For the connection between Eq. (2.69) and the discrete dynamics equation (2.6) see Appendix B .

[28] If $f : A \to \mathbb{R}^N$ is differentiable at $c \in A$, then there exists strictly positive numbers δ, K such that $\|x - c\| < \delta$, then $\|f(x) - f(c)\| \leq K\,\|x - c\|$. In particular, it follows that f is continuous.

where the second and third addend in the last equality are obviously constant. Therefore, in that neighborhood $Basin(\xi^\mu)$ we get $E^\mu(\xi^\mu) = 0$ and $E^\mu(V) > 0$ for every $V \in\ Basin(\xi^\mu)$.

3. The time derivative of E is given by

$$\frac{dE}{dt} = -\sum_i \frac{dV_i}{dt}\Big(\sum_j T_{ij}V_j - \frac{u_i}{R_i} + I_i\Big) \tag{2.72}$$

Using (2.69) and then the inverse relation that defines V_i in terms of u_i we gain

$$\begin{aligned} \frac{dE}{dt} &= -\sum_{i=1}^{N} C_i\Big(\frac{dV_i}{dt}\Big)\Big(\frac{du_i}{dt}\Big) = -\sum_{i=1}^{N} C_i\Big(\frac{dV_i}{dt}\Big)\frac{d}{dt}g_i^{-1}(V_i) \\ &= -\sum_{i=1}^{N} C_i\Big(\frac{dV_i}{dt}\Big)^2\frac{d}{dV_i}g_i^{-1}(V_i) \end{aligned} \tag{2.73}$$

Since C_i is positive and $g_i^{-1}(V_i)$ is a monotone increasing function, it follows that

$$\frac{d}{dV_i}g_i^{-1}(V_i) \geq 0, \qquad \forall\ V_i \tag{2.74}$$

We also note that

$$\Big(\frac{dV_i}{dt}\Big)^2 \geq 0, \qquad \forall\ V_i \tag{2.75}$$

Therefore all the factors that make up the sum on the right-hand side of Eq. (2.73) are nonnegative. In other words, for the energy function E defined in Eq. (2.70), we have

$$\frac{dE}{dt} \leq 0 \qquad \text{and} \qquad \frac{dE}{dt} = 0 \implies \frac{dV_i}{dt} = 0\ \ \forall i \tag{2.76}$$

Remark. The condition $\frac{dV_i}{dt} = 0\ \ \forall i$ corresponds to the definition of networks stability (Def. 2.2.4). Thus it implies that the vectors V satisfying this property are the fixed points of the dynamics.

Thus we can write

$$\frac{dE}{dt} < 0 \quad \text{except at fixed points} \tag{2.77}$$

In conclusion, we have proved that E is a Lyapunov function in the sense of Def. (2.2.8) and can be made a classical Lyapunov function (as in Def. (2.2.7)) only locally in neighborhoods of the stored patterns[29]. □

Therefore, the time evolution of the continuous Hopfield model described by the system of nonlinear first-order differential equation (2.69) is a motion in state space, which seeks out minima of the energy (Lyapunov) function E and comes to a stop at such fixed points. Furthermore, from the Lyapunov theorem 2.2.10 the Hopfield network is asymptotically stable; the attractor fixed-points are the minima of the energy function and vice versa.

[29] Also in the case of local Lyapunov function, exists a theorem that guarantees the asymptotic Lyapunov stability analogous to theorem 2.2.10, as discussed by Giesl and Hafstein in [23].

2.3.3 Relation between the stable states of the two models

The Hopfield network may be operated in a continuous mode or discrete mode, depending on the model adopted for describing the neurons. The continuous mode of operation is based on an additive model, on the other hand the discrete mode of operation is based on the McCulloch-Pitts model.
Following the contributions of Hopfield [29] and Haykin [25], pp. 706-708, we may readily establish the relationship between the stable states of the continuous Hopfield model and the corresponding ones of the Hopfield discrete model by redefining the input-output relation for a neuron such that two simplifying characteristics are satisfied:

- The output of a neuron V_i has the asymptotic values

$$V_i = \begin{cases} +1 & \text{for } u_i = +\infty \\ -1 & \text{for } u_i = -\infty \end{cases} \tag{2.78}$$

- The midpoint of the activation function of a neuron lies at the origin

$$g_i(0) = 0 \tag{2.79}$$

Correspondingly, we may set the bias I_i equal to zero for all i.
In formulating the energy function E for a continuous Hopfield model, the neurons are permitted to have self-loops. A discrete Hopfield model, on the other hand, needs not to have self-loops. We may therefore simplify our discussion by setting $T_{ii} = 0$ for all i in both models.
In light of these observations, the energy function of the continuous Hopfield model given in Eq. (2.70) is given by

$$E = -\frac{1}{2}\sum_i \sum_{j,\ j\neq i} T_{ij}V_iV_j + \sum_i \frac{1}{R_i}\int_0^{V_i} g_i^{-1}(V)dV \tag{2.80}$$

whereas the discrete energy is given by

$$E = -\frac{1}{2}\sum_i \sum_{j,\ j\neq i} T_{ij}V_iV_j \tag{2.81}$$

We may rewrite Eq. (2.80) by scaling the gain function g in the following way

$$V_i = g_i(u_i) \quad \text{replaced by} \quad V_i = g_i(\lambda u_i) \tag{2.82}$$

and

$$u_i = g_i^{-1}(V_i) \quad \text{replaced by} \quad u_i = \frac{1}{\lambda} g_i^{-1}(V_i) \tag{2.83}$$

This scaling changes the steepness of the sigmoid gain curve without altering the output asymptotes. Therefore, Eq. (2.80) becomes

$$E = -\frac{1}{2}\sum_i \sum_{j,\ j\neq i} T_{ij}V_iV_j + \frac{1}{\lambda}\sum_i \frac{1}{R_i}\int_0^{V_i} g_i^{-1}(V)dV. \tag{2.84}$$

The integral is zero for $V_i = 0$ and positive otherwise, getting very large as V_i approaches ±1.

- In the high-gain limit $\lambda \to \infty$ the second term on Eq. (2.84) becomes negligible and the location of the maxima and minima of the full energy expression in the continuous case become the same of (2.81). The only stable points of the very high gain, continuous system correspond to the stable points of the discrete system. In particular, the maxima and minima lie at the corners of the N-dimensional hypercube space $[-1, 1]^N$, representing the domain of operation of the equation of motion ($-1 \leq V_i \leq 1$).

- For large but finite λ the second term in (2.80) begins to contribute. In particular, this contribution is large and positive near all the surfaces, edges and corners of the unit hypercube that defines the state space of the model. On the other hand, the contribution is negligibly small at points that are far from the surface. This leads to an energy surface that still has its maxima at corners but the minima become displaced slightly toward the interior of the hypercube.

- As λ decreases minima disappear one at time making the minimum and a saddle point join together. For very small λ this term dominates and the only minimum is at $V_i = 0$.

Figure 2.3 depicts the *energy contour map* or *energy landscape* for a continuous Hopfield model using two neurons. The outputs of the neurons define the two axes of the map. The lower left-and upper right-hand corners of Fig. 2.3 represent stable minima for the limiting case of infinite gain; the minima for the case of finite gain are displaced inward. The flow to fixed points (i.e. stable minima) may be interpreted as the solution to the minimization of the energy function E defined in Eq. (2.70).

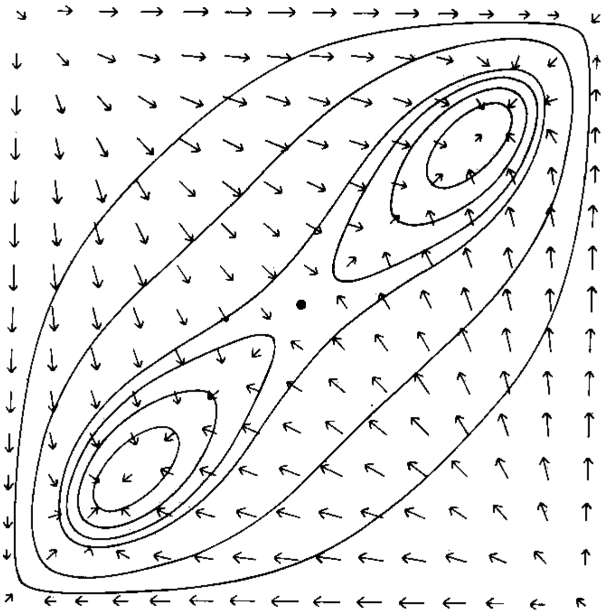


Figure 2.3: An energy contour map for a two-neuron, two-stable state system. The ordinate and abscissa are the output of the two neuorns. Stable states are located near the lower left and upper right corners and unstable extrema at the other two corners. The arrows show the motion of the state. This motion is not generally perpendicular to the energy contours. (Figure taken from Hopfield, 1984 [29])

2.3.4 Cohen-Grossberg Theorem

Hopfield was not the first person to use a Lyapunov energy function to prove that a system reaches equilibrium. Cohen and Grossberg [14], have proven an even stronger result for certain continuous neural networks with less restriction on the functions and parameters. Furthermore, this model represents a generalization of many neural network models that are capable for content addressable memory such as additive neural networks, cellular neural networks and bidirectional associative memory networks, Hopfield model and also biological models such as Lotka-Volterra models of population dynamics.
In Cohen and Grossberg (1983) [14], a general principle for assessing the stability of a certain class of neural networks is described by the following system of coupled non linear differential equations[30]:

$$\frac{d}{dt}u_i = a_i(u_i(t))\Big[b_i(u_i(t)) - \sum_{j=1}^{N} c_{ij}g_j(u_j(t))\Big], \quad i = 1, \dots, N \tag{2.85}$$

where $u_i(t)$ denotes the state of the ith neuron at time t, c_{ij} are the interaction coefficients, $g_j(\cdot)$ is the activation function, $a_i(\cdot)$ and $b_i(\cdot)$ represent the amplification function and behaved function at time t.
According to Grossberg and Cohen, this class of neural networks admits a Lyapunov function defined as

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}g_i(u_i)g_j(u_j) - \sum_{i=1}^{N}\int_0^{u_i} b_i(\lambda)g_i'(\lambda)d\lambda \tag{2.86}$$

where

$$g_i'(\lambda) = \frac{d}{d\lambda}(g_i(\lambda)) \tag{2.87}$$

In these definitions we require the following conditions to hold:

- The synaptic weights of the network are *symmetric*: $c_{ij} = c_{ji}$.
- The function $a_i(u_i)$ satisfies the condition for *nonnegativity*: $a_i(u_i) \geq 0$.
- The nonlinear input-output function $g_i(u_i)$ satisfies the condition for *monotonicity*:

$$g_i'(u_i) = \frac{d}{du_i}g_i(u_i) \geq 0$$

We may now formally state *Cohen-Grossberg theorem*, as presented in [25], pp. 723-724 :

Theorem 2.3.2 (Cohen-Grossberg theorem). *Provided that the system of nonlinear differential equation (2.85) satisfies the conditions of symmetry, non negativity and monotonicity, the Lyapunov function E of the system defined by Eq. (2.86) satisfies the condition*

$$\frac{dE}{dt} \leq 0$$

Once this basic property of the Lyapunov function E is in place, global stability of the system follows from Lyapunov's Theorem (2.2.10).

[30] This equation was derived by Grossberg in the 1960s and 1970s generalizing the *Additive and Shunting models* to a class of dynamical systems that included these models as well as non-neural biological models. He proved content addressable memory theorems for this problem and then together with Cohen he discovered in 1981 a Lyapunov function useful for studying the stability of the network.
We recall that the additive models are models like the one described in section 2.3.1, whereas the shunting model is a similar one, that takes into account that the inputs can be excitatory and inhibitory.

Table 2.1: Correspondence between Cohen-Grossberg theorem and Hopfield model

Cohen-Grossberg Theorem	Hopfield model
u_i	$C_i u_i$
$a_i(u_i)$	1
$b_i(u_i)$	$-(u_i/R_i) + I_i$
c_{ij}	$-T_{ij}$
$g_i(u_i)$	$g_i(u_i)$

Hopfield model as a special case of the Cohen-Grossberg theorem

By comparing the general system of Eq. (2.85) with the system of Eq. (2.69) for a continuos Hopfield model, we make the correspondence between Hopfield model and Cohen-Grossberg theorem, that are summarized in Table 2.1. The use of this table in Eq. (2.86) yields the following Lyapunov function for the continuous Hopfield model:

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij} g_i(u_i) g_j(u_j) - \sum_{i=1}^{N}\int_0^{u_i}\Big(\frac{u_i}{R_i} - I_i\Big) g_i'(u)du \tag{2.88}$$

We observe that:

1. $g_i(u_i) = V_i$
2. $\int_0^{u_i} g_i'(u)du = \int_0^{V_i} dV = V_i$
3. $\int_0^{u_i} u g_i'(u)du = \int_0^{V_i} u dV = \int_0^{V_i} g_i^{-1}(V)dV$

Basically, relations 2 and 3 result from the use of $V = g_i(u)$. Note that although $g_i(u)$ must be a nondecreasing function of the input u, it does not need to have an inverse in order for the generalized Lyapunov function of Eq. (2.88) to hold.
Thus, the use of these observations in the Lyapunov function of Eq. (2.88) yields a result identical to that we defined earlier in Eq. (2.70), but more general.

2.4 Improvements of the Hopfield model

From previous discussion it is clear how the Hopfield model with Hebb's learning rule has many weaknesses: the symmetry of the synapses, the full-connectivity, the limitation of the model to random uncorrelated patterns and many others. All these characteristics restrict the applicability of Hopfield model to biological and practical situations. No amount of lifting of simplifications will closely approximate the full glory of an assembly of real live neurons. Yet, as the grossest assumptions are replaced by more realistic ones and as the model is modified to account for more complex types of behavior without a significant loss in its basic functional features and in its effectiveness, the model gains plausibility. The lifting of simplifications is not performed as an end in itself. If the more complicated system functions in a qualitative way that can be captured by the simplified system, then the complication is removed and analysis continues with the simplified system (Amit, [4]).
We will present now some changes that can be made to the model to improve it, following also Amit's book [4] (in particular Chapters 7 and 8) .

2.4.1 Asymmetric neural networks

Hopfield network [28] has a synaptic matrix which is symmetric. In the previous discussion we have used strongly the hypothesis of having symmetrical connections, but this is not the case of human brain. From medical studies it is well known that for a given neuron the synapses are of only one type (excitatory or inhibitory) and with the requirement of symmetry this would imply two populations of neurons not connected to each other.
Another drawback of symmetric networks is the presence of spurious stable states. Much work has been done on the nature and origin of spurious states (e.g. [4]). A small amount of noise can be useful for escaping from them. Such noise can be added by introduction of weak random asymmetry into a symmetric well functioning Hopfield network, as done by Feigelman and Ioffe in [17] and Sompolinsky in [55]. However, such addition of noise does not change the qualitative performance very drastically, even at high level of asymmetry.
For these reasons various attempts go in the direction of generalizing the Hopfield model to the case of asymmetric synapses, i.e. such that the value of the synaptic connection in the $i \to j$ direction is not correlated with the value of the connection in the direction $j \to i$.
Hopfield network with asymmetric synaptic matrix can also be referred to as *Asymmetric Hopfield Network*. The dynamics of that type of network has been discussed by Derrida, Gardner and Zippellius in [16], Parisi in [44], Crisanti, Sompolinsky and Sommers in [56] and by many other researchers. One of the results of these studies is the enormous robustness of the Hopfield model to structural changes in the model. Asymmetries, dilution, delays, etc. do not affect the retrieval properties significantly and sometimes even improve the performance of the network, as we will see.
Asymmetry in the synaptic matrix can be introduced by

1. *Asymmetric learning rule.* There is no differentiation in the post-synaptic and pre-synaptic neuron in a symmetric learning rule. If pre-synaptic and post-synaptic neurons are differentiated, then the synaptic matrix will be asymmetrical.

2. *Random changes.* In a symmetric synaptic matrix, asymmetry can be introduced by randomly selecting and changing the synaptic weights to some other value. Random selection of a pair of synaptic weights and swapping the values will also introduce asymmetry. The number of random changes depends on the degree to which asymmetry is to be introduced.

3. *Dilution.* Dilution is a special case of random changes. Synaptic values are randomly selected from the symmetric synaptic matrix and these selected values are changed to zero. These changes in the synaptic matrix can be viewed as representing *Diluted Hopfield networks*, i.e. networks which are not fully connected. The study of this type of networks is reported for example in [4] and in [16]. Although dilution and asymmetry are two different and distinct concepts most of the relevant works reported in literature uses dilution as a means for introducing asymmetry in Hopfield network. In other words, dilution is used to achieve asymmetry but symmetric dilution does not lead to asymmetry.

We now proceed to describe all this in detail.

Dale's principle

As we mentioned, all the synapses, which connect a given neuron to others, are either excitatory or inhibitory. Even though exceptions are known, this so-called *Dale's law* seems

widely accepted for biological networks. What is implied is that a given type of neuron can release one type of neurotransmitter and hence can produce a single type of synapse, either excitatory or inhibitory. Clearly, a network obeying this law cannot have a symmetric synaptic connectivity matrix: neurons may receive inputs excitatory and inhibitory, but send outputs of a single type.
It is customary to use the following version of Dale's principle [48]:

Statement (Dale's law)**.** *A given neuron has the same physiological action at all its synapses. If all its synapses are excitatory (inhibitory), then the neuron is said to be excitatory (inhibitory).*

Another version of this law is mathematical [30]: Suppose u_i denotes the activity of the i-th neuron, $i = 1, \dots, N$. We assume that increasing (decreasing) u_i corresponds to depolarization (hyper-polarization). Suppose the network of such neurons is governed by a dynamical system of the form given in Eq. (2.69), where T_{ij} represents the synaptic weight from neuron j to neuron i. Applying Dale's principle we gain:

- If $T_{ij} \geq 0$ for all i, then the j-the neuron is *excitatory.*
- If $T_{ij} \leq 0$ for all i, then the j-the neuron is *inhibitory.*

It is easy to see that if $T_{ij} > 0$, then increasing the j-th neuron activity u_j facilitates increasing the i-th neuron activity u_i. Hence, the j-th neuron has an excitatory effect on the i-th neuron. Thus, it follows that the effect is the same for any other neuron, provided that the neurons are connected (that is, if the synaptic coefficient is not identically zero).

Remark. We observe that the coefficient T_{jj} is not subject to Dale's principle.

Asymmetry and Dilution

An asymmetric and diluted version of Hopfield network was proposed by Derrida et al. in [16], where an exact analytic solution of the dynamics of a standard ANN (*Attractor Neural Network*), whose connections have been randomly diluted to a level where the mean number of remaining synapses per neuron is smaller than lnN, is proposed. The significance is not that it presents a state affairs closer to biological reality than the standard model. Quite the contrary, it is probably less so. After all, while cortical connectivity is certainly not symmetrical and even in small regions is not complete, it remains true that average connectivity is a few thousand per neuron. For a network to have $lnN > 1000$, N must be a super-astronomical number.
Starting from the synaptic matrix defined in Eq. (2.16), the *diluted version* is represented by:

$$J_{ij} = c_{ij} N T_{ij} = c_{ij} \sum_{\mu=1}^{p} (\xi_i^\mu \xi_j^\mu - \delta_{ij}) \tag{2.89}$$

where $c_{ij} \in \{0, 1\}$ is an independent random number which represents dilution and asymmetry. The dilution matrix is chosen with the distribution

$$Pr(c_{ij}) = \frac{c}{N} \delta(c_{ij} - 1) + \Big(1 - \frac{c}{N}\Big) \delta(c_{ij}) \tag{2.90}$$

where c is a constant representing mean connectivity per neuron and δ is the Dirac delta function, that satisfies $\delta(0) = 1$ and $\delta(x) = 0$ for $x \neq 0$. Thus, this is a random dilution which gives the probability c/N for a synaptic efficacy to remain intact and $(1 - c/N)$ for it

to be set to zero.
Furthermore since $c/N \to 0$ as $N \to \infty$, the network is heavily diluted and the analysis is carried out under the restriction that

$$c \ll lnN \tag{2.91}$$

In agreement with [16], the neurons are updated according to the following rule

$$V_i(t+\Delta t) = \begin{cases} +1 & \text{with probability } \ (1+\exp(-2h_i(t)/T_0))^{-1} \\ -1 & \text{with probability } \ (1+\exp(2h_i(t)/T_0))^{-1} \end{cases} \tag{2.92}$$

where $h_i(t) = \sum_{j=1}^{N} J_{ij} V_j(t)$ and T_0 measures the strength of the noise and is therefore called "temperature"[31].
In this context it is natural to redefine the load parameter (and thus the storage capacity), relative to the number of connections per neuron, at a reduced temperature T as

$$\bar{\alpha} \equiv \frac{p-1}{c} \quad \text{and} \quad T = \frac{T_0}{c} \tag{2.93}$$

Through simulation at 0 temperature[32], it is found that the critical value α_c of α is

$$\alpha_c = \frac{2}{\pi} = 0.6366\ldots \tag{2.94}$$

which is larger than the value 0.15 of the non-diluted symmetric case.

- For $\alpha < \alpha_c$ two initial configurations close to a stable state vector remain close to the stable state vector but do not become identical.
- Above α_c ($\alpha > \alpha_c$) the system does not remember (i.e. it does not converge to a stored pattern).

Therefore, in the Derrida-Gardner-Zippellius's study [16] they proved that the retrieval properties of an ANN with a standard storage prescription sustain any degradation that is not intelligently intended to damage its retrieval properties and an extremely randomly diluted ANN can perform at least as well as the standard, fully connected, symmetric network.

Asymmetric Learning Rule

We now proceed to describe an asymmetric version of Hopfield network using asymmetric learning rule, as reported by Gardner-Mertens-Zippellius in [19].
In 1949 Hebb [26] suggested that the efficacy of an excitatory synapse should increase when both the pre- and the postsynaptic neuron are active and that it should decrease when the postsynaptic neuron is active while the presynaptic neuron is silent. Taking into account that type of learning rule and following [19], a synaptic matrix for a network of bipolar neurons can be constructed in that way

$$\Delta\tilde{J}_{ij} = \frac{1}{2}\sum_{\mu=1}^{p}(\xi_i^\mu + 1)(\xi_j^\mu) \tag{2.95}$$

[31] We called this parameter temperature in analogy with the Ising model, which can be shown to be isomorphic to that of Hopfield, and the thermodynamic approach developed for that model.

[32] If we consider Ising model with temperature $T = 0$, we have the Hopfield model, always referring to the isomorphism that exists between them.

In other words, we consider a Hebbian learning mechanism, which gives rise to a change in the synaptic efficacy $\tilde{J}_{ij}$, operating between the presynaptic neuron j and the postsynaptic neuron i, only if the postsynaptic neuron is active[33]. An excitatory synapse is changed as suggested by Hebb, whereas inhibition becomes less effective if both neurons are simultaneously active and becomes more effective if the postsynaptic one is active and the presynaptic one is not.

Remark. We note that the synaptic efficacies in Eq. (2.95) are not symmetric, because our learning rule differentiates between post- and presynaptic neurons. Hence there is no a priori Lyapunov functional, which decreases monotonically during the time evolution of the network.

Each active neuron $V_j = 1$ contributes to the postsynaptic potential of neuron i through

$$h_i(t) = \frac{1}{2}\sum_{j\neq i} \tilde{J}_{ij}(V_j(t) + 1) \tag{2.96}$$

according to the synaptic strength $\tilde{J}_{ij}$ (2.95). As before, if the postsynaptic potential exceeds a threshold U_i then the postsynaptic neuron is activated

$$V_i(t + \Delta t) = sgn(h_i(t) - U_0) \tag{2.97}$$

where for simplicity we have considered a uniform threshold $U_i = U_0 > 0$ and the updating procedure is given by

$$V_i(t+\Delta t) = \begin{cases} +1 & \text{with probability} \quad \left[1 + \exp\left(-\dfrac{2h_i(t) - 2U_0}{T_0}\right)\right]^{-1} \\ -1 & \text{with probability} \quad \left[1 + \exp\left(\dfrac{2h_i(t) - 2U_0}{T_0}\right)\right]^{-1} \end{cases} \tag{2.98}$$

By introducing a strong dilution and asymmetry parameter c_{ij} the final synaptic weights are obtained as

$$J_{ij} = c_{ij}\tilde{J}_{ij} \tag{2.99}$$

with $c_{ij} \in \{0, 1\}$, chosen with the same distribution as before (2.90).
Under the condition $c/N \to 0$ as $N \to \infty$ with reduced temperature $T = T_0/c$ and reduced threshold $U = U_0/c$, the load parameter with respect to existing coupling is obtained as

$$\alpha = \frac{p-1}{c} \tag{2.100}$$

We note that this relation for α is the same as that given in Eq. (2.93). Then, from simulations the capacity of the network is found to depend on the threshold U_i of the post-synaptic neuron: it is optimal for $U_i \approx 0.1$ and no retrieval is possible for $U_i > 0.5$.

2.4.2 Biased Patterns

As discussed in section 2.2.3, we have stored p patterns ξ_i^μ, $\mu = 1, \ldots, p$, $i = 1, \ldots, N$, where $\xi_i^\mu = \pm 1$ with equal probability. Hence, each stored pattern have on average 50% of

[33] We should remember that in the case in exam, if we change the value of $\tilde{J}_{ij}$, the value of $\tilde{J}_{ji}$ is not affected.

the neurons active (+1) and 50% passive (−1). Consequently, in the process of retrieval when the network enters an attractor, 50% of the neurons are active on average. These vectors are termed as *unbiased or uncorrelated patterns*; namely, in a large network

$$\frac{1}{N}\sum_i \xi_i^\mu \xi_i^\nu = 0, \qquad \text{if } \mu \neq \nu. \tag{2.101}$$

This situation is unsatisfactory for several reasons:

1. Neurophysiological evidence indicates that mean firing rates are significantly lower than 50%.

2. Hopfield dynamics is symmetric with respect to the interchange of firing by non-firing ($V_i \to -V_i$). Hence, with every stored pattern the network stores the reversed pattern as well. Usually, this type of symmetry, when undesired, is lifted by an external field (threshold). But here, if the learned patterns are of Hopfield random type, it is difficult to find a natural way of suppressing the doubling of the stored patterns.

3. More realistic networks have to confront the presence of correlated patterns in models of associative memory. A model, which deals with correlated patterns, has been proposed for example by Amit, Gutfreund and Sompolinksy in 1986, [5]. This model involves a much more complicated dependence of T_{ij} on the learned patterns. In order to learn and retrieve correlated patterns, one has to introduce non locality, either in the learning process or in the dynamics.

In [5], Amit et al. have studied associative memory of random patterns whose mean activities differ from 50%. Patterns of this type are called *biased patterns.* For instance, every component ξ_i^μ in a learned pattern can be chosen independently with probability $Pr(\xi_i^\mu)$:

$$Pr(\xi_i^\mu) = \frac{1}{2}(1+a)\delta(\xi_i^\mu - 1) + \frac{1}{2}(1-a)\delta(\xi_i^\mu + 1) \tag{2.102}$$

where $\delta(\cdot)$ is the delta Dirac function.

Definition 2.4.1. *The average*[34] *of each* ξ^μ*, also called* bias parameter*, is*

$$a = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu, \quad \textit{with } -1 \leq a \leq 1 \tag{2.103}$$

and the mean activity in each stored pattern *is*

$$A = \frac{1}{2}(1+a) \tag{2.104}$$

With such an asymmetric distribution of stored patterns, memories are necessarily *correlated* in the following way

$$\langle\langle \xi_i^\mu \xi_i^\nu \rangle\rangle = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu = \delta^{\mu\nu} + a^2(1-\delta^{\mu\nu}) \tag{2.105}$$

[34] As discussed in [5], this can also be generalized to the case

$$\frac{1}{N}\sum_i \xi_i^\mu = a_\mu, \quad \mu = 1, \dots, p$$

→ A naive application of Hopfield dynamics with the ξ^μ's chosen according to the asymmetric distribution (2.102) is catastrophic in that, even at small values of the bias parameter a, the stored patterns become unstable at very low storage levels. This is due to the fact that the noise generated by the other patterns in the retrieval of each pattern does not average to zero.

In fact, if we consider Eq. (2.19) in the usual Hopfield context, but with correlated stored patterns, we have

$$h_i = [T\xi^\mu]_i = \sum_{j=1}^{N} T_{ij}\xi_j^\mu = \frac{1}{N}\sum_{j=1}^{N}\sum_{\beta=1}^{p}\xi_i^\beta\xi_j^\beta\xi_j^\mu = \xi_i^\mu + \frac{1}{N}\sum_{\beta\neq\mu}\sum_{j}\xi_i^\beta\xi_j^\beta\xi_j^\mu \tag{2.106}$$

where we have separated the expression into a signal and a noise term and we have neglected in the summation $j \neq i$, taking into account that the diagonal terms are equal zero and so we are adding zero terms.

The noise term can be rewritten as

$$\frac{1}{N}\sum_{\beta\neq\mu}\sum_{j}\xi_i^\beta\xi_j^\beta\xi_j^\mu = \langle\langle\xi^\mu\xi^\beta\rangle\rangle\sum_{\beta\neq\mu}\xi_i^\beta = a^2\sum_{\beta\neq\mu}\xi_i^\beta \tag{2.107}$$

where the last factor on the right hand side can take values between $-(p-1)$ and $(p-1)$.

Arguing in the same way of section 2.2.4, the crosstalk term $C_i^\mu = \xi_i^\mu a^2 \sum_{\beta\neq\mu}\xi_i^\beta$ has not the same sign of ξ_i^μ if $C_i^\mu \geq 1$. Using that condition and that

- $\xi_i^\mu \geq -1, \ \forall\, i$
- $\sum_{\beta\neq\mu}\xi_i^\beta \geq -(p-1)$,

we have

$$-a^2(-1)(p-1) \leq \xi_i^\mu a^2\sum_{\beta\neq\mu}\xi_i^\beta \leq 1 \qquad \Longrightarrow \qquad a^2(p-1) \leq 1$$

and thus the patterns become destabilized when

$$(p-1)a^2 \leq 1. \tag{2.108}$$

To appreciate the scale of disaster consider, for example, the case where the mean level of activity is 5%. The value of the bias a can be deduced from the mean activity A

$$a = 2A - 1 \approx -0.9 \tag{2.109}$$

Substituting this in Eq. (2.108) we gain that p must not exceed 2 irrespective of the value of N and this is an absurd.

→ To overcome this difficulty, Hebb's rule is replaced with the following non-local learning rule

$$\tilde{T}_{ij} = \frac{1}{N}\sum_{\mu}(\xi_i^\mu - a)(\xi_j^\mu - a) \tag{2.110}$$

This equation is an expression of the non locality in this learning scheme. In getting modified each synapse has to be aware of the mean, allowed activity rate of the entire network. Then, these synaptic weights avoid catastrophe by shifting the noise back to a zero mean. Considered from a biological point of view, Eq. (2.110) can be interpreted as learning by modification of synaptic efficacies due to neural activity. If the network persists in some activity states $V_i = \xi_i^\nu$, then

$$\Delta\tilde{T}_{ij} = \frac{1}{N}(\xi_i^\nu - a)(\xi_j^\nu - a) \tag{2.111}$$

Since

$$\sum_i \xi_i^\nu = Na \tag{2.112}$$

we have

$$\sum_j \Delta\tilde{T}_{ij} = 0, \tag{2.113}$$

which states that the total modification of synapses on a given neuron is unchanged during learning. This is, of course, also a property of the original Hopfield description, if the patterns are uncorrelated and unbiased pattern, i.e. $\frac{1}{N}\sum_i \xi_i^\mu = 0$.

With this prescription (2.110), the system will then follow the dynamics of the Hamiltonian energy function

$$E = -\frac{1}{2N}\sum_{i,j}\sum_\mu (\xi_i^\mu - a)(\xi_j^\mu - a)V_iV_j \tag{2.114}$$

and the local field on neuron i is given by

$$h_i = \sum_{j,j\neq i}\tilde{T}_{ij}V_j = \frac{1}{N}\sum_{j\neq i}\sum_\mu (\xi_i^\mu - a)(\xi_j^\mu - a)V_j \tag{2.115}$$

Consider now the stability of the stored patterns by the simple signal-to-noise estimation. Thus, taking $V_i = \xi_i^\nu$ and substituting in Eq. (2.115), we find that the signal term is

$$S = \frac{1}{N}\sum_{j\neq i}(\xi_i^\nu - a)(\xi_j^\nu - a)\xi_j^\nu = (\xi_i^\nu - a)\frac{1}{N}\sum_{j\neq i}(\xi_i^\nu - a)\xi_i^\nu = (\xi_i^\nu - a)(1-a^2) \tag{2.116}$$

whereas the noise is determined by

$$W = \frac{1}{N}\sum_{j\neq i}\sum_{\mu\neq\nu}(\xi_i^\mu - a)(\xi_j^\mu - a)\xi_j^\nu = \sum_{\mu\neq\nu}(\xi_i^\mu - a)\frac{1}{N}\sum_{j\neq i}(\xi_j^\mu - a)\xi_j^\nu \tag{2.117}$$

Since in W, $j \neq i$ and $\mu \neq \nu$, its mean vanishes[35]. The effect of the noise has to be estimated by its variance, i.e. the mean of the square of the noise term[36] $\langle\langle W^2\rangle\rangle$, which is

$$\rho^2 \equiv \langle\langle W^2\rangle\rangle = \frac{(N-1)(p-1)}{N^2}(1-a^2)^2 \approx \frac{p}{N}(1-a^2)^2 \tag{2.118}$$

where the brackets indicate an average over all the components $j \neq i$ and the stored patterns $\mu \neq \nu$. Clearly, this network can have a macroscopic number of stable patterns. In fact, the lower limit of the signal is

$$|S| \geq (1-a^2)(1-|a|) \tag{2.119}$$

and hence

$$\frac{\rho}{|S|} \leq \sqrt{\frac{p}{N}}\Big(\frac{1}{1-|a|}\Big) \tag{2.120}$$

[35] For the proof of this statement see section 2.2.4.

[36] For a complete proof of this relation see Appendix 8.5.1, pp. 423-425, in Amit's book [4].

In contrast to the naive introduction of biased patterns, if

$$\alpha = \frac{p}{N} \ll (1-|a|)^2$$

then using that relation in (2.120) we get

$$\rho^2 \ll \frac{(1-|a|)^2}{(1-|a|)^2}|S|^2 \quad \Longrightarrow \quad \rho^2 \ll |S|^2 \tag{2.121}$$

Thus, the noise is essentially zero and the original patterns are stable at $T = 0$ in absence of fast noise. On the other hand, we have here an indication, even on this simple level, that the storage capacity decreases as a increases. Furthermore, the storage capacity is reduced compared to the capacity of the network storing unbiased patterns.

A more careful treatment of the probability distribution of the noise term has led in [5] to the conclusion that patterns are retrieved with no errors if

$$\alpha < \alpha_c = \frac{(1-|a|)^2}{2lnN} = (1-|a|)^2\alpha_c(0) \tag{2.122}$$

where $\alpha_c(0)$ denotes the storage efficiency provided by the naive signal-to-noise estimate. This decrease in the storage capacity with the bias a is not a feature of a proper network storing low activity patterns: a reasonable functioning network would increase its capacity as the level of activity is lowered (as the bias is increased).

Therefore, this modification restores the ability of the model to store a macroscopic number of patterns, but as proved in [5], spurious states are found to plague the dynamics. As a increases, spurious states start dominating the dynamics, even though the stored patterns remain locally stable. In addition, the number of spurious states is found to increase with a and they become the absolute minima of the energy as a increases. All this indicates that to have a network which can effectively store and retrieve biased patterns, it does not suffice to modify the couplings. In fact, from a biological perspective, the dynamical process should be modified as well to be consistent with the levels of activity of the stored patterns. This is implemented by adding a global constraint, which restricts the configuration space to states V_i, such that they obey

$$\frac{1}{N}\sum_i V_i = a \tag{2.123}$$

In other words, the dynamics of the network is restricted to a part of the full space of 2^N possible network states, where the mean activity is in the neighborhood of the activity in the biased patterns.

An immediate consequence of Eq. (2.123) is the breaking of the symmetry $V_i \to -V_i$. Thus, symmetry has led to the doubling of every solution for the attractors, as was discussed in section 2.2.6. Two memorized states which have all the corresponding components reverse are no longer equivalent[37]. In particular, it implies that on storing a set of patterns, their reversed states are no longer attractors unless $a = 0$. Note that these reversed patterns have not become unstable. They have only been pushed out of the space of states visited by the dynamics.

The consequences of the restricted dynamics are then analyzed by Gutfreund-Amit-Sompolinsky [5] using the replica symmetric mean-field theory. From that, they observed that the global constraint suppresses spurious states and leads to the unexpected result that the storage

[37] We recall that in the Hopfield model if V is a stored pattern, V and $-V$ are equivalent because they are eigenvectors relative to the same eigenvalue and have the same characteristics.

capacity is higher than that of the unbiased network, $\alpha_c(a) > \alpha_c(0)$, up to very high values of the bias ($|a| \simeq 0.99$).
Technically, the imposition of the constraint can be affected either in a hard or in a soft way. A biological system is not expected to impose such a global constraint rigidly. Hence, the rigid constraint (2.123) is relaxed by adding a cost term to the energy (2.114)

$$E_c = \frac{\lambda}{2N}\Big(\sum_i V_i - Na\Big)^2 \tag{2.124}$$

where λ is a positive number, that measures the stiffness of the constraint. Such a term represents an extra uniform contribution to the local field of a neuron, proportional to the deviation of the level of activity from the normal level. The membrane potentials (local fields) are then modified by adding

$$\tilde{h}_i = -\lambda\Big(\frac{1}{N}\sum_{j=1}^{N} V_j - a\Big) \tag{2.125}$$

which represents a uniform level depending on the momentary activity in the network. This additional term can be interpreted in two ways:

- As an external control by potential inputs, which ensures that the activity in the network be constantly corrected towards the desired mean.

  → If it is too high, $\tilde{h}_i$ dominates on h_i and becomes inhibitory (hyper-polarizing).

  → If it is too low, h_i prevails and all neurons receive an equal excitatory (depolarizing) input.

  Moreover, the magnitude of these inputs increases with the magnitude of the deviation from the desired mean activity a.

- As a uniform inhibitory contribution to the synapses of efficacy $-\lambda/N$ and a constant, uniform excitatory potential $a\lambda$. This can be interpreted as a uniform lowering of the thresholds.

Clearly, the properties of the network should not be too sensitive to the particular value of the constraint parameter λ. For very large values of λ, it is showed in [5] that the results become essentially independent of λ and that one returns to the rigid constraint with reduced numbers of spurious sites, high α_c and excellent retrieval quality.

Non-local learning rules: The pseudo-inverse learning rule

Hopfield model has been constructed using Hebb's learning rule, whose main characteristic is that of being a *local law.* Each T_{ij} depends only on the values of ξ_i^μ and ξ_j^μ. Therefore, the synapses are modified presumably by past activities of the neurons they connect, represented by the quenched values ξ_i^μ and ξ_j^μ. Locality of the learning rule is a reasonable condition for biological networks, but it is not unreasonable that, even in biological systems, some long-term synaptic modifications which depend on the history of activity of a large group of neurons might occur in addition to the main local modifications.
As seen in section 2.2.4, this type of learning rule has some limitations. Indeed while using this rule it may either be difficult or not possible to deal with correlated candidate state vectors. For this reason, we introduce *non-local learning rules*, that can suppress the adverse

effects of the overlaps among the candidate state vectors. An example is given in (2.110), where we have taken into account the mean activity rate of the entire network. An alternative is represented by the *pseudo-inverse learning rule.*

In section 2.2.3 we have introduced Hebbian learning rule and seen how it works well when the stored patterns are orthogonal or nearly orthogonal. A more complex learning rule can be chosen to ensure that T has a single degenerate eigenvalue regardless of whether the memory vectors are orthogonal. We can thus use in Hopfield *pseudo inverse learning rules* to encode the pattern information if pattern vectors are even non orthogonal. Such "spectral schemes" have been explored by Personnaz, Guyon and Dreyfus in [47] and then by Venkatesh and Psaltis in [64][38]. In these works they have shown constructive schemes for the generation of the weight matrix which yield a larger capacity than the outer-product scheme and the memorized vectors ξ^μ are true eigenvectors for the linear operator T and not only "pseudo-eigenvectors". Furthermore, they have constructed a nonlocal model in which any set of patterns, correlated or not, can be memorized without errors as long as they are linearly independent.

Firstly, consider the *Moore-Penrose generalized inverse* of $\Xi = [\xi^1, \dots, \xi^p]$, the $N \times p$ matrix containing all the stored memory,

$$\Xi^\dagger = \begin{bmatrix} \xi_1^\dagger \\ \vdots \\ \xi_p^\dagger \end{bmatrix} = (\Xi^T\Xi)^{-1}\Xi^T \tag{2.126}$$

This matrix has the property that is orthogonal to the starting matrix Ξ

$$\Xi^\dagger\Xi = \mathbb{I} \qquad \text{or} \qquad \xi_\mu^\dagger\xi^\beta = \delta_{\mu\beta} \tag{2.127}$$

where $\mathbb{I}$ is the $p \times p$ identity matrix and $\delta_{\mu\beta}$ is the Kronecker delta. Hence, the pseudo inverse weight matrix is given by

$$\tilde{T} = \Xi\Xi^\dagger = \sum_{\mu=1}^{p} \xi^\mu\xi_\mu^\dagger \tag{2.128}$$

This matrix (2.128) is always symmetric: $\tilde{T}^T = \tilde{T}$ and satisfies the following property:

$$\tilde{T}\Xi = \Xi\Xi^\dagger\Xi = \Xi \quad \Longrightarrow \quad \tilde{T}\xi^\mu = \xi^\mu \tag{2.129}$$

In other words, each memory pattern is an eigenvector of the weight matrix $\tilde{T}$ with eigenvalue 1 and this is true as long as the ξ^μ's are linearly independent (as we have supposed in Hopfield model).

The pseudo-inverse learning rule can then be reformulated using the *correlation matrix* between stored patterns, as described by Kanter and Sompolinsky in [31]:

$$C_{\mu\nu} \equiv \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu\xi_i^\nu, \qquad \mu,\nu = 1,2,\dots,p \tag{2.130}$$

The model consists of a set of synapses

$$\tilde{T}_{ij} = \frac{1}{N}\sum_{\mu,\nu=1}^{p} \xi_i^\mu\xi_j^\nu(C^{-1})_{\mu\nu} \tag{2.131}$$

[38] Venkatesh and Psaltis have started from the article and theory of Personnaz et al. and have developed a theory using this pseudo-inverse learning rule.

where C^{-1} is the inverse of the matrix defined in Eq. (2.130). Also in this case, the network stability condition holds for any set of patterns, provided they are linearly independent.
In both models, the self-coupling term $\tilde{T}_{ii}V_i$ is not neglected, when considering the local field $\tilde{h}_i$ of neuron i

$$\tilde{h}_i = \sum_{j=1}^{N} \tilde{T}_{ij} V_j \tag{2.132}$$

because this term is needed for the validity of the network stability relation. The presence of the self-coupling restricts severely the size of the basins of attraction of the patterns especially for large α. In [31], it is proven, through numerical simulations, that although the memories are stable up to $\alpha = 1$ the maximum load parameter of the system for providing associative memory is

$$\alpha_c = \frac{1}{2} \tag{2.133}$$

A modification of the above model is proposed in [31] by eliminating the self-coupling term in the expression of the local field of neuron i

$$\tilde{h}_i = \sum_{j, j \neq i} \tilde{T}_{ij} V_j \tag{2.134}$$

In this case, they have found that for synchronous as well as asynchronous update the radius of attraction R is unity at $\alpha \to 0$, then it decreases monotonically and vanishes at $\alpha_c = 1$, as can be understood from relation (2.66). Then, there is an increase in the number of spurious stable states as α increases, but the occurrence of linear combinations of stable state vectors as spurious states is very rare. Thus the presence of spurious states does not affect the basins of attraction of the stored stable states.
Hence, we have presented a model capable of storing both correlated and uncorrelated state vectors, overcoming the assumption of uncorrelated patterns of Hopfield. Unfortunately, this is achieved at the expense of a huge increase in computational complexity, a problem which plagues all the alternatives to the outer product rule.

2.4.3 Inferring Learning rule from *In Vivo* data

In the previous discussion, we have proposed the Hopfield model as an attractor network capable of learning and retrieving memories. In this model, synaptic connectivity is set up in such a way that the network dynamics have multiple attractor states, each of which represents a particular item that is stored in memory. In the original model, the appropriate synaptic connectivity is assumed to be generated through a Hebbian learning process according to which synaptic efficacies are modified by the activity of pre- and post-synaptic neurons. The attractor network scenario was originally instantiated into highly simplified, fully connected networks of binary neurons (Hopfield, 1982 [28]). Although theorists have since strived to incorporate more neurophysiological realism into associative memory models, using, e.g., asymmetric and sparse connectivity (Derrida et al. [16]), sparse coding of memories and correlated patterns (Kanter and Sompolinsky [31], Tsodyks and Feigel'Man [63]), spiking neuron (Gerstner and Van Hemmen [20]), there is still a large gap between the models and experimental data. First, most models use bimodal distributions of firing rates with neurons either activated by a stimulus or not, whereas there is no indication of such a bimodality in the data. Second, the connectivity matrices used in these models are essentially engineered (and sometimes highly fine-tuned) to produce attractor dynamics but are totally unconstrained by data.

In a recent article of Brunel and Pereira (July 2018, [46]), they have compared the distribution of responses to novel and familiar stimuli to allow the inference of the dependence of the learning rule on the post-synaptic firing rates. The inferred learning rule is Hebbian, but with some differences from the classical one: the post-synaptic dependence of the rule is dominated by depression so that the vast majority of external inputs leads to a net decrease in total synaptic inputs to a neuron with learning, leading to a sparser representation of external stimuli, and the dependence of the rule on post-synaptic firing rates is highly non-linear.
Consider a model of N neurons, whose firing rates[39] are described by variables r_i. Firing rates obey a standard rate equation (a slight change of Eq. (2.69))

$$\tau \dot{r}_i = -r_i + \phi\Big(I_i + \sum_{j \neq i} T_{ij} r_j\Big) \tag{2.135}$$

where τ is the time constant of firing rate dynamics, ϕ is the input-output single-neuron transfer function. The connectivity matrix is sparse and existing connections are shaped by external inputs ("patterns") through a non-linear unsupervised Hebbian synaptic plasticity rule.
In this context, external synaptic input ξ_i^μ to neuron i are supposed to be generated randomly and independently from a Gaussian distribution. The assumption of independence is reasonable, because it is consistent with the data. The external inputs shape the connectivity matrix through

- the firing rates $\phi(\xi_i^\mu)$ generated by such inputs
- through two non-linear functions f_1 and f_2, that characterize the dependence of the learning rule on the post-synaptic rate (f_1) and pre-synaptic rate (f_2), respectively.

When p patterns are learned by the network, the final connectivity after learning is structured as

$$J_{ij} = \frac{Ac_{ij}}{cN} \sum_{\beta=1}^{p} f_1[\phi(\xi_i^\beta)] f_2[\phi(\xi_j^\beta)] \tag{2.136}$$

where $c_{ij} \ll 1$ is a sparse random connectivity matrix with a distribution similar to that seen previously in Eq. (2.90)

$$Pr(c_{ij}) = c\delta(c_{ij} - 1) + (1 - c)\delta(c_{ij}) \tag{2.137}$$

This synaptic connectivity matrix can be obtained by a learning rule that changes the synaptic connectivity matrix by a factor $\Delta J_{ij} \propto f_1[\phi(\xi_i^\beta)] f_2[\phi(\xi_j^\beta)]$ when a pattern ξ^μ is presented to the network, starting from an initial tabula rasa $J_{ij} = 0$ and neglecting the contributions of recurrent connections during learning.
This rule represents a generalization of the classical Hebbian rule with two important differences:

- patterns have a Gaussian distribution instead of binary
- the dependence of the rule on firing rates is non-linear instead of linear.

The model defined by Equations (2.135) and (2.136) depends on three functions: ϕ, f_1 and f_2, that define the single-neuron transfer function and synaptic learning rule, respectively. In [46], it is used a method to infer the transfer function (ϕ) and the postsynaptic dependence

[39] Firing rates are defined as the averaged number of spikes in a unit interval time.

of the learning rule f_1 from electrophysiological data recorded in ITC (*inferior temporal cortex*[40]), whereas f_2 cannot be inferred. As an additional step to previous studies, sigmoidal functions for f_1 and ϕ are used, because they provided good fits to the data.

A question that can be addressed is whether the model exhibits attractor dynamics or not. Consider two order parameters, that describe how network states are correlated (or not) with stored patterns:

- the overlap m^μ between the current state of the network and the learned patterns of interest, defined as in Eq. (2.10)
- the interference I due to the other stored patterns in the connectivity matrix, that is proportional to the averaged squared firing rates of the network

$$I = \frac{1}{N}\sum_{i=1}^{N} r_i^2 \tag{2.138}$$

Through numerical simulations of large networks and mean field theory, in [46] it is proved that this type of network behaves as an associative memory when constrained by ITC data, without any need for parameter variation or fine tuning. The learning rule inferred from this data is then close to maximizing the number of stored patterns ($\alpha_c \approx 0.85$) in a space of unsupervised Hebbian learning rules with sigmoidal dependence on pre- and post-synaptic firing rates.

2.5 A probabilistic approach

Hopfield developed a model that makes it possible to explore the global natures of large neural networks without losing the information of essential biological functions. For symmetric neural circuits, an energy landscape can be constructed and thus a memory retrieval process from a cue (incomplete initial information) to the corresponding memory (complete information) is obtained. However, in real neural networks, the connections among neurons are mostly asymmetric rather than symmetric, as we have discussed previously in section 2.4.1. Under this more realistic biological situation, the original Hopfield model fails to apply, because there is no easy way of finding out the underlying energy function and studying the global stability. In order to overcome this problem, following the contributions of Yan et al. in [67] we will study the global behavior of neural circuits with synaptic connections from symmetric to general asymmetric ones. Thus, we will develop a potential and flux landscape theory for neural networks based on the statistical probabilistic description of non-equilibrium dynamical systems.

In general, when exploring the global dynamics of a neural network, there are several approaches: one is to follow individual deterministic trajectories and the other one is to describe the system from thermodynamics perspectives with energy, entropy and free energy for a global characterization of the system. Hopfield in [28] proposed a model that quantifies how individual neurons evolve with time based on their biological features, i.e. Eq. (2.69), that can be improved as explained in section 2.4. However, neural networks are often under fluctuations from intrinsic source and external environments. Therefore, rather than individual trajectory evolution, the probabilistic evolution can characterize the dynamics globally and often more appropriate.

[40] The IT cortex (ITC) in humans, also known as the Inferior Temporal Gyrus, is the anterior region of the temporal lobe located underneath the central temporal sulcus, that processes visual stimuli of objects in our field of vision and is involved in memory and memory recall to identify that object.

2.5.1 Symmetric case

We start by considering Hopfield continuous model (2.69) with the assumptions $C_j = 1$, $R_j = 1$ for all $j = 1, \dots, N$.

$$\frac{du_j}{dt} = -u_j + \sum_i T_{ji} V_i + I_j, \qquad \text{with } V_j = g(u_j) \tag{2.139}$$

Then, under these hypothesis the energy function (2.70) becomes

$$E = -\frac{1}{2} \sum_i \sum_j T_{ij} V_i V_j + \sum_i \int_0^{V_i} g_i^{-1}(V) dV - \sum_i I_i V_i \tag{2.140}$$

It is easy to observe that E can be written as the gradient of the r.h.s of Eq. (3.102):

$$\begin{aligned} \nabla_k E = \frac{\partial E}{\partial u_k} = \frac{\partial E}{\partial V_k} \frac{\partial V_k}{\partial u_k} &= g'(u_k) \Big(-\sum_j T_{kj} V_j + g_k^{-1}(V_k) - I_k \Big) \\ &= g'(u_k) \Big(-\sum_j T_{kj} g_j(u_j) + u_k - I_k \Big) \end{aligned} \tag{2.141}$$

Note that $g'(u_k)$ is a constant since g is a sigmoidal function and for this type of function the following property holds:

$$g'(u) = \frac{dg(u)}{du} = g(u)(1 - g(u))$$

Thus in our case $g'(u_k) = V_k(1 - V_k) \in \mathbb{R}$ and using Eq. (2.141) in (3.102) we gain

$$\frac{du_j}{dt} = -\frac{1}{g'(u_j)} \nabla_j E \tag{2.142}$$

with E that represents the potential energy.
We then obtain a set of *Langevin equations* considering the stochastic dynamics of the neural network under fluctuations:

$$\frac{du_j}{dt} = -\frac{1}{g'(u_j)} \nabla_j E + \zeta_j \tag{2.143}$$

where ζ represents the stochastic fluctuation force with an assumed Gaussian distribution. The autocorrelations of the fluctuation are assumed to be

$$\langle \zeta_j(u,t), \zeta_i(u,\tilde{t}) \rangle = 2 D_{ji}(u) \delta_{ji} \delta(t - \tilde{t}) \tag{2.144}$$

where δ is the δ-function and the diffusion matrix $D(u)$ is defined as $D(u) = \mathcal{D} G(u)$, where $\mathcal{D}$ is a scale constant giving the magnitude of the fluctuations and $G(u)$ represents the scaled diffusion matrix, that is symmetric and positive definite. Then, we can define the vector field X as $X(u) = -\nabla E(u)$, which represents the driving force for the dynamics of the underlying neural network.
Starting from the Langevin equation (2.143), we want now to derive a Fokker-Planck equation following the approach presented in [18], [52] and [54]. We recall that in general we can write a Fokker-Planck equation for the multivariate case as

$$\partial_t P(\tilde{u}, t | u_0, t_0) = (-1) \sum_j \frac{\partial}{\partial \tilde{u}_j} [M_j^{(1)}(\tilde{u}, t) P(\tilde{u}, t | u_0, t_0)] + \sum_{j,i} \frac{\partial^2}{\partial \tilde{u}_j \partial \tilde{u}_i} [M_{ji}^{(2)}(\tilde{u}, t) P(\tilde{u}, t | u_0, t_0)] \tag{2.145}$$

by considering a *Kramers-Moyal expansion*[41] of the probability density. We recall that $M^{(1)}(u,t)$ is a vector and $M^{(2)}(u,t)$ a semidefinite positive and symmetric matrix.
In order to establish the corresponding Fokker-Planck equation we have to determine the drift term $M^{(1)}$ and the diffusion term $M^{(2)}$. This can be done by computing the moments of the variable $\tilde{u} = (\tilde{u}_1, \dots, \tilde{u}_N)$ starting from u_0 at time t_0 directly from the Langevin equation. Those moments will be then identified with the definitions

$$\begin{aligned} M^{(m)}_{j_1,\dots,j_m}(\tilde{u},t) &\equiv \frac{1}{m!}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{\mathbb{R}^N} du(u_{j_1}-\tilde{u}_{j_1})\cdots(u_{j_m}-\tilde{u}_{j_m})P(u,t+\Delta t|\tilde{u},t) \\ &= \frac{1}{m!}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\Big\{\prod_{\eta=1}^{m}(u_{j_\eta}(t+\Delta t)-u_{j_\eta}(t))|u_{j_1}(t)=\tilde{u}_{j_1},\dots,u_{j_m}(t)=\tilde{u}_{j_m}\Big\} \end{aligned} \tag{2.146}$$

Let us consider the finite difference version of Eq. (2.143)

$$\Delta u_{j_\eta}(t) \equiv u_{j_\eta}(t+\Delta t) - u_{j_\eta}(t) = \frac{1}{g'(u_{j_\eta})}X_{j_\eta}(u)\Delta t + (2D_{j_\eta,j_\mu})^{1/2}\Delta W(t) \tag{2.147}$$

where $\Delta W(t) = W(t+\Delta t) - W(t)$ are increments of the related Wiener process. We then define $F_{j_\eta}(u) = \frac{1}{g'(u_{j_\eta})}X_{j_\eta}(u)$ to simplify the notation.
By using the properties $\mathbb{E}\{\Delta W(t)\} = 0$, $\mathbb{E}\{(\Delta W(t))^2\} = \Delta t$ and the independence of the increments $\Delta W(t)$ one obtains:

$$\begin{aligned} \mathbb{E}\{F_{j_\eta}(u)\Delta t|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} &= \mathbb{E}\{F_{j_\eta}(u)|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\}\Delta t && (2.148) \\ &= F_{j_\eta}(\tilde{u})\Delta t && (2.149) \\ \mathbb{E}\{(2D_{j_\eta j_\mu})^{\frac{1}{2}}\Delta W(t)|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} &= (2D_{j_\eta j_\mu})^{\frac{1}{2}}\mathbb{E}\{\Delta W(t)|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} && (2.150) \\ &= (2D_{j_\eta j_\mu})^{\frac{1}{2}}\mathbb{E}\{\Delta W(t)\} = 0 && (2.151) \\ \mathbb{E}\{2D_{j_\eta j_\mu}(\Delta W(t))^2|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} &= 2D_{j_\eta j_\mu}\mathbb{E}\{(\Delta W(t))^2|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} && (2.152) \\ &= 2D_{j_\eta j_\mu}\mathbb{E}\{(\Delta W(t))^2\} = 2D_{j_\eta j_\mu}\Delta t && (2.153) \end{aligned}$$

By using the above relations one gets

$$\begin{aligned} \mathbb{E}\{\Delta u_{j_\eta}(t)|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} &= \mathbb{E}\{F_{j_\eta}(u)\Delta t + (2D_{j_\eta,j_\mu})^{\frac{1}{2}}\Delta W(t)|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} \\ &= F_{j_\eta}(\tilde{u})\Delta t \end{aligned} \tag{2.154}$$

$$\begin{aligned} &\mathbb{E}\{(\Delta u_{j_\eta}(t))^2|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} = \\ &= \mathbb{E}\{(F_{j_\eta}(u))^2(\Delta t)^2 + F_{j_\eta}(u)(2D_{j_\eta,j_\mu})^{\frac{1}{2}}\Delta t\Delta W(t) + \\ &\qquad\qquad + 2D_{j_\eta,j_\mu}(\Delta W(t))^2|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t)\} = \\ &= 2D_{j_\eta,j_\mu}\Delta t + O((\Delta t)^2) \end{aligned} \tag{2.155}$$

[41] The Kramers-Moyal expansion refers to a Taylor series expansion of the Master equation for the probability density, that can thus be transformed into a partial differential equation:

$$\partial_t P(\tilde{u},t|u_0,t_0) = \sum_{m=1}^{N}(-1)^m\sum_{i_1\cdots i_m}\frac{\partial^m}{\partial\tilde{u}_{i_1}\cdots\partial\tilde{u}_{i_m}}[M^{(m)}_{i_1,\dots,i_m}(\tilde{u},t)P(\tilde{u},t|u_0,t_0)]$$

where $M^{(m)}$ is the m-th coefficient of the expansion. If we consider only the first two terms $M^{(1)}(u,t)$ and $M^{(2)}(u,t)$ we get the Fokker-Planck equation.

and

$$\begin{aligned}
&\mathbb{E}\{\Delta u_{j_\eta}(t)\Delta u_{j_\mu}(t)|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t),u_{j_\mu}(t)=\tilde{u}_{j_\mu}(t)\}=\\
=\;&\mathbb{E}\{F_{j_\eta}(u)F_{j_\mu}(u)(\Delta t)^2+2F_{j_\eta}(u)(2D_{j_\eta,j_\mu})^{\frac{1}{2}}\Delta t\Delta W(t)+F_{j_\mu}(u)(2D_{j_\mu,j_\eta})^{\frac{1}{2}}\Delta t\Delta W(t)+\\
&\qquad+2D_{j_\eta,j_\mu}(\Delta W(t))^2|u_{j_\eta}(t)=\tilde{u}_{j_\eta}(t),u_{j_\mu}(t)=\tilde{u}_{j_\mu}(t)\}=\\
=\;&2D_{j_\eta,j_\mu}\Delta t+O((\Delta t)^2)
\end{aligned} \tag{2.156}$$

where we use the symmetry of the matrix D, i.e. $D_{j_\eta,j_\mu}=D_{j_\mu,j_\eta}$. Hence

$$\begin{aligned}
M_j^{(1)}(\tilde{u},t) &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int du(u_j-\tilde{u}_j)P(u,t+\Delta t|\tilde{u},t)\\
&= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\Big\{\Delta u_j|u_j(t)=\tilde{u}_j\Big\}=F_{j_\eta}(\tilde{u})
\end{aligned} \tag{2.157}$$

and

$$\begin{aligned}
M_{ji}^{(2)}(\tilde{u},t) &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\frac{1}{2}\int du(u_j-\tilde{u}_j)(u_i-\tilde{u}_i)P(u,t+\Delta t|\tilde{u},t)\\
&= \frac{1}{2}\lim_{\Delta t\to 0}\frac{1}{\Delta t}\mathbb{E}\Big\{\Delta u_j\Delta u_i|u_j(t)=\tilde{u}_j,u_i(t)=\tilde{u}_i\Big\}=\frac{2D_{ji}}{2}=D_{ji}
\end{aligned} \tag{2.158}$$

Therefore, the corresponding Fokker-Planck equation becomes

$$\partial_t P(u,t|u_0,t_0)=-\sum_j\frac{\partial}{\partial u_j}[F_j(u)P(u,t|u_0,t_0)]+\sum_{j,i}\frac{\partial^2}{\partial u_j\partial u_i}[D_{ji}(u,t)P(u,t|u_0,t_0)] \tag{2.159}$$

Using the definition of $F_j(u)$ and $D_{ji}(u)$ given above we obtain

$$\partial_t P(u,t|u_0,t_0)=-\sum_j\frac{\partial}{\partial u_j}[\frac{1}{g'(u_j)}X_j(u)P(u,t|u_0,t_0)]+\mathcal{D}\sum_{j,i}\frac{\partial^2}{\partial u_j\partial u_i}[G_{ji}(u,t)P(u,t|u_0,t_0)] \tag{2.160}$$

where we recall that $X_j(u)=-\nabla_j E(u)$.

Now since we can find an energy function E and the driving force X is the gradient of that energy function, we know that for the Fokker-Planck equation (2.160) exists an equilibrium stationary solution

$$P_{ss}(u)=e^{-\frac{E(u)}{\mathcal{D}}} \tag{2.161}$$

that represents the long time limit solution and it is easy to understand that its point of maximum u^* corresponds to a point of minimum of E, i.e. a stored pattern.

2.5.2 General asymmetric case

For the Hopfield model with symmetric synapses we can always find an energy function E and use it to study the global stability of the system. We have also seen that we can derive a Fokker-Planck equation with an associated equilibrium solution P_{ss}. However, in reality, the connections of neurons in neural networks are often asymmetric. Hopfield model with symmetric connections does not apply in this regime, because there is no easy way of finding out the energy function determining the global nature of the neural networks in such general case.

Realistic neural networks are open systems with constant exchanges of energy and information with the environment. Thus they are not conserved systems. The driving force therefore cannot often be written as a gradient of an energy function in such non-equilibrium conditions.

However, finding an energy-like function such as a Lyapunov function (monotonically going down as the dynamics) is essential for quantifying the global stability of the system. Following [67] and [69], we want to prove whether such a potential function exists and can be used to study the stability.
As done for the symmetric case in section 2.5.1, starting from the continuous version of the Hopfield equation for the dynamics with an asymmetric interconnection matrix J, defined for example as in Eq. (2.89) or in Eq. (2.136),

$$\frac{du_j}{dt} = -u_j + \sum_i J_{ji} V_i + I_j, \qquad \text{with } V_j = g(u_j) \tag{2.162}$$

and adding a stochastic fluctuation force ζ, we obtain the following Langevin equation

$$\frac{du_j}{dt} = F_j(u) + \zeta_j \tag{2.163}$$

where ζ_j is defined as in the symmetric case and satisfies the property given in Eq. (2.144), $F(u)$ is the driving force for the dynamics of the underlying neural networks and in our case

$$F_j(u) = -u_j + \sum_i J_{ji} V_i + I_j \qquad \text{for } j = 1, \dots, N \tag{2.164}$$

Then also in this case we can establish the corresponding Fokker-Planck diffusion equation for probability evolution of the state variable u_j by following the same reasoning of the symmetric case, where we never use the hypothesis of symmetry of the interconnection matrix. Thus we obtain

$$\frac{\partial P(u,t)}{\partial t} = -\sum_{j=1}^{N} \frac{\partial}{\partial u_j}[F_j(u)P(u,t)] + \mathcal{D}\sum_{j=1}^{N}\sum_{i=1}^{N} \frac{\partial^2}{\partial u_j \partial u_i}[G_{ji}(u,t)P(u,t)] \tag{2.165}$$

The Fokker-Planck equation (2.165) can be written in the form of the probability conservation:

$$\frac{\partial P}{\partial t} + \nabla \cdot \Phi = 0 \tag{2.166}$$

where Φ is the probability flux: $\Phi = FP - \nabla \cdot (DP)$. The driving force of the neural network systems F can be decomposed as follows:

$$F = \frac{\Phi_{ss}}{P_{ss}} + D \cdot \frac{\nabla(P_{ss})}{P_{ss}} + \nabla \cdot D = \frac{\Phi_{ss}}{P_{ss}} - D \cdot \nabla \mathcal{U} + \nabla \cdot D \tag{2.167}$$

where the non-equilibrium potential $\mathcal{U}$ here is naturally defined as

$$\mathcal{U}(u) = -\ln(P_{ss}(u)) \tag{2.168}$$

and is related to the steady-state probability distribution[42]. Then we note that for steady state the divergent condition of the flux $\nabla \cdot \Phi_{ss} = 0$ is satisfied. Thus, F can be decomposed

[42] To obtain the potential landscape characterized by $\mathcal{U}$ we have to calculate the probability distribution of steady state P_{ss}. The corresponding Fokker-Planck equation describes the evolution of the system that can give exact solution of steady-state probability. However, it is hard to solve this equation directly due to the huge dimensions of the network. Therefore, in [67] a self-consistent mean field approximation is used to reduce the dimensionality. This method can effectively reduce the dimensionality from exponential to polynomial by approximating the whole probability as the product of the individual probability for each variable and be carried out in a self-consistent way (treating the effect of other variables as a mean field). The interested reader can refer to [67].

into a gradient of a non-equilibrium potential and a divergent free flux, modulo to the inhomogeneity of the diffusion, which can be absorbed and redefined in the total driving force. The divergent free force Φ_{ss} has no source or sink to go to or come out, therefore it has to rotate around and become curl.
When the flux is divergent free $\nabla \cdot \Phi = 0$ in steady state, it does not necessarily mean $\Phi_{ss} = 0$. There are indeed two possibilities:

1. $\Phi_{ss} = 0$: the detailed balance[43] is satisfied and the system is in equilibrium. Furthermore, the dynamics is determined by purely the gradient of the potential. This is exactly the equilibrium case for the Hopfield model for neural networks of learning and memory assuming underlying symmetrical connections between neurons.

2. $\Phi_{ss} \neq 0$: the detailed balance is broken and we are in the non-equilibrium state. The steady-state probability distribution can quantify the global nature of the neural network, whereas the local dynamics is determined by both the gradient of the non-equilibrium potential landscape and the nonzero curl flux. This curl flux can lead to spiral motion with the underlying network dynamics deviating from the gradient path and even generate coherent oscillations[44], which are not possible under pure gradient force.

To quantitatively study the global stability of the neural network, we need to find out whether a Lyapunov function, that monotonically goes down in time, exists and how it is related to the potential landscape constructed. Following [67], expand $\mathcal{U}(u)$ with respect to the parameter $\mathcal{D}$ for the case of weak fluctuations, $\mathcal{D} \ll 1$ in realistic neural networks

$$\mathcal{U} = \frac{1}{\mathcal{D}}\psi = \frac{1}{\mathcal{D}}\sum_{j=0}^{\infty}\mathcal{D}^i\psi_i$$

and replace it with the steady-state Fokker-Planck diffusion equation. Then we obtain the $\mathcal{D}^{-1}$ order part and this leads to the Hamilton-Jacobi equation for ψ_0, [67]:

$$\sum_{j=1}^{N} F_j(u)\frac{\partial\psi_0(u)}{\partial u_j} + \sum_{j=1}^{N}\sum_{i=1}^{N} G_{ji}(u)\frac{\partial\psi_0(u)}{\partial u_j}\frac{\partial\psi_0(u)}{\partial u_i} = 0 \tag{2.169}$$

that can also be rewritten in a more compact form as

$$F \cdot \nabla\psi_0 + \nabla\psi_0 \cdot G \cdot \nabla\psi_0 = 0 \tag{2.170}$$

Proposition 2.5.1 ([67]). *ψ_0 is a Lyapunov function in the sense of definition 2.2.8.*

Proof. ψ_0 satisfies the Lyapunov properties given in definition 2.2.8:

1. In the weak fluctuation assumptions $\psi_0 = \mathcal{U}\mathcal{D} = -\mathcal{D}\ln(P_{ss}(u))$, thus it a continuous function in u.

2. ψ_0 has a lower bound:

$$\psi_0 = -\mathcal{D}\ln(P_{ss}(u)) \geq -\mathcal{D}|\ln(P_{ss}(u))| \geq -\mathcal{D}$$

   where $\mathcal{D}$ is a constant.

[43] The principle of detailed balance states that at equilibrium each elementary process should be equilibrated by its reverse process.

[44] A coherent state is a state which has a dynamics most closely resembling the oscillatory behavior of classical harmonic oscillator.

3. Consider the time evolution of ψ_0.

$$\frac{d\psi_0(u)}{dt} = \nabla\psi_0 \cdot \dot{u} = \nabla\psi_0 \cdot F = -\nabla\psi_0 \cdot G \cdot \nabla\psi_0 \leq 0$$

   Since this derivation is based on the weak fluctuation assumptions ($\mathcal{D} \ll 1$), the temporal evolution $\dot{u}$ can be written as $F(u)$ here and therefore is deterministic without the noise term ζ. Then, the diffusion matrix G is positive definite and thus ψ_0 decreases monotonously with time.

□

From that property, we gain that the dynamical process will not stop until the system reaches a minimum that satisfies $\nabla\psi_0 = 0$. In this case there are two possible situations: we are in a point attractor or in a limit cycle.

- For point attractors, ψ_0 will settle down at the minimum value as happened for the classical Hopfield model.
- For limit cycles, the values of ψ_0 on the attractors must be constant. Therefore, the limit cycle cannot emerge from Hopfield model for neural networks with a pure gradient dynamics because there is no driving force for the coherent oscillation, that is provided from the nonzero flux.

ψ_0, being a Lyapunov function, is closely associated with the non-equilibrium potential $\mathcal{U}$ under the small fluctuation limit, therefore it reflects the intrinsic properties of the steady state without the influences from the magnitude of the fluctuations. When the underlying fluctuations are not weak, we can no longer expand the non-equilibrium potential $\mathcal{U}$ with respect to the fluctuation amplitude parameter $\mathcal{D}$ as shown above and the noise term ζ cannot be neglected in the derivation. Thus, there is no guarantee that ψ_0 is a Lyapunov function for finite fluctuations.

In analogy with equilibrium thermodynamics, we can construct non-equilibrium thermodynamics and apply it to neural networks following Yan et al. [67].

Starting from ψ_0, define the steady-state probability P_{ss} as

$$P_{ss}(u) = \frac{1}{Z} e^{-\frac{\psi_0}{\mathcal{D}}} \tag{2.171}$$

where Z can be defined as the time-independent (steady-state) non-equilibrium partition function

$$Z = \int e^{-\frac{\psi_0}{\mathcal{D}}} du \tag{2.172}$$

and the diffusion scale $\mathcal{D}$ measures the strength of the fluctuations. We also note that Z is not dependent on time t.

As presented in [67], the entropy S, the energy E and the free-energy $\mathcal{F}$ of the non-equilibrium neural network can then be defined as follows:

$$S = -\int P(u,t) \ln P(u,t) du \tag{2.173}$$

$$E = \int \psi_0 P(u,t) du = -\mathcal{D} \int \ln[Z P_{ss}] P(u,t) du \tag{2.174}$$

$$\mathcal{F} = E - \mathcal{D}S = \mathcal{D}\Big(\int P \ln\Big(\frac{P}{P_{ss}}\Big) du - \ln Z\Big) \tag{2.175}$$

As an open system that constantly exchanges energy and materials with the environment, the entropy does not necessarily increase all of the time. Through some algebra it can be seen that the system entropy evolution in time is contributed by two terms[45]:

$$\dot{S} = -\frac{d}{dt}\int P(u,t)\ln P(u,t)du = -\int \frac{\partial P(u,t)}{\partial t}\ln P(u,t)du = \dot{S}_t - \dot{S}_e$$

where

- $\dot{S}_t = \int du(\Phi \cdot G^{-1} \cdot \Phi)/P^2$ is the entropy production rate, which is either positive or zero since G is positive definite.
- $\dot{S}_e = \int du(\Phi \cdot G^{-1} \cdot F')/P$ is the heat dissipation rate or entropy flow rate to the non-equilibrium neural network from the environment, that can either be positive or negative. $\tilde{F} = F - \nabla \cdot G$ is the effective force.

Hence, the system entropy change rate $\dot{S}$ is not necessarily positive. Therefore, rather than entropy maximization, it is better to consider the free-energy minimization as the global principle and optimal design criterion for neural networks, as stated in [67].

Proposition 2.5.2 ([67], [69])**.** *$\mathcal{F}$ is a Lyapunov function in the sense of definition 2.2.8.*

Proof. $\mathcal{F}$ satisfies the Lyapunov property of definition 2.2.8:

1. Consider the time derivative of the free-energy

$$\begin{aligned}\frac{d\mathcal{F}}{dt} &= \mathcal{D}\frac{d}{dt}\Big(\int P\ln\Big(\frac{P}{P_{ss}}\Big)du\Big) = \mathcal{D}\int\frac{dP}{dt}\ln\Big(\frac{P}{P_{ss}}\Big)du + \mathcal{D}\int\frac{dP}{dt}du = \\ &= -\mathcal{D}\int(\nabla\cdot\mathcal{J})\ln\Big(\frac{P}{P_{ss}}\Big)du + \mathcal{D}\frac{d}{dt}\int Pdu = \\ &= -\mathcal{D}\Big(\int\nabla\cdot\Big[\mathcal{J}\ln\Big(\frac{P}{P_{ss}}\Big)\Big]du - \int\mathcal{J}\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)du\Big)\end{aligned}$$

Here $\mathcal{J}$ is defined as $\mathcal{J} = \Phi|_{\mathcal{D}\to 0}$. Now using Gauss's theorem and the boundary condition $n \cdot \Phi = 0$, which is the result of the conservation of total probability, we obtain

$$\frac{d\mathcal{F}}{dt} = \mathcal{D}\Big(\int\mathcal{J}\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)du\Big)$$

From $\Phi = FP - \mathcal{D}\nabla\cdot(GP) = (F - \mathcal{D}\nabla\cdot G - \mathcal{D}G\cdot\nabla\ln P)P$ we can get

$$\frac{\mathcal{J}}{P} + \mathcal{D}G\cdot\nabla\ln P = \frac{\mathcal{J}_{ss}}{P_{ss}} + \mathcal{D}G\cdot\nabla\ln P_{ss}$$

Consequently, we gain

$$\mathcal{J} = P\Big[\frac{\mathcal{J}_{ss}}{P_{ss}} - \mathcal{D}G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]$$

[45] To see that this equality holds, just think of the definition of Φ and substitute it in this equality and then after some calculations it is easy to obtain the derivative of S.

So,

$$\begin{aligned}
\frac{d\mathcal{F}}{dt} &= \mathcal{D}\Big(\int P\Big[\frac{\mathcal{J}_{ss}}{P_{ss}} - \mathcal{D}G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)du\Big)\\
&= \mathcal{D}\Big(\int P\frac{\mathcal{J}_{ss}}{P_{ss}}\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)du\Big)-\mathcal{D}^2\Big(\int\Big[\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\cdot G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]Pdu\Big)\\
&= \mathcal{D}\Big(\int \mathcal{J}_{ss}\cdot\nabla\Big(\frac{P}{P_{ss}}\Big)du\Big)-\mathcal{D}^2\Big(\int\Big[\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\cdot G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]Pdu\Big)\\
&= \mathcal{D}\Big(\int \nabla\cdot\Big(\mathcal{J}_{ss}\frac{P}{P_{ss}}\Big)du - \int\frac{P}{P_{ss}}\nabla\cdot\mathcal{J}_{ss}du\Big)-\mathcal{D}^2\Big(\int\Big[\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\cdot G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]Pdu\Big)\\
&= \mathcal{D}\Big(\int \nabla\cdot\Big(\mathcal{J}_{ss}\frac{P}{P_{ss}}\Big)du - \int\frac{P}{P_{ss}}\frac{\partial P_{ss}}{\partial t}du\Big)-\mathcal{D}^2\Big(\int\Big[\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\cdot G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]Pdu\Big)\\
&= \mathcal{D}\Big(\int \nabla\cdot\Big(\mathcal{J}_{ss}\frac{P}{P_{ss}}\Big)du\Big)-\mathcal{D}^2\Big(\int\Big[\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\cdot G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\Big]Pdu\Big)
\end{aligned}$$

Using again Gauss's theorem and the boundary condition $n\cdot\Phi = 0$ we obtain

$$\frac{d\mathcal{F}}{dt} = -\mathcal{D}^2\int\nabla\ln\Big(\frac{P}{P_{ss}}\Big)\cdot G\cdot\nabla\ln\Big(\frac{P}{P_{ss}}\Big)Pdu \leq 0$$

where to conclude we have used the positive definiteness of the diffusion matrix G.

2. $\mathcal{F}$ decreases in time until reaching the minimum value at steady state, that is a lower bound and is given by

$$\mathcal{F} = \mathcal{D}\Big(\int P\ln\Big(\frac{P}{P_{ss}}\Big)du - \ln Z\Big) \geq -\mathcal{D}\ln Z$$

since we know from the equilibrium thermodynamics that $\hat{H}[P](t) = \int P\ln\Big(\frac{P}{P_{ss}}\Big)du$ is a positive definite function, equal to zero only when $P = P_{ss}$.

□

Therefore, the free-energy function $\mathcal{F}$ related to the stationary state probability P_{ss} is a Lyapunov function for stochastic neural network dynamics at finite fluctuations and is suitable to explore the global stability of non-equilibrium neural networks. $\mathcal{F}$ is thus the natural extension of the energy landscape description to situations with noise and asymmetry. As discussed in [67], the minimization of $\mathcal{F}$ leads to the steady state probability distribution, that corresponds to the attractor probability distribution of states.

CHAPTER 3 Kuramoto Model

The object of this chapter is to study a network of N neurons in terms of oscillators, taking into account the rhythmical activity of each element and also of the brain. In order to do so, we will derive and then describe the Kuramoto model following the contributions of Acebrón et al. [2], Mori and Kuramoto [41] and Strogatz [58].

3.1 The role of synchronization in nature

Time plays a key role for all living beings. In fact their activity is governed by cycles of different duration which determine their individual and social behavior. There are biological processes and specific actions which require precise timing. Some of these actions demand a level of expertise that can only be acquired after a long period of training, but others take place spontaneously. How do these actions occur? Possibly through *synchronization* of individual actions in a population.

Synchronization has attracted the interest of scientists for centuries, because it is observed in biological, chemical, physical and social systems. A paradigmatic example is the synchronous flashing of fireflies observed in some South Asian forests. At night these fireflies start to emit flashes in an uncoordinated way, but after a certain period of time, their pulsations spontaneously synchronize, so that the entire population emits light perfectly in unison. This surprising phenomenon has been the subject of several studies and only in the 60s of the twentieth century has found a first explanation: the rhythm of flashes is not set by a single leader firefly, but depends on the interaction between all components of the population. In fact, in some way, the frequency of each firefly corrects itself spontaneously in order to blink in unison with the others. Synchronism, of course, is not limited to the case of fireflies, but is present in many other areas of reality: in physics we find it in the arrays of lasers and Josephson junctions, in sociology in the applause of a vast audience, in medicine, for example, in heartbeat due to pacemaker cells and in the dynamics of neuronal networks[46].

The mathematical study of synchrony arises from the need/will of describing, and therefore understanding, through a rigorous theory, these phenomena of collective self-organization that are clearly evident in various scientific and engineering fields. Schematically, these phenomena consist of a huge system of different oscillators that spontaneously, at a certain time step in their evolutions, collectively converge at a common frequency, in spite of the different individual frequencies from which they are characterized in the initial configuration. Therefore, two or more oscillating systems are said to be *totally synchronized* if their trajectories, in phase space, converge asymptotically in time to the same trajectory with exponentially fast convergence. A less stringent definition of synchronization has been proposed by Pikovsky et al. in *"Synchronization, an universal concept in nonlinear sciences"* (2001, [49]): "an adjustment of rhythms of oscillating objects due to their weak interaction".

Research on synchronization phenomena focuses inevitably on ascertaining the main mechanisms responsible for collective synchronous behavior among members of a given population. Generally to obtain a correlated global activity, interacting oscillating elements are required: the "oscillating objects" are described as systems with oscillating motion caused by internal

[46] This last topic is at the center of our interest for this model and it will be discussed in section 3.6.

processes or external energy sources, able to oscillate autonomously (self-sustaining oscillators) and to interact with each other with an intensity according to the coupling strength that distinguishes them.
There are different ways to tackle this problem:

- Suppose that the rhythmical activity of each element is described in terms of a physical variable that evolves regularly in time. When such a variable reaches a certain threshold, the element emits a pulse (action potential for neurons), which is transmitted to the neighborhood. Later on, a resetting mechanism initializes the state of this element and then a new cycle starts. Essentially the behavior of each element is similar to that of an oscillator. Assuming that the rhythm has a certain period, it is convenient to introduce the concept of *phase*, a periodic measure of the elapsed time. The effect of the emitted pulse is to alter the current state of the neighbors by modifying their periods, lengthening or shortening them. This disturbance depends on the current state of the oscillator receiving the external impulse and it can also be studied in terms of a phase shift. The analysis of the collective behavior of the system can be carried out in this way under two conditions:
  - the phase shift induced by an impulse is independent of the number of impulses arriving within an interspike interval
  - the arrival of one impulse affects the period of the current time interval, but memory thereof is rapidly lost and the behavior in future intervals is not affected.
- There is another scenario in which synchronization effects have been studied extensively. Let us consider an ensemble of nonlinear oscillators moving in a globally attracting limit cycle of constant amplitude. These are *phase- or limit-cycle oscillators.* We now couple them weakly to ensure that no disturbance will take any of them away from the global limit cycle. Therefore only one degree of freedom is necessary to describe the dynamic evolution of the system: the *phase.*

Even at this simple level of description it is not easy to propose specific models and the description becomes much more complicated for a large number of coupled elements. Hence, we are forced to consider models that are mathematically tractable, with continuous time and specific nonlinear interactions between oscillators.
In the last decades numerous authors have been looking for a “solvable” model of this type for years. Winfree, for example, persistently sought a model with nonlinear interactions. He realized that synchronization can be understood as a threshold process: when the coupling among oscillators is strong enough, a macroscopic fraction of them synchronizes to a common frequency. The model he proposed was hard to solve in its full generality, although a solvable version has been recently found by Ariaratnam and Strogatz in 2001.
The most successful attempt to find such solvable model was due to Kuramoto (1975), who analyzed a model of phase oscillators running at arbitrary intrinsic frequencies and coupled through the sine of their phase differences. Kuramoto not only noted that, under appropriate hypotheses, the model exhibits a transition behavior in which a critical coupling constant K_c separates two types of regime: one in which the system is inconsistent ($K < K_c$) and the other in which there is a partial and progressive synchronization, but it was also able to determine the exact value of this critical constant. The Kuramoto model is simple enough to be mathematically tractable, yet sufficiently complex to be nontrivial. It is also rich enough to display a large variety of synchronization patterns and sufficiently flexible to be adapted to many different contexts.

3.2 The Kuramoto model

We now proceed with the derivation of the Kuramoto model following Mori and Kuramoto's book *"Dissipative Structures and Chaos"* [41].

3.2.1 Phase method

Consider the N-dimensional differential equation

$$\frac{dX}{dt} = F(X) \qquad F : \mathbb{R}^N \to \mathbb{R}^N \tag{3.1}$$

that describes the motion of a single *limit-cycle oscillator* with angular frequency ω. Let $t \mapsto X_0(t)$ be an asymptotically stable and T-periodic solution, i.e.

1. $\Re(Spect(F'(X_0(t)))) < 0$

2. $\frac{dX_0(t)}{dt} = F(X_0(t))$ and $X_0(t+T) = X_0(t), \quad \forall t$

Following the idea of Winfree in *"The Geometry of Biological Time"* [66], used also by Kuramoto, we first introduce the *phase* of a periodic event.
Every periodic solution $t \mapsto \hat{X}(t)$ is represented by a closed orbit $\mathcal{O}$ and each of these closed orbits is in bijective correspondence with a circumference. The crucial point of this argument lies in the correspondence

$$\{\text{closed orbit}\} \longleftrightarrow \mathbb{S}^1$$

where the phase is a point in $\mathbb{S}^1$, that is defined on the circumference and moved on the closed orbit $\mathcal{O}$ of our interest. It can be interpreted both as a spatial coordinate (it gives the direction in which the orbit is traveled) and temporal, on which we will focus.
Consider the unit circle as the reference circumference:

$$\begin{cases} x = cos(2\pi t) \\ y = sin(2\pi t) \end{cases}$$

with $x, y, t \in \mathbb{R}$, or using the complex notation

$$z = x + iy = e^{2\pi t}$$

Note that while t varies between 0 and 1, the point indicating the phase travels the circumference counterclockwise.
Since the phase describes a certain rhythm, it is a function of t: it is the interval of time elapsed since the event we have marked with $\phi = 0$ occurred for the last time. In other words:

Definition 3.2.1 (Phase). *The* phase *of a periodic motion is defined as*

$$\phi(z(t)) = f(t) = \frac{time\ elapsed\ since\ \phi = 0}{period\ T} \quad (mod\ 1) \tag{3.2}$$

where mod 1 *let* $\phi = 0$ *be equivalent to* $\phi = 1$.

During a non-perturbed motion of period T, the definition implies that ϕ grows constantly:

$$\dot{\phi} = \frac{df(t)}{dt} = \frac{\omega}{T}$$

For simplicity we consider a unitary period T, so that the above expression becomes:

$$\dot{\phi} = \omega$$

Defined in this way, the phase is a convenient measure of time in periodic phenomena.
Now go back to our initial problem. If the dynamical system (3.1) is equipped with a stable, periodic solution of period T, with cycle-limit $\mathcal{C}$, then the motion on $\mathcal{C}$ can be described by the phase $\phi(X)$. Let $\mathcal{C}$ be this closed orbit that corresponds to the periodic solution $X_0(t)$. Using what we have just seen, we can define the phase on $\mathcal{C}$ through:

$$\frac{d\phi(X_0(t))}{dt} = \omega, \qquad X_0 \in \mathcal{C}$$

In our case it is very restrictive that the phase is defined only for the points that are on $\mathcal{C}$. In fact, if we consider the perturbed system, it can happen that points belonging to $\mathcal{C}$ go away, due to the perturbation. We therefore see the need to understand how the phase is influenced by it and to extend its definition to points that are "close" to $\mathcal{C}$ (it is enough to focus attention on the vicinity of $\mathcal{C}$, because we have assumed that the perturbation is weak).

Statement. *The phase is defined also for every point in a tubular region $\mathcal{T}$ around the closed orbit $\mathcal{C}$.*

Proof. Imagine that the closed orbit $\mathcal{C}$ crosses a thin tube with a circular section. We want to define $\phi(X(t))$ for every X that belongs to the tubular region. Denote with $\mathcal{T}$ this N-dimensional tubular region, containing the neighborhoods close to $\mathcal{C}$. Moreover, suppose that the basin of attraction[47] of $\mathcal{C}$ contains $\mathcal{T}$. We take a point P such that: $P \in \mathcal{T}$ but $P \notin \mathcal{C}$, and a point $Q \in \mathcal{C}$. From definition, we can associate the phase ϕ_Q to Q.
Suppose that P and Q start moving following the non-perturbed equation (3.1). For $t \to \infty$, P ends on $\mathcal{C}$, thus ϕ_P is the phase related to P. If, on $\mathcal{C}$, P and Q are infinitely close, we have $\phi_P = \phi_Q$. Since P is an arbitrary point on $\mathcal{T}$, this proves that we can associate a phase to every point on $\mathcal{T}$. □

We can imagine that $\mathcal{T}$ is covered by a mono-parametric family of $(N-1)-$dimensional surfaces, where the phase is constant. These surfaces are called *isochronous* and are indicated with $I(\phi)$. It follows directly from the definition of ϕ that all the points on $\mathcal{T}$, which are in a certain instant in the same isochronous, will remain in the same isochronous until the system remains unperturbed. In particular, the variation of ϕ in a given point is the same as it belongs to $\mathcal{C}$ or $\mathcal{T}$. Therefore:

$$\frac{d\phi(X(t))}{dt} = \omega, \qquad X \in \mathcal{T} \tag{3.3}$$

Now we want to find an equation that describes the phase behavior. We have the obvious identity:

$$\frac{d\phi(X(t))}{dt} = grad_X \phi \cdot \frac{dX}{dt} \tag{3.4}$$

[47] We recall the definition of basin of attraction. An equilibrium point x^* is said to be *attractive* if there exists a neighborhood U of x^* such that for every orbit $x(t)$ that starts from an interior point of U we have

$$\lim_{t\to\infty} x(t) = x^*$$

The biggest U for which this happens is called *basin of attraction* of x^*.

Combining Eq. (3.3) and Eq. (3.4) with Eq. (3.1), we gain

$$grad_X\phi \cdot F(X) = \omega, \qquad X \in \mathcal{T} \tag{3.5}$$

Since we have extended the definition of phase to all the X near the orbit $\mathcal{C}$, we can deduce something about the perturbed motion of ϕ.

A perturbation on the initial equation leads to a deviation of ϕ from the linear growth $\dot{\phi} = \omega$, leading to *phase locking* or *phase diffusion* effects.

The phase method is extremely effective in describing a collection of weakly coupled oscillators, which, in the hypothesis of weak coupling strength, are geometrically confined in a N-dimensional $\mathbb{T}^N$ torus. Denote with $q(t)$ the weak external force, which is due to the overall effect of the interactions on the single oscillator. Thus, we perturb the vector field F with a small perturbation $\epsilon q(t)$, that in general depends on t and we obtain:

$$\frac{dX}{dt} = F(X) + \epsilon q(t) \tag{3.6}$$

Our aim is to study what happens at the periodic solution $t \mapsto X_0(t)$. The motion remains periodic even having introduced the perturbation, only the period deviates slightly from the original T. We want to see that this variation is of order ϵ.

If we substitute Eq. (3.6) in Eq. (3.4) and using Eq. (3.5), we have

$$\frac{d\phi(X(t))}{dt} = grad_X\phi \cdot [F(X) + \epsilon q(t)] = \omega + \epsilon grad_X\phi \cdot q(t) \tag{3.7}$$

The right-hand side of Eq. (3.7) depends on the location of the point X on the isochronous $I(\phi)$. We would like to find an equation that describes the motion of the phase in terms of ϕ itself. We therefore try to obtain an equation in a closed form in ϕ, without dependence on X. To do this we use the "perturbative idea". Even if we specify the value of the phase, for example $\phi = \bar{\phi}$, we cannot locate the point X, since this is on a surface where the phase is constantly $\bar{\phi}$. Let $X_0(\phi)$ be the intersection point between $I(\phi)$ and $\mathcal{C}$. We know that X is close to $X_0(\phi)$ and the distance $|X - X_0(\phi)|$ tends to zero when $\epsilon \to 0$. In this way, approximating to the minimum order the right member of (3.7) we can replace X with $X_0(\phi)$. Thus, we obtain a differential equation in the only variable ϕ, which describes its dynamics:

$$\frac{d\phi}{dt} = \omega + \epsilon W(\phi) \tag{3.8}$$

where

$$W(\phi) = Z(\phi) \cdot \Pi(\phi),$$

with

$$Z(\phi) = grad_X\phi\big|_{X=X_0(\phi)}, \qquad Z(\phi + T) = Z(\phi)$$

and

$$\Pi(\phi) = q(X_0(\phi))$$

$Z(\phi)$ is the sensitivity function (already introduced by Winfree) and indicates how the oscillators respond to external perturbations. Figure 3.1 shows the geometric interpretation of $Z(\phi)$. It is represented by a vector which is normal to the surface $I(\phi)$ in X_0, because the vector $grad_X\phi$ is perpendicular to the isochronous $I(\phi)$. Furthermore, its module is given by the density of surfaces I through $X_0(\phi)$.

Remark. $Z(\phi)$ and $\Pi(\phi)$ are T-periodic functions in ϕ; this means that the instantaneous frequency, expressed in the right-hand side of Eq. (3.8), is T-periodic in ϕ.

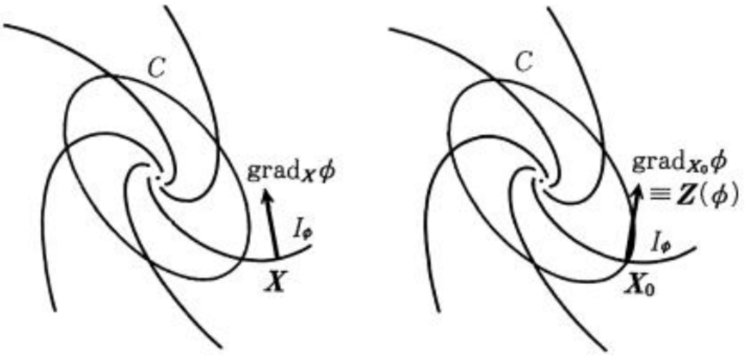


Figure 3.1: (a) Geometrical interpretation of the gradient vector $grad_X\phi$. (b) When the point-state X is close to the limit cycle, the gradient vector can be evaluated assuming that X belongs to the trajectory $\mathcal{C}$.

3.2.2 Derivation of the model

In the following, according to the theory proposed by Kuramoto, the dynamics in the trivial case of two identical and interacting oscillators are derived; from this follows the generalization to N elements. The interaction between two oscillators is represented by the perturbation $q = \tilde{q}(X, X')$ in which X and X' are respectively the vector-state of the oscillator i and j. Now, with the same reasoning seen above, we can replace X and X' in V with $X_0(\phi) = I(\phi) \cap \mathcal{C}$ and $X_0(\phi') = I(\phi') \cap \mathcal{C}$, obtaining the following equation that depends only on the phases:

$$\frac{d\phi}{dt} = \omega + G(\phi, \phi') \tag{3.9}$$

with G a T-periodic function in ϕ and ϕ' expressed by

$$G(\phi, \phi') = Z(\phi)\tilde{q}(X_0(\phi), X_0(\phi'))$$

To obtain the average frequency, we introduce the disturbances ψ and ψ' to the phases through $\phi = \omega t + \psi$ and $\phi' = \omega t + \psi'$. Equation (3.9) becomes:

$$\frac{d\psi}{dt} = G(\omega t + \psi, \omega t + \psi') \tag{3.10}$$

Note that since the coupling factor between the oscillators is weak, G is small and ψ and ψ' vary very slowly, so they change very hardly during a period T that they can be considered constant without loss of generality. Then, Eq. (3.10) can be mediated with respect to time and we gain

$$\frac{d\psi}{dt} = \Gamma(\psi, \psi') \tag{3.11}$$

with

$$\Gamma(\psi - \psi') = \frac{1}{T}\int_0^T G(\omega t + \psi, \omega t + \psi')d(\omega t) \tag{3.12}$$

that in terms of ϕ gives:

$$\frac{d\phi}{dt} = \omega + \Gamma(\phi - \phi') \tag{3.13}$$

with Γ a T-periodic function in $\phi - \phi'$.
It is easy to observe that the generalization to the case of N interacting oscillators leads to

N differential equations:

$$\frac{d\phi_i}{dt} = \omega_i + \sum_{\substack{j=1 \\ j\neq i}}^{N} \Gamma_{ij}(\phi_i - \phi_j) \qquad i = 1, 2, \ldots, N \tag{3.14}$$

where, as before, the function Γ_{ij} determines the coupling between the oscillators i and j. Equation (3.14) cannot be treated analytically: it becomes solvable if we take further and appropriate simplifications. In particular, the simplest case is represented by *Kuramoto model*, which needs the following additional hypothesis:

1. The function Γ_{ij} is the same for every couple of oscillators (i, j) and is of order N^{-1}, i.e. $\Gamma_{ij}(\phi_i - \phi_j) = N^{-1}\Gamma(\phi_i - \phi_j)$;

2. The function Γ is trigonometric: $\Gamma_{ij}(\phi_i - \phi_j) = -K \sin(\phi_i - \phi_j)$ with $K \geq\ 0$ *coupling strength*. We have to observe that the choice of this trigonometric function is crucial for the analytical solvability of the model.

3.2.3 Analysis of the Kuramoto model

As seen, the *mean-field* Kuramoto model consists of a collection of N oscillators lying on the unit circle $\mathbb{S}^1$ and with natural frequencies ω_i, whose phases ϕ_i are governed by the differential equations:

$$\dot{\phi}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N} \sin(\phi_j - \phi_i) \qquad i = 1, 2, \ldots, N \tag{3.15}$$

where K is the coupling constant and N the total number of oscillators. The presence of the multiplicative factor $\frac{1}{N}$ in the equation is necessary to maintain the model well-defined if we work with $N \to \infty$.
To simplify the theoretical passages, some assumptions are made:

- The natural frequencies ω_i are in general distributed according to a probability density $g(\omega)$ symmetric with respect to the average frequency $\omega_0 = \int_{-\infty}^{\infty} \omega g(\omega) d\omega$, so as to have $g(\omega_0 + \omega) = g(\omega_0 - \omega)$.

- In most cases it is required also that $g(\omega)$ is monotonically increasing in $(-\infty, \omega_0]$ and monotonically decreasing in $[\omega_0, \infty)$.

- We also assume that the probability density g is an unimodal function.

Under these hypotheses Kuramoto formulates the following conjecture: for $K < K_c$ the oscillators do not seem to be affected by the mutual interaction and therefore persist in their motion around the unit circle with a frequency close to the intrinsic one, generating a *de-synchronous* state asymptotically stable; exceeding the critical threshold K_c, however, the oscillators are divided into two groups. Those with natural frequency sufficiently far from the center of the function g continue to rotate with their natural frequency, while those with frequency near the center become *phase-locked* and rotate with frequency ω_0 and average phase θ.
The intrinsic difficulties related to the demonstration of this conjecture are essentially due to the impossibility of using standard analysis techniques, such as spectral theory and

asymptotic stability criteria, because of the presence of a linear operator with a continuous spectrum in the imaginary axis. Among the ideas emerged to overcome the problem there is that of Chiba, who was able to prove the linear stability of stationary solutions by first developing the spectral theory on a space of generalized functions with the help of Hilbert's theory and later admitting a spectral decomposition for the linear operator consisting of a set of countable cardinality autofunctions. For a theoretical study we refer to [13].
As regards the transition process that underlies the model, it is possible to determine a direct equation able to supply the exact value of the critical coupling constant K_c, necessary for the synchronization phenomena to occur.
To investigate the behavior of Eq. (3.15) we now follow the analysis procedure proposed by Strogatz in *"From Kuramoto to Crawford: Exploring the onset of synchronization in the population of coupled oscillators"*, [58].

3.2.4 Order Parameter

The first step is to introduce the following complex order parameter:

$$re^{i\theta} = \frac{1}{N}\sum_{j=1}^{N} e^{i\phi_j} \tag{3.16}$$

where $0 \leq r \leq 1$ measures the global *phase coherence* level of the system and θ gives the average of the phases of all the oscillators. The right interpretation key to be given to Eq. (3.16) is the following: interpreting the phase of each oscillator as a point on the unit circumference, the order parameter $re^{i\theta}$ is represented by an arrow whose length depends on the state of the system and increases as the phases of the oscillators approach each other as shown in Fig. 3.2.

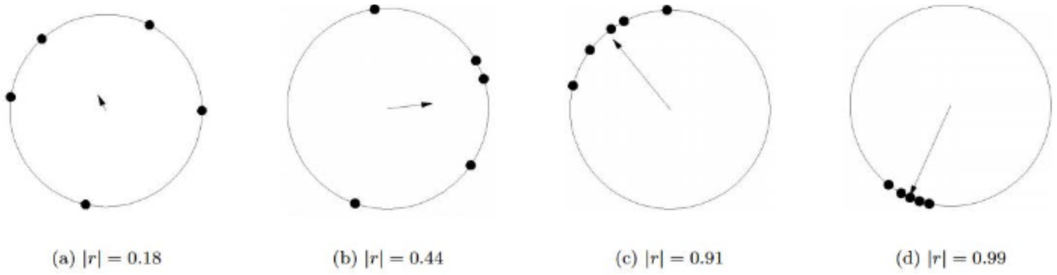

(a) $|r| = 0.18$ (b) $|r| = 0.44$ (c) $|r| = 0.91$ (d) $|r| = 0.99$

Figure 3.2: The order parameter is represented by an arrow from the center of the unit circle. Notice how the length of the order parameter increases as the oscillator phases approach each other. The point identified by $re^{i\theta}$ is called *centroid.*

The order parameter can be interpreted as the macroscopic quantity that indicates the collective rhythm of the oscillator system. In particular,

- $r \approx 0$ corresponds to a configuration in which no macroscopic rhythm is produced, because the oscillators are located uniformly on the circle and moves individually,
- $r \approx 1$ indicates that the phases are almost all similar and the system will act as a single macro-oscillator.
- In the case where $r(t)$ assumes an intermediate value, $0 < r(t) < 1$, there is partial synchronization, that is a part of the oscillators moves in a compact group, while the others continue their individual motion.

Using the order parameter it is possible to rewrite Eq. (3.15) in a more convenient form. Multiply both members of Eq. (3.16) for $e^{-i\phi_i}$ to obtain

$$re^{i(\theta-\phi_i)} = \frac{1}{N}\sum_{j=1}^{N} e^{i(\phi_j-\phi_i)} \qquad i = 1, \dots, N \tag{3.17}$$

whose imaginary part is

$$r\sin(\theta - \phi_i) = \frac{1}{N}\sum_{j=1}^{N} \sin(\theta - \phi_i) \tag{3.18}$$

Substituting this last relation in Eq. (3.15) we obtain the *governing equation* as a function of the order parameter:

$$\dot{\phi}_i = \omega_i + Kr\sin(\theta - \phi_i) \qquad i = 1, 2, \dots, N \tag{3.19}$$

It should be noted, in particular, how the interaction between the oscillators is described only in terms of the "mean-field" quantities r and θ. The coupling term shows how the phase of each individual oscillator is pushed towards the middle phase of the whole ensemble with a force proportional to the magnitude of the order parameter and the coherence r. This establishes a deep link between coupling and coherence. When the population becomes more and more coherent, the increase of r and therefore of the coupling tends to bring more oscillators in the synchronized group. If consistency continues to increase, the synchronization process continues, otherwise it stops.

3.2.5 Synchronization

The qualitative and heuristic considerations presented so far are analytically summarized in the following theorem, as stated and proved in [58]:

Theorem 3.2.2 ([58])**.** *Let be given a system of N identical oscillators that satisfy the model of Kuramoto, i.e. Eq. (3.19) with the assumptions described before. In the continuous thermodynamic limit, $N \to \infty$, there is a threshold value $K_c = \frac{2}{\pi g(\omega_0)} \geq 0$ for the coupling constant such that:*

$$\begin{cases} \lim_{t\to+\infty} r(t) = 0 & \textit{for } K < K_c \\ \lim_{t\to+\infty} r(t) = r_\infty \ (\textit{with } 0 < r_\infty \leq 1) & \textit{for } K \geq K_c \end{cases} \tag{3.20}$$

and where $g(\omega_0)$ is the value assumed in ω_0 by the distribution $g(\omega)$ of the frequencies (Fig. 3.48).

Proof. Following Kuramoto, we are looking for stable solutions of Eq. (3.19), in which $r(t)$ is constant and $\theta(t)$ rotates uniformly at the frequency ω_0, i.e.:

$$\begin{cases} \theta(t) = \theta_0 + \Omega t \\ r(t) = r, \quad \text{constant } (\leq 1) \end{cases} \tag{3.21}$$

Our system is described by Eq. (3.19):

$$\dot{\phi}_j = \omega_j + Kr\sin(\theta - \phi_j) \quad (= \Phi^j(\phi)) \quad j = 1 \dots, N \tag{3.22}$$

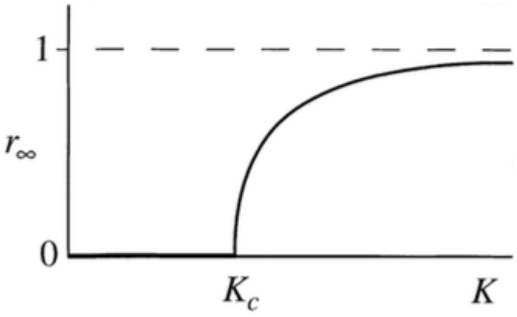


Figure 3.3: Below K_c there is only one stable solution, $r = 0$. This will be a completely incoherent state. As the coupling reaches K_c there is a nonzero solution, which branches off. As K increases to infinity r tends to 1. This will thus be a completely coherent state where every oscillator will be in phase due to the strong coupling. The above image was taken from Strogatz [58].

where $r(t)$ and $\theta(t)$ are defined by the definition (3.16) of the order parameter.
Since $r(t)$ is assumed to be constant in Eq. (3.19), the equations are decoupled and, therefore, the oscillators are all independent of each other. Our work now consists in solving the resulting motions of all the oscillators, which will depend on r as a parameter. These motions must give values for r and θ consistent with those assumed. The stratagem of this method for analyzing the system is the condition of *self-consistency* for the order parameter.
We start by checking the *compatibility* of (3.21) with the asymptotic motion of the oscillators:

$$\phi_j \to \phi_j^{(0)} + \Omega t \qquad \text{for } t \to +\infty \tag{3.23}$$

and by finding the necessary stability conditions.
We will proceed by exploiting the results in Appendix C, since this can be considered a simple case of that theory. The fundamental difference lies in the work environment. Now, in fact, we are not in $\mathbb{R}^N$, but in $\mathbb{T}^N$.
Consider the model of Kuramoto, that is described by the system:

$$\begin{cases} \dot{\phi}_j = \omega_j + \dfrac{K}{N}\displaystyle\sum_{j=1}^{N} \sin(\phi_i - \phi_j) \\ \dot{z}_j = 0 \end{cases}$$

where z_j, for $j = 1, \dots, N$, indicates the radius. The second equation only says that the radius remains stationary. For this reason it is useless for the dynamics of the problem, since it does not provide any information, but only clarifies the new geometry of the system. We will then work on the "universal covering" of $\mathbb{T}^N$, which is $\mathbb{R}^N$.

- Consider the t-dependent diffeomorphism:

  $$\psi_j = \phi_j - \phi_j^{(0)} - \Omega t \quad (= \psi(\phi, t))$$

  and calculate the conjugate field $\Psi(\psi)$ of $\Phi(\phi)$:

  $$\dot{\psi}_l = \Psi^l(\psi, t) = \sum_{j=1}^{N} \frac{\partial \psi^l}{\partial \phi_j}\Phi^j + \frac{\partial \psi^l}{\partial t}$$

that gives

$$\dot{\psi}_l = \omega_l + Kr\sin(\theta(t) - \psi_l - \phi_l^{(0)} - \Omega t) - \Omega$$

Substituting $\theta_0 = \theta(t) - \Omega t$, derived from the first equation in (3.21), in the previous equation, we gain:

$$\dot{\psi}_l = \omega_l - \Omega + Kr\sin(\theta_0 - \psi_l - \phi_l^{(0)}) \qquad (= \Psi^l(\psi)) \tag{3.24}$$

- We want the solution $\psi_l = 0$ to be an asymptotically stable equilibrium solution for Eq. (3.24):

  → $\psi_l = 0$ is an equilibrium for Eq. (3.24) if it is true:

$$0 = \omega_l - \Omega + Kr\sin(\theta_0 - \phi_l^{(0)}) \quad \Longrightarrow \quad \phi_l^{(0)} = \arcsin\frac{\omega_l - \Omega}{Kr} + \theta_0 \tag{3.25}$$

Hence, the condition that guarantees the existence of equilibria is given by

$$\left|\frac{\omega_l - \Omega}{Kr}\right| \leq 1 \qquad \Longrightarrow \qquad |\omega_l - \Omega| \leq Kr \tag{3.26}$$

Now examine the stability condition for $\psi_l^* = 0$. Consider Eq. (3.24) as

$$\frac{d\psi}{dt} = X(\psi(t)) \qquad \text{with} \qquad X(\psi^*) = 0$$

From theorem C.0.2, if $\Re(Spect(X'(\psi^*))) < 0$, ψ^* is an asymptotically stable equilibrium, i.e.

$$\left.\frac{\partial\Psi^l}{\partial\psi_m}\right|_{\psi_l^*=0} = -Kr\cos(\theta_0 - \phi_l^{(0)})\delta_{lm} < 0 \tag{3.27}$$

If we consider the weak condition: $-Kr\cos(\theta_0 - \phi_l^{(0)})\delta_{lm} \leq 0$, then we have "linear stability".
Starting from Eq. (3.27), the condition, that must satisfy ψ_l^*, $l = 1, \dots, N$, to be asymptotically stable, is:

$$\cos(\theta_0 - \phi_l^{(0)}) > 0, \qquad \text{i.e.} \qquad |\theta_0 - \phi_l^{(0)}| < \frac{\pi}{2} \tag{3.28}$$

thus: $\theta_0 - \phi_l^{(0)} \in]-\frac{\pi}{2}, \frac{\pi}{2}[$.
To sum up: if

1. $|\omega_l - \Omega| \leq Kr$
2. $\theta_0 - \phi_l^{(0)} \in]-\frac{\pi}{2}, \frac{\pi}{2}[$

then the solution $\psi_l^* = 0$ is asymptotically stable for Eq. (3.24). Hence, those oscillators, whose frequencies satisfy (3.26), will tend to the asymptotic motion (3.23). Then the quiet motions in the equilibria:

$$\psi_l(t) = \psi_l^*(t) \equiv 0 \qquad l = 1, \dots, N$$

correspond to the synchronization for those l indices that satisfy (3.26). These oscillators are called *locked*, because in their original reference system their phase is blocked at the

frequency Ω. If, however, it does not apply (3.26), that is $|\omega_l - \Omega| > Kr$, oscillators are called *drifting*; they move on the circumference unevenly, accelerating in the vicinity of some phases and slowing down near others.
The condition (3.26) is, therefore, fundamental to understand which oscillators will form the synchronized group and which ones do not. The system will be composed of a 'cluster' of synchronized oscillators (the oscillators that correspond to the l indices that satisfy (3.26)) and a set of individual oscillators that move around the circumference (corresponding to the indices l for which it does not apply (3.26)).

- Consider now the particular case where we choose

  (a) $\theta_0 = 0$

  (b) $\phi_j^{(0)}$, $j = 1, \dots, N$ are distributed symmetrically around zero in $]-\frac{\pi}{2}, \frac{\pi}{2}[$.

  We want to evaluate a posteriori if, in this simplified case, we obtain values of r and θ acceptable with those we have assumed (see (3.21)).
  Taking into account these hypotheses, the order parameter is determined through the definition (3.16):

$$re^{i\theta} = \frac{1}{N}\sum_{j=1}^{N} e^{i\phi_j(t)} = \frac{1}{N}e^{i(\phi_j^{(0)}+\Omega t)} = \Big[\frac{1}{N}\sum_{j=1}^{N}\cos\phi_j^{(0)} + \underbrace{i\frac{1}{N}\sum_{j=1}^{N}\sin\phi_j^{(0)}}_{=0 \text{ for } (a)}\Big]e^{i\Omega t}$$

By comparing the first and last member, we get:

$$\begin{cases} \theta(t) = \Omega t \\ r = \frac{1}{N}\sum_{j=1}^{N}\cos\phi_j^{(0)} < 1 \qquad (= 1 \text{ only if } \phi_j^{(0)} \equiv 0 \ \ \forall j) \end{cases}$$

As we can see what results is perfectly consistent with the (3.21) (remembering that we have placed $\theta_0 = 0$) and with what was expected. In this case, $\psi_l = 0$ is an equilibrium if, placed $\theta_0 = 0$ in (3.25), we get

$$0 = \omega_l - \Omega - Kr\sin(\phi_l^{(0)}) \qquad \Longrightarrow \qquad \phi_l^{(0)} = \arcsin\frac{\omega_l - \Omega}{Kr} \tag{3.29}$$

where it must be valid the existence condition:

$$|\omega_l - \Omega| \leq Kr \tag{3.30}$$

Remark. From (3.29) we see that the phases $\phi_l^{(0)}$, $l = 1, \dots, N$ are distributed according to condition (a), if the frequencies ω_l, $l = 1, \dots, N$, are symmetrically distributed around Ω.

If (3.30) holds for every $l = 1, \dots, N$ there is *total synchronization*, because all the oscillators tend to their asymptotic motions:

$$\phi_j \to \phi_j^{(0)} + \Omega t \qquad \forall\, l$$

We see that when the system totally synchronizes, the order parameter r is about 1. Since the frequencies of the oscillators are distributed according to the probability density $g(\omega)$, in order that (3.30) is satisfied for each $l = 1, \dots, N$, we can think that

$K \to +\infty$, so that the interval $[-Kr, Kr]$ can cover the whole range of frequencies. If $K \to +\infty$, since $r \leq 1$, from (3.29) we get $\phi_l^{(0)} \to 0$ and therefore $\omega_l \approx \Omega$. Hence $r \to 1$: in fact, since $\cos\phi_l^{(0)} \to 1$, one has that $r = \frac{1}{N}\sum_{l=1}^{N}\cos\phi_l^{(0)} \to 1$. The oscillators, considered as points on the unitary circumference, tend to overlap in a neighborhood of the same point and move in a compact block with frequency Ω.

- What remains to be understood is how $r(t)$ keeps constant even in the presence of errant oscillators that run disorderly on the circumference. Kuramoto required that oscillators of this type formed a uniform distribution on the circumference. In this way the centroid is fixed, even though the oscillator is still moving.
Denote with $\rho(\phi, \omega, t)d\phi$ the fraction of oscillators with angular frequency ω and angular position between ϕ and $\phi + d\phi$ at time t. For the stationary hypothesis we have that $\rho(\phi, \omega, t)$ is inversely proportional to the velocity in ϕ

$$\rho = \frac{C}{v} \tag{3.31}$$

that is the oscillators accumulate where they go slower and thin out when they are faster along the circumference. Hence:

$$\rho = \begin{cases} \delta\Big(\phi - \psi - \arcsin\Big(\dfrac{\omega - \Omega}{Kr}\Big)\Big)H(cos(\phi)) & |\omega - \Omega| < Kr \\ \dfrac{C}{|\omega - \Omega + Kr\sin(\phi)|} & \text{otherwise} \end{cases} \tag{3.32}$$

where H is the Heaviside function

$$H(x) = \begin{cases} 1 & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.33}$$

and C is found by imposing the normalization condition of the function ρ:

$$\int_{-\pi}^{\pi} \rho(\phi, \omega, t)d\phi = 1 \quad \text{for every } \omega.$$

This conclusion is never shown in Kuramoto's works, since he had and worked on an intuition. The first mathematician to deal with this gap was Strogatz, as we find in the article *"From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators"*, [58]. The proof is shown in Appendix D.

- Now calculate the C constant using the residuals method (for a theoretical reference see Appendix E). We want to evaluate

$$\int_{-\pi}^{\pi} \frac{C}{\omega - \Omega + Kr\sin(\phi)}d\phi \quad 0 < Kr < \omega - \Omega$$

$$\begin{aligned} \int_{-\pi}^{\pi} \frac{C}{\omega - \Omega + Kr\sin(\phi)}d\phi &= \int_{0}^{2\pi} \frac{C}{\omega - \Omega + Kr\sin(\phi)}d\phi = \\ = \int_{\gamma} \frac{C}{(\omega - \Omega) + \frac{Kr}{2i}(z - \frac{1}{z})}\frac{dz}{iz} &= 2iC\int_{\gamma} \frac{dz}{iz(2i(\omega - \Omega) + Kr(z - \frac{1}{z}))} = \\ &= 2C\int_{\gamma} \frac{dz}{Krz^2 + 2i(\omega - \Omega)z - Kr} \end{aligned}$$

We look for the poles of the integrand function:

$$z_{\pm} = \frac{-2i(\omega - \Omega) \pm \sqrt{-4(\omega - \Omega)^2 + 4K^2r^2}}{2Kr} = \frac{-i(\omega - \Omega) \pm i\sqrt{(\omega - \Omega)^2 - K^2r^2}}{Kr}$$

with $(\omega - \Omega)^2 - K^2r^2 > 0$. There are two distinct poles, so they are of the first order. Now we have to evaluate the modules, since to calculate the integral with the residuals method we have to consider the ones that have module less than 1.

→ Surely $|z_-| > 1$.

→ Now consider the other pole. $z_\pm$ are solutions of the equation:

$$(z - z_+)(z - z_-) = z^2 + \frac{2i(\omega - \Omega)}{Kr}z + 1$$

then we have $z_+z_- = 1$; then, going to the modules $|z_+||z_-| = 1$, since $|z_-| > 1$, we obtain $|z_+| < 1$.

In conclusion, the pole that interests to calculate the integral is z_+. Therefore we gain

$$\int_{-\pi}^{\pi} \frac{C}{\omega - \Omega + Kr\sin(\phi)}d\phi, \qquad 0 < Kr < \omega - \Omega$$

$$\begin{aligned}\int_{-\pi}^{\pi} \frac{C}{\omega - \Omega + Kr\sin(\phi)}d\phi &= (2C)(2\pi i)Res_{z_+}\frac{1}{Krz^2 + 2i(\omega - \Omega)z - Kr} &=\\ &= (2C)(2\pi i)\frac{1}{2Krz + 2i(\omega - \Omega)}\Big|_{z=z_+} = 2\pi Ci\frac{1}{i\sqrt{(\omega - \Omega)^2 - K^2r^2}} &=\\ &= \frac{2\pi C}{\sqrt{(\omega - \Omega)^2 - K^2r^2}}\end{aligned}$$

hence

$$C = \frac{1}{\int_{-\pi}^{\pi}\rho(\phi, \omega, t)d\phi} = \frac{\sqrt{(\omega - \Omega)^2 - K^2r^2}}{2\pi} \tag{3.34}$$

If $C(\omega) = 0$, in order that ρ is normalizable, it is necessary that ρ be a Dirac delta function with a peak corresponding to a certain ϕ':

$$C(\omega) = \rho v = \delta(\phi - \phi')v = \delta(\phi - \phi')(\omega - \Omega + Kr\sin(\phi)) = 0 \tag{3.35}$$

Integrating with respect to ϕ we gain

$$\omega - \Omega + Kr\sin(\phi') = 0 \tag{3.36}$$

- We have identified two groups of oscillators: the *locked oscillators* and the *drifting oscillators*. In the resolution with respect to the order parameter it is therefore possible to divide the sum into two parts, each representing the contribution of one of these groups. We obtain

$$\langle e^{i\phi}\rangle = re^{i\theta} = \langle e^{i\phi}\rangle_{lock} + \langle e^{i\phi}\rangle_{drift} \tag{3.37}$$

where we consider the average over the entire population of oscillators. Using the self-consistent condition in order to gain a constant value for r in agreement with the definition (3.16) of the order parameter we have

$$r = e^{-i\theta}\langle e^{i\phi}\rangle_{lock} + e^{-i\theta}\langle e^{i\phi}\rangle_{drift} \tag{3.38}$$

The contribution of the locked oscillators is given by

$$e^{-i\theta}\left\langle e^{i\phi}\right\rangle_{lock}=\int_{-Kr+\Omega}^{Kr+\Omega}e^{i(\phi-\theta)}g(\omega)d\omega=\int_{-Kr+\Omega}^{Kr+\Omega}e^{i(\phi^{(0)}+\Omega t-\theta)}g(\omega)d\omega \tag{3.39}$$

where it is taken into account that for these oscillators the asymptotic motions, described by Eq. (3.23), are valid. Applying a change of variables, from (3.25) we obtain

$$\omega=\omega(\phi^{(0)})=\Omega+Kr\sin(\phi^{(0)}-\theta_0) \tag{3.40}$$

and substituting in Eq. (3.39) we obtain an integral in $\phi^{(0)}$:

$$e^{-i\theta}\left\langle e^{i\phi}\right\rangle_{lock}=Kr\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}e^{i(\phi^{(0)}-\theta_0)}g(\Omega+Kr\sin(\phi^{(0)}-\theta_0))\cos(\phi^{(0)}-\theta_0)d\phi^{(0)} \tag{3.41}$$

Set $\phi^{(0)}-\theta_0=x$:

$$e^{-i\theta}\left\langle e^{i\phi}\right\rangle_{lock}=Kr\int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0}e^{ix}g(\Omega+Kr\sin x)\cos x\ dx$$

For the contribution due to the errant oscillators we have

$$e^{-i\theta}\left\langle e^{i\phi}\right\rangle_{drift}=\int_{-\pi}^{\pi}\int_{|\omega-\Omega|>Kr}e^{i(\phi-\theta)}\rho(\phi,\omega)g(\omega)d\omega d\phi \tag{3.42}$$

This integral is zero, because it is valid: $\rho(\phi,\omega)=\rho(\phi+\pi,2\Omega-\omega)$.
Therefore the self-consistent equation we have reached is given by:

$$r=(e^{-i\theta}\left\langle e^{i\phi}\right\rangle_{lock}=)Kr\int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0}e^{ix}g(\Omega+Kr\sin x)\cos x\ dx \tag{3.43}$$

Equation (3.43) has a trivial solution for $r=0$ for every value of K, that corresponds to the state of complete incoherence with $\rho(\phi,\omega)=\frac{1}{2\pi}$ for every ϕ and ω.
What is surprising is that this inconsistent state is always possible, regardless of how frequencies are distributed. Even if they are all the same, the inconsistency can last forever, once it is established as an initial condition.
Eq. (3.43) is equivalent to the following pair of equations, obtained by equating the imaginary and the real part of its two members:

$$0=\int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0}g(\Omega+Kr\sin x)\sin x\ \cos x\ dx \tag{3.44}$$

$$1=K\int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0}g(\Omega+Kr\sin x)\cos^2 x\ dx \tag{3.45}$$

From these last two equations we obtain the values of r and Ω. By hypothesis, $g(x)$ is symmetric with respect to ω_0; therefore, Eq. (3.44) can be satisfied if and only if

$$\Omega=\omega_0 \tag{3.46}$$

Eq. (3.45) gives a family of solutions corresponding to the partially synchronized state. These solutions bifurcate continuously starting from $r=0$, when the coupling constant is K_c. To calculate it, take the limit of Eq. (3.45) for $r \to 0^+$; we want to see under which conditions we begin to come out of the completely incoherent state.

$$\begin{aligned} 1 &= \lim_{r\to 0^+} K_c \int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0} g(\omega_0 + Kr\sin x)\cos^2 x\ dx = \\ &= K_c g(\omega_0) \int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0} \cos^2 x\ dx = K_c g(\omega_0)\frac{\pi}{2} \end{aligned} \tag{3.47}$$

Hence:

$$K_c = \frac{2}{\pi g(\omega_0)} \tag{3.48}$$

which is exactly the critical value needed to have states in which synchronization phenomena start to manifest. Unlike what happens for the inconsistency, this state is not always possible. It exists only above a certain K_c threshold.

□

3.2.6 Initial Growth of K_c

After calculating the value of K_c, we want to study the initial growth of that value, as presented by Strogatz in [58].
Assume for simplicity that $\Omega = \omega_0 = 0$. To understand how r behaves near the critical point K_c, expands $g(Kr\sin(x))$ around 0 with Taylor:

$$g(Kr\sin(x)) \approx g(0) + g'(0)(Kr\sin(x)) + \frac{1}{2}g''(0)(Kr\sin(\phi))^2 \tag{3.49}$$

Since for hypothesis $g(\omega)$ has a maximum in 0, we have $g'(0) = 0$ and thus Eq. (3.49) becomes

$$g(Kr\sin(x)) \approx g(0) + \frac{1}{2}g''(0)(Kr\sin(\phi))^2 \tag{3.50}$$

If we substitute this relation in Eq. (3.45) we gain

$$1 = K\int_{-\frac{\pi}{2}-\theta_0}^{\frac{\pi}{2}-\theta_0} \cos^2(x)[g(0) + \frac{1}{2}g''(0)(Kr\sin(\phi))^2]\ dx \tag{3.51}$$

Solving the integral and using Eq. (3.48) we obtain

$$1 = K\Big[g(0)\frac{\pi}{2} + \frac{K^3r^2g''(0)}{2}\frac{\pi}{8}\Big] = \frac{K}{K_c} + \frac{\pi K^3r^2g''(0)}{16} \tag{3.52}$$

Now multiply both members by K_c and assume that $K \approx K_c$ since we are working near the critical point. We get

$$K_c = K + K_c\frac{\pi K^3r^2g''(0)}{16} \tag{3.53}$$

If we set $\mu \equiv \frac{K-K_c}{K_c} = -\frac{\pi K^3r^2g''(0)}{16}$ the distance from the critical point, solving Eq. (3.53) for the order parameter we obtain

$$r = \sqrt{\frac{-16\mu}{\pi g''(0)K_c^3}} \tag{3.54}$$

or equivalently

$$r = \sqrt{\frac{-16}{\pi g''(0) K_c^4}} (K - K_c)^{1/2} \tag{3.55}$$

For $K \searrow K_c$, then r is proportional to the square root of the distance from K_c as exemplified by Figure 3.4. In particular, for g unimodal and sufficiently smooth around $\omega = 0$ (implying $g''(0) < 0$), the partially synchronized state bifurcates super-critically from the incoherent state for $K > K_c$ indicating a so-called *second-order phase transition*.

In the absence of a well defined and known distribution function $g(\omega)$, it is not possible to go beyond Eq. (3.55), expliciting $r(K)$. As an example, we consider the particular case that g is *Lorentzian* (or Cauchy):

$$g(\omega) = \frac{\Delta}{\pi(\Delta^2 + \omega^2)} \tag{3.56}$$

where $\Delta > 0$ is the scale parameter, that define the amplitude of the distribution. The integral in (3.45) can then be evaluated explicitly, resulting in

$$r = \sqrt{1 - \frac{2\Delta}{K}} \tag{3.57}$$

or equivalently, considering that $K_c = \frac{2}{\pi g(0)} = \frac{2}{\pi \frac{1}{\pi\Delta}} = 2\Delta$,

$$r = \sqrt{1 - \frac{K_c}{K}} \tag{3.58}$$

for $K \geq K_c$, with $K_c = 2\Delta$. According to Figure 3.4, this result for the continuous limit model (D.1)-(D.11) coincides quite well with simulations carried out for the original model (3.15) with large but finite N.

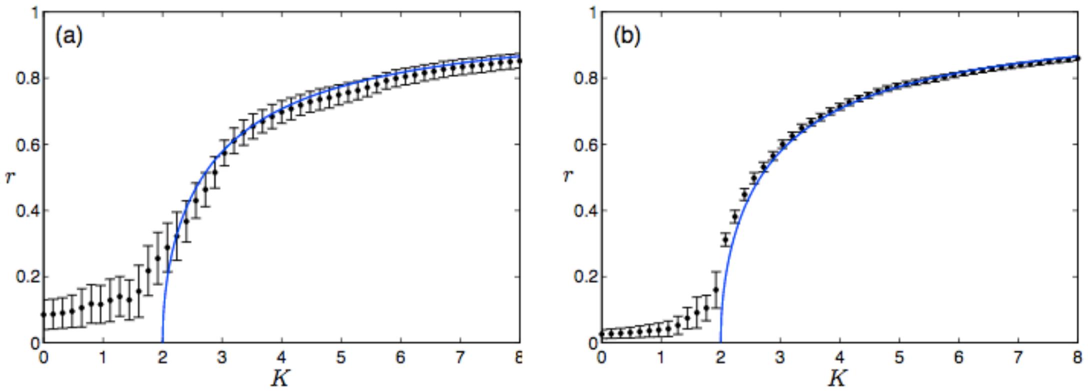


Figure 3.4: Phase diagram for the Kuramoto model (3.15) in the case of the Lorentzian frequency distribution (3.56) with $\Delta = 1$. Figures (a) and (b) show the simulation results obtained with $N = 100$ and $N = 1000$ oscillators respectively, having natural frequencies that are the same for each evaluated K. In both figures, the blue plot corresponds to (3.58). For each investigated K, the corresponding r is the result from averaging over a time period after the system has settled into a stationary state, with error bars representing standard deviations. Figure taken from the bachelor's thesis of Zeegers [68], where all the details about the numerical scheme used and the Matlab code are provided.

3.2.7 The bimodal case

So far we have extensively discussed the phenomenology of the mean-field Kuramoto model when g is symmetric and unimodal. A natural question is to what extent these results still apply when g is shaped differently. An obvious choice is to consider a g that consists of two peaks. This bimodal frequency distribution has been investigated in the literature (see for instance [34] and [36]), but the model proved to be considerably more difficult to analyze than in the unimodal setting.

We will restrict ourselves to the results presented in [36]. In this study, exact results have been obtained for the case that g is equal to the sum of two identical Lorentzian densities:

$$g(\omega) = \frac{\Delta}{2\pi}\Big(\frac{1}{\Delta^2 + (\omega - \omega_0)^2} + \frac{1}{\Delta^2 + (\omega + \omega_0)^2}\Big) \tag{3.59}$$

In (3.59), Δ is again the scale parameter of the two Lorentzians and these are centered at $\pm\omega_0$ (see Figure 3.5). The analytical and numerical results of [36] are summarized in Fig. 3.6 and show new phenomena compared to the unimodal case.

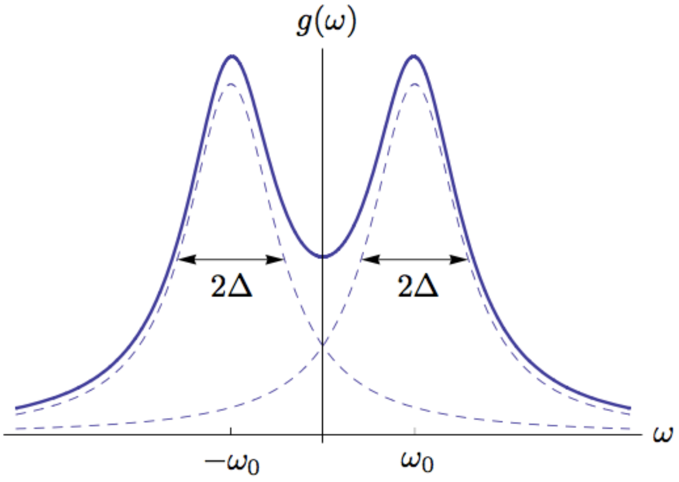


Figure 3.5: The bimodal distribution given by (3.59), being the sum of two identical Lorentzians. Adapted from [36].

- For $\omega_0/\Delta < 1/\sqrt{3}$, we simply get the previously considered phase transition between incoherence and partial synchronization. Indeed, this is precisely the region for which the distribution (3.59) is unimodal, since the peaks are not far enough from each other, in relation to their widths, to break unimodality.
- In the region $1/\sqrt{3} < \omega_0/\Delta < 1$, the distribution still is hardly bimodal. However, what is new is that close to the phase transition boundary both the incoherent state and the partially synchronized state are stable. In this so-called *bistable* region, it depends on the initial distribution of the oscillators whether incoherence or synchronization prevails.
- For $\omega_0/\Delta > 1$, the peaks of g are sufficiently separated such that stable *standing waves* emerge in the phase diagram. In these states, the population is split into two counter-rotating clusters of oscillators. This happens in the intermediate region where the

coupling is strong enough to lock oscillators with frequencies around the same peak of g, but too weak to lock the two clusters. As we see from Figure 3.6, the standing waves coexist with partially synchronized states for $1 < \omega_0/\Delta < 1.18$ and this bistability has fully disappeared for $\omega_0/\Delta > 1.81$.

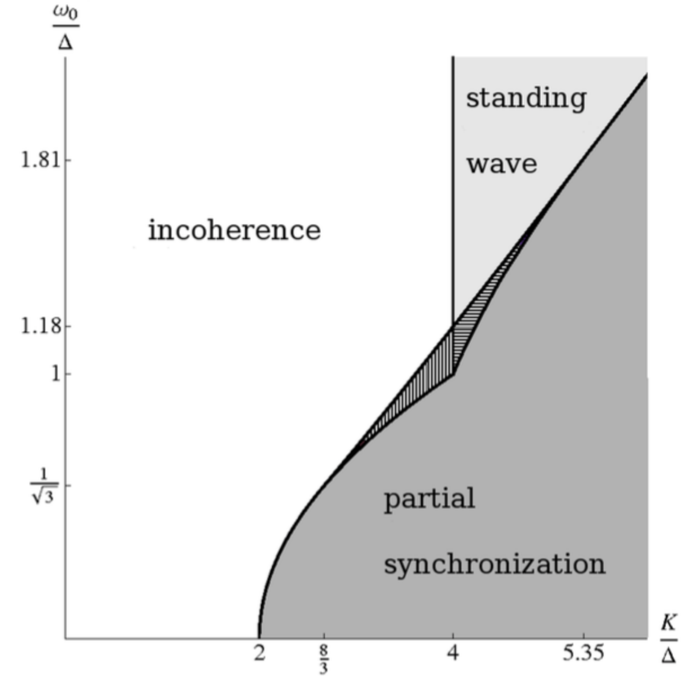


Figure 3.6: Phase diagram for the mean-field Kuramoto model with g given by (3.59). It consists of regions of incoherence (white), partial synchronization (dark gray), standing waves (light gray) and bistability, where in the last case partially synchronized states coexist either with incoherent states (vertical lines) or with standing waves (horizontal lines). Figure taken from [36].

The question on how the states in Figure 3.6 bifurcate from each other is more delicate than in the unimodal setting. A detailed bifurcation analysis can be found in [36], where in addition similar results have been obtained for the case that g is a sum of two identical Gaussians.

3.2.8 Summary of the Kuramoto Model

We now summarize what we have seen up to now about the Kuramoto model.
Kuramoto proposed a model that can be solved precisely for the synchronization of weakly coupled oscillators and this is governed by

$$\dot{\phi}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N}\sin(\phi_j - \phi_i) \qquad i = 1, 2, \ldots, N$$

where $\phi_i(t)$ is the phase of the i-th oscillator and ω_i the natural frequency, chosen randomly from the following Lorentzian probability distribution

$$g(\omega) = \frac{\Delta}{\pi(\Delta^2 + \omega^2)}$$

Using the self-consistent hypothesis, the model is solved for $N \to \infty$ and $t \to \infty$ with the order parameter expressed by

$$r(t) = \left| \frac{1}{N} \sum_{j=1, j \neq i}^{N} e^{i\phi_j} \right|$$

Then the following relation

$$r = \begin{cases} 0 & K < K_c \\ \sqrt{1 - \dfrac{K_c}{K}} & K \geq K_c \end{cases}$$

where $K_c = 2\Delta$, expresses the condition for which a synchronized state is verified or not. We have then seen in section 3.2.5 that the conditions to have an asymptotically stable point are:

1. $|\omega_l - \Omega| \leq Kr$
2. $\theta_0 - \phi_l^{(0)} \in]-\frac{\pi}{2}, \frac{\pi}{2}[$

Oscillators whose frequency satisfies condition 1 tend to the asymptotic synchronization motion (3.23), therefore they are said to be *locked*, since their phase ϕ coincides with the frequency Ω. Oscillators not respecting the equilibrium condition move along the circumference unevenly. Condition 1 is, thus, fundamental to understand which oscillators form the synchronized group and which are not.

What has been done so far is a mere investigation of the validity and plausibility of the model, which highlights the conditions of existence for the asymptotic stability of weakly interacting oscillator systems. To verify the stability of the equilibrium of the system, we used the *Lyapunov Spectral Method*, where a uniformly rotating reference system was introduced with the synchronous asymptotic angular velocity (Ω). The value that it assumes is therefore of fundamental importance: for excessively high or reduced values, the model does not work. However, a fundamental condition for the occurrence of synchronism is that there must not be too much difference in the oscillator population.

3.3 Structural stability of the Kuramoto model

The following section analyzes the robustness of the model treated, focusing attention on the properties of the system's equilibrium points which are preserved, completely or partially, in the presence of a weak perturbation. The perturbation will concern both the vector field, with a $\mathcal{C}^0$ function close to the original, and the initial condition of the Cauchy Problem.

Intuitively, the behavior of a system is stable if, subject to perturbation, the resulting movement is "quite" similar to the original one. In order to define quantitatively the concept of "close enough" it is necessary to fix a distance, and therefore a norm, in the space of the states. In this case it is possible to give the following topological definitions:

Definition 3.3.1 (Absolute Stability and Total Stability of Dubovin)**.** *Given the vector field $X(x,t)$ related to the differential equation $\dot{x} = X(x,t)$, let x^* be an equilibrium point: $X(x^*,t) = 0$ and let $x(x_0,t)$ be a solution for the equation with initial data x_0.*

x^* is said to be absolutely stable *for $\dot{x} = X(x,t)$ if, for every $\epsilon > 0$, exists $\delta > 0$ such that, for every $x_0 \in B(x^*, \delta)$, we have that:*

- *$x(x_0,t) \in B(x^*, \epsilon)$ for every $t > 0$, i.e. x^* is stable;*

- $\lim_{t \to +\infty} x(x_0, t) = x^*$.

x^* is said to be totally stable *for $\dot{x} = X(x,t)$ if, for every $\epsilon > 0$, exists $\delta > 0$ such that*

- *for every initial data x_0: $|x_0 - x^*| < \delta$*
- *for every perturbation $Y(x,t)$ of the vector field characterized by the condition*

$$\sup_{t \in \mathbb{R}} \sup_{x \in \mathbb{R}^N} |Y(x,t)| < \delta$$

then the solutions of

$$\begin{cases} \dot{x} = X(x,t) + Y(x,t) \\ x(0) = x_0 \end{cases}$$

are ϵ-limited near x^*, that is $|x(x_0,t) - x^*| < \epsilon$ for every $t \geq 0$.

Remark. Note tha $X(x^*,t) = 0$ does not imply necessarily that $X(x^*,t) + Y(x^*,t) = 0$.

Malkin theorem aims to provide a characterization of the structural stability of a dynamic system: in the presence of perturbations, in particular, it ensures at least the total stability around the asymptotically stable equilibria of the unperturbed system. The proof of the theorem follows from the *Gronwall's Lemma* or from the *Lyapunov function of asymptotic stability.* The first alternative is developed below.

3.3.1 Malkin Theorem

In this section we present the proof of Malkin's Theorem as it guarantees us the sufficient condition to obtain the total stability of the equilibrium model. This theorem is exploited since the asymptotic stability has been verified previously by means of the Spectral Method[48]. To develop the proof of this theorem, Gronwall's Lemma is used, that is so defined.

Lemma 3.3.2 (Gronwall's Lemma)**.** *Considering the particular case in which parameters a and b are not time-dependent, then take the parameters $a, b > 0$ and $\phi(t) \geq 0$ for every $t \in [0, +\infty)$. If by hypothesis the following differential inequality holds*

$$\dot{\phi}(t) \leq a\phi(t) + b \tag{3.60}$$

then we have:

$$\phi(t) \leq e^{at}\Big(\phi(0) + \frac{b}{a}\Big) - \frac{b}{a} \tag{3.61}$$

Proof. Consider the following Cauchy problem

$$\begin{cases} \dot{\psi} = a\psi + b \\ \psi(0) = \phi(0) \end{cases}$$

whose solution is

$$\psi(t) = e^{at}\phi(0) + \int_0^t e^{a(t-s)} b ds = e^{at}\Big(\phi(0) + \frac{b}{a}e^{-as|_{s=0}}\Big) - \frac{b}{a} = e^{at}\Big(\phi(0) + \frac{b}{a}\Big) - \frac{b}{a}$$

[48] Malkin's Theorem guarantees the total stability around the asymptotically stable equilibrium points. Therefore, if the asymptotically stable equilibrium points have been identified with the Spectral Method, the total stability of the model will occur around those points.

For hypothesis we have

$$\dot{\phi}(t) \leq \dot{\psi}(t), \quad \forall\, t > 0, \qquad \text{and} \qquad \phi(0) = \psi(0)$$

thus the thesis is verified

$$\phi(t) \leq e^{at}\Big(\phi(0) + \frac{b}{a}\Big) - \frac{b}{a}$$

□

Remark. If we take $a, b > 0$ for every $t \in [0, \infty)$ we have

$$|\dot{\phi}(t)| \leq a|\phi(t)| + b \tag{3.62}$$

then

$$|\phi(t)| \leq e^{at}\Big(|\phi(0)| + \frac{b}{a}\Big) - \frac{b}{a} \tag{3.63}$$

Proof. The thesis follows immediately from[49]

$$\frac{d}{dt}|\phi(t)| \leq |\dot{\phi}(t)| \leq a|\phi| + b \tag{3.64}$$

thus we use the previous Lemma 3.3.2 for $\phi(t) \to |\phi(t)|$. □

We now state and prove *Malkin theorem* using relation (3.63).

Theorem 3.3.3 (Malkin Theorem)**.** *Given the differential equation for the vector field* $X : \mathbb{R}^N \to \mathbb{R}^N$

$$\begin{cases} \dot{x}_X = X(x, t) \\ x_X(0) = x_0 \end{cases}$$

with x^* an equilibrium point asymptotically stable, then x^* is totally stable, i.e. $\forall\ \epsilon > 0$ exist a $\eta(\epsilon) > 0$ and $\delta(\epsilon) > 0$ such that

- *for every vector field of perturbation $Z = X + Y$, where this field is $\mathcal{C}^0$-close to X, with*

$$\|X(\cdot) - Z(\cdot)\|_{\mathcal{C}^0} \coloneqq \sup_{x \in \mathbb{R}^N} |X(x) - Z(x)| < \eta(\epsilon) \tag{3.65}$$

- *and for every initial data x_0*

$$|x_0 - x^*| < \delta(\epsilon)$$

we have that $x_Z(x_0, t)$ is a solution for the differential equation of the perturbed field

$$\begin{cases} \dot{x}_Z = Z(x_Z, t) \\ x_Z(0) = x_0 \end{cases}$$

such that

$$|x_Z(x_0, t) - x^*| < \epsilon \qquad \textit{for every } t \geq 0.$$

Remark. We can state Malkin theorem also referring to a unique $\hat{\delta}(\epsilon)$: it is sufficient to consider, instead of $\eta(\epsilon)$ and $\delta(\epsilon)$, $\hat{\delta}(\epsilon) = \min\{\eta(\epsilon), \delta(\epsilon)\}$.

[49] This relationship always occurs because the derivative of an absolute value must also take into account the sign function.

Proof. Consider the vector field Z η-uniformly close to X, i.e. in $\mathcal{C}^0$ norm where the value of η will be tuned subsequently:

$$\|X(\cdot) - Z(\cdot)\| < \eta$$

We define the function $\phi(t)$ as

$$\phi(t) := x_X(x_0, t) - x_Z(x_0, t) \tag{3.66}$$

and the modulus of the derivative of that function (3.66) is given by

$$|\dot{\phi}(t)| = |X(x_X) - Z(x_Z)| \tag{3.67}$$

Remark. 1. In order that for the differential equation

$$\begin{cases} \dot{x} = X(x,t) \\ x(0) = x_0 \end{cases} \tag{3.68}$$

there is a theorem of existence and uniqueness, $X : \mathbb{R}^N \to \mathbb{R}^N$ has to be uniformly Lipschitz $\forall\ t \in [0,T]$: that is, locally at x_0, in a neighborhood U of x_0, there must exist a number $L > 0$ such that $\forall\ x_1, x_2 \in U$ and $\forall\ t \in [0,T]$ the following relation holds

$$|X(x_1) - X(x_2)| \leq L|x_1 - x_2|$$

2. It is known that if X is $\mathcal{C}^1$ then X is Lipschitz.

3. If $X \in \mathcal{C}^1$ in the neighborhood U, with U connected, then

$$L = \sup_{x \in U} |\nabla X(x)|$$

Therefore if $X \in \mathcal{C}^1$ then the function is Lipschitz and there exists a positive number L such that

$$|\dot{\phi}(t)| = |X(x_X) - Z(x_Z)| = |X(x_X) - X(x_Z) + X(x_Z) - Z(x_Z)| \leq L|x_X - x_Z| + \eta$$

where

$$L|x_X - x_Z| + \eta = L|\phi(t)| + \eta$$

and thus we have

$$|\dot{\phi}(t)| \leq L|\phi(t)| + \eta$$

From Gronwall's Lemma (3.63) we obtain

$$|\phi(t)| \leq e^{Lt}\left(\frac{\eta}{L}\right) - \frac{\eta}{L}$$

since for $t = 0$, (3.66) becomes $\phi(0) = x_X(x_0, 0) - x_Z(x_0, 0) = 0$.
From $e^x - 1 \leq xe^x$, we get

$$|\phi(t)| \leq \frac{\eta}{L}\left(e^{Lt} - 1\right) \leq \eta t e^{Lt} \tag{3.69}$$

By hypothesis we have that x^* is an asymptotically stable equilibrium and therefore primarily stable. From the definition of simple stability of x^* for X we have that, $\forall\ \epsilon > 0$, $\exists\ \delta > 0$ such that for every initial datum

$$|x_0 - x^*| < \delta(\epsilon)$$

we have

$$|x_X(x_0,t) - x^*| < \frac{\epsilon}{2}, \qquad \forall\, t \geq 0$$

Furthermore, the asymptotic stability of x^* gives for X that: $\forall\, \epsilon > 0$ and for $\delta > 0$ determined before, exists a time $\tau > 0$ such that for $t \geq \tau$

$$|x_X(x_0,t) - x^*| < \frac{\delta}{2} \quad \left(\leq \frac{\epsilon}{2}\right)$$

At ϵ, δ and τ like above and for initial data x_0 such that

$$|x_0 - x^*| < \delta$$

we have that $\forall\, t \in [0,\tau]$:

$$|x_Z(x_0,t) - x^*| \leq |x_X(x_0,t) - x^*| + |x_Z(x_0,t) - x_X(x_0,t)| \leq \frac{\epsilon}{2} + \underbrace{\eta t e^{Lt}}_{\leq \frac{\delta}{2} \leq \frac{\epsilon}{2}} \leq \epsilon$$

In order to make this works, it is sufficient to tune the value of $\eta = \eta(\epsilon)$, which measures the perturbation of the vector field, in order that the following relation holds:

$$\eta \tau e^{L\tau} \leq \frac{\delta}{2} \tag{3.70}$$

□

To better understand the implications of the theorem, we focus our attention on what happens at time $t = \tau$. We have that

$$|x_Z(x_0,\tau) - x^*| \leq \underbrace{|x_X(x_0,\tau) - x^*|}_{\leq \frac{\delta}{2} \text{ for asymptotic stability}} + \underbrace{|x_Z(x_0,\tau) - x_X(x_0,\tau)|}_{\leq \frac{\delta}{2}} \leq \delta \tag{3.71}$$

As expressed by the relation (3.71) we note that the perturbed solution $x_Z(x_0,\tau)$ at time $t = \tau$ is (*re*-)entered in the ball of radius δ centered in x^* and this occurs cyclically for $t = m\tau$, for $m = 1, \ldots, \infty$; this indicates that the simple stability behavior around x^* has been gained for the vector field Z, although x^* is not necessarily an equilibrium point for it. Thanks to the mathematical proof just presented, we reach an important result for the existence of the model; through Malkin theorem is confirmed that although the ruling function is disturbed, the model is still valid. This is of fundamental importance, especially in light of a real application, where the representative function will never be the theorized one. In conclusion we can see how the strength of this model resides precisely in the possible practical applicability and it is not reduced only in the theoretical plane.

3.4 Improvements of the Kuramoto model

The Kuramoto model represents a famously tractable, but also fairly generic, model of mass synchronization in biological systems. It has been applied to a plethora of phenomena and is sometimes proposed in connection with neural network dynamics. Due to the complexity of interactions in neural networks there have been relatively few direct applications of the Kuramoto model to neurobiological data. To perhaps begin to remedy this situation, following the article of Timms and English [61] we propose a combination of two extensions or modifications to the Kuramoto model that are very important in neuroscience: the inclusion of spatial embedding through time delays and the inclusion of variable synaptic strength through dynamically changing coupling. These two modifications together are most likely indispensable if this model is to be successful in describing many aspects of neural processes.

3.4.1 Time delays

Time delays represent an essential characteristic because neuronal axons have finite signal transmission speeds of the electric signal in the axon. In fact, it arrives at the synapses of another neuron not instantaneously, but after the period of time τ which has served to cover the spatial distance between one neuron and another. The fact that signals often take physiologically significant time to reach their destinations is fundamental to the design of neural networks and the lack of time delays in the original Kuramoto model has been identified as a significant obstacle. The easiest way to include a time delay is to modify Eq. (3.15) to

$$\dot{\phi}_i(t) = \omega_i + \frac{K}{N}\sum_{j=1}^{N} \sin(\phi_j(t-\tau_{ij}) - \phi_i(t)) \tag{3.72}$$

Here, $\tau_{ij} = d_{ij}/v$, where d_{ij} is the distance from oscillator i to oscillator j and v is the velocity of the signal. Thus in this formulation the relative spatial positions of the oscillators become important. Note that a change in τ_{ij} can correspond to the modification of the distance between components or to the variation of the velocity v of the signal.
As presented in [10], the time delay τ_{ij} can be translated into a corresponding phase offset η_{ij} obtaining

$$\dot{\phi}_i(t) = \omega_i + \frac{K}{N}\sum_{j=1}^{N} \sin(\phi_j - \phi_i - \eta_{ji}) \tag{3.73}$$

In terms of synchrony, an elaborate synchronization behavior is now observed. For $\tau_{ij} = 0$, $\forall\ i, j$ the original model is found; for $\tau_{ij} \neq 0$, $\forall\ i, j$, a given K and $N \to \infty$, we have multiple stable solutions for Eq. (3.72) with the presence of several clusters, each of which consists of a large number of oscillating units with different frequencies and various basins of attraction. Furthermore, in a regime of intense coupling strength and strong delay, the 2D system has evolved towards a state characterized by the lowest possible frequency among those that are solutions of Eq. (3.72) (Niebur et al., 1991) and the delay, depending on the distance, induces in the solutions various spatial structures, such as spirals and traveling rolls (Jeong et al., 2002). The more complex dynamics due to η suggests the notion of *frustration*, whereby the sinusoidal interaction functions require some phase offset $\phi_j - \phi_i \neq 0$ in order to vanish. Thus, the presence of η causes the interaction functions to pull the phases away from absolute synchrony, even when the natural frequencies are identical. On closer inspection, the dynamics of this new model can be approximated with that of networks in which there is no delay, but in which the strength of the connections varies periodically with the spatial distance. For this purpose, it is therefore crucial to order the time delay according to a spatial metric $\eta_{ij} \propto |x_j - x_i|$ either in one dimension or in higher dimensions. In this way, we obtain the following equation[50]

$$\dot{\phi}_i(t) = \omega_i + \frac{K}{N}\sum_{j=1}^{N} cos(\eta_{ji}) \sin(\phi_j - \phi_i) \tag{3.74}$$

where it can be observed that the coupling strength has a maximal amplitude for $\cos(\eta_{ji}) = \pm 1$.

[50] A detailed analysis of the dynamics of the latter system is outside our scope, therefore the interested reader can refer to [10].

3.4.2 Dynamically changing coupling strength

The brain must adapt in some way to accommodate new functions, memories and skill. The adaptation of individual neural networks occurs through synaptic modification according to mechanisms on the scale of single neurons. These modifications are theorized to be the cellular basis of learning and long-term memory and may be required for the creation of neurocomputers. We must thus allow network plasticity in order to examine the role of learning and generally investigate the mutual interaction between oscillator dynamics and network dynamics.

First, we relax the condition that the coupling strengths between all pairs of oscillators be equal. To this end, define K_{ij} to be the connection strength from oscillator i to oscillator j; in general this matrix will not be symmetric (especially with the inclusion of time delays). Then following [61], a version of the Hebbian learning rule, that increases connection strengths between those oscillators whose phases are close in value, is implemented. Thus, a second set of differential equations is introduced

$$\dot{K}_{ij}(t) = \epsilon\{\gamma \cos(\phi_i(t) - \phi_j(t - \tau_{ij})) - K_{ij}\} \tag{3.75}$$

Here, ϵ is a constant that allows us to arbitrarily adjust how "dynamic" the coupling is, i.e. how fast changes in coupling strengths can occur and the cosine function forces the coupling strengths to increase when the difference in phase approaches zero and decrease when it approaches π(exactly out of phase). Additionally, the parameter γ in Eq. (3.75) provides the fixed-point value, that K_{ij} would approach if the ij pair was perfectly phase matched. Thus, conceptually γ and $-\gamma$ are the maximum and minimum values that the coupling is allowed to attain respectively.

Remark. Note that we are using the expression "Hebbian learning" somewhat liberally here; this is not the traditional rule but rather its Delta formulation (i.e. we are considering the time derivative of the coupling strength). Furthermore, we are making another simplification here by not restricting the coupling to positive or negative values. This means that the interaction between oscillator pairs can easily switch from excitatory to inhibitory.

3.5 Learning with oscillators

From what we have described in the previous section, we can infer that Kuramoto model can be used in order to store information. When many oscillators are coupled together, the whole system can evolve towards different synchronized states with specific phase relationships among them and these phase patterns can be used to store information. In fact, oscillatory associative memory models usually consist of interacting oscillators, where memorized patterns are stored as phase-locked oscillations.

Starting from a generic oscillatory neural network[51], Hoppensteadt and Izhikevich in [30] have proved how under some conditions, these networks can have interesting neurocomputational properties. In particular, they can act as multiple attractor neural networks, where attractors are limit cycle.

[51] An oscillatory neural network is an ensemble of neurons responsible for a wide variety of periodic behavior patterns. It can be seen as a generalization of the Kuramoto model, where the interaction function H_{ij} is not said to be sinusoidal and has the property of zero average; i.e.

$$\int_0^{2\pi} H_{ij}(\chi)d\chi = 0$$

Theorem 3.5.1 (Convergence Theorem for Oscillatory Neural Networks). *Consider the oscillatory neural network*

$$\dot{\phi}_i = \omega_i + \sum_{j=1}^{N} H_{ij}(\phi_j - \phi_i), \qquad \phi_i \in \mathbb{S}^1 \tag{3.76}$$

and suppose that $\omega_1 = \cdots = \omega_N = \omega$ and

$$H_{ij}(-\chi) = -H_{ji}(\chi), \qquad \chi \in \mathbb{S}^1 \tag{3.77}$$

for all $i, j = 1, \ldots, N$. Then the network dynamics converges to a limit cycle attractor. On the limit cycle, all neurons oscillate with equal frequencies and constant phase deviations. This corresponds to synchronization of the network activity.

Proof. Let $\phi = \omega t + \varphi$. Then

$$\dot{\varphi}_i = \sum_{j=1}^{N} H_{ij}(\varphi_j - \varphi_i), \qquad \varphi_i \in \mathbb{S}^1 \tag{3.78}$$

If this system always converges to an equilibrium, say φ^*, then Eq. (3.76) converges to a limit cycle $\phi(\tau) = \omega\tau + \varphi^*$. Obviously, phase deviations (vector $\varphi^* \in \mathbb{T}^N$) are constant on the limit cycle.
Let

$$R_{ij}(\chi) = \int_0^{\chi} H_{ij}(t)dt$$

be the antiderivative of $H_{ij}(\chi)$; that is $dR_{ij}/d\chi = H_{ij}$. Due to the property that characterizes H_{ij} we have

$$\int_0^{2\pi} H_{ij}(\chi)d\chi = 0 \qquad \Longrightarrow \qquad R_{ij}(2\pi) = 0$$

and therefore $R_{ij}(\chi)$ is continuous. Let us check that the function $E : \mathbb{T}^N \to \mathbb{R}$ given by

$$E(\varphi) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} R_{ij}(\varphi_j - \varphi_i) \tag{3.79}$$

is a global Lyapunov function for (3.76). Indeed, since it is continuous and $\mathbb{T}^N$ is compact, E is bounded below. Moreover, it satisfies

$$\frac{\partial E}{\partial \varphi_l} = \frac{1}{2}\Big(\sum_{i=1}^{N} H_{il}(\varphi_l - \varphi_j) - \sum_{j=1}^{N} H_{lj}(\varphi_j - \varphi_l)\Big) = -\sum_{j=1}^{N} H_{lj}(\varphi_j - \varphi_l) = -\dot{\varphi}_l$$

and hence

$$\frac{dE(\varphi)}{d\tau} = \sum_{l=1}^{N} \frac{\partial E}{\partial \varphi_l}\dot{\varphi}_l = -\sum_{l=1}^{N} |\dot{\varphi}_l|^2 \leq 0$$

Notice that $\frac{dE}{d\tau} = 0$ precisely when $\dot{\varphi}_1 = \cdots = \dot{\varphi}_N = 0$, i.e. at an equilibrium point of (3.76). □

There could be many equilibria of (3.76) depending on the connection functions H_{ij}. Therefore, the network can remember and reproduce many previously memorized oscillatory patterns. We now study the problem connected with Kuramoto model, where we consider specific H_{ij}, namely $H_{ij}(\phi_j - \phi_i) = (1/N)K_{ij}\sin(\phi_j + \eta_{ij} - \phi_i)$ for some constants K_{ij} and η_{ij}.

3.5.1 Application to the Kuramoto model

When all the connection functions in the neural network are equal, namely, $H_{ij}(\chi) = H_{ji}(\chi) = H(\chi)$ for some function $H(\chi)$, then condition (3.77) implies that $H(\chi)$ is an odd function. Any odd function on $\mathbb{S}^1$ can be represented as a Fourier series

$$H(\chi) = \sum_{l=1}^{+\infty} a_l \sin(l\chi)$$

Kuramoto suggested that the first term, $a_1 \sin(\chi)$, is dominant, and he disregarded the rest of the series. Thus, as just presented in Eq. (3.15), the Kuramoto's phase model is

$$\dot{\phi}_i = \omega_i + \sum_{j=1}^{N} \frac{K_{ij}}{N} \sin(\phi_j - \phi_i)$$

where $K_{ij} \in \mathbb{R}$ are the synaptic coefficients. The convergence theorem 3.5.1 is applicable to Kuramoto's model when $\omega_1 = \cdots = \omega_N$ and the synaptic matrix $K = (K_{ij})$ is symmetric; that is, $K_{ij} = K_{ji}$ for all i and j. Indeed, condition (3.77) has to be satisfied not only by K but by the entire function H described above. Thus, since the *sine* function is an odd function, this implies that in order to satisfy that condition the matrix K need to be symmetric.

Since, as discussed in section 3.4.1, to have a better model we have to take into account time delays, one can adjust Kuramoto's model to include the natural phase differences η_{ij}, so that the model is

$$\dot{\phi}_i = \omega_i + \sum_{j=1}^{N} \frac{K_{ij}}{N} \sin(\phi_j + \eta_{ij} - \phi_i), \qquad \phi_i \in \mathbb{S}^1 \tag{3.80}$$

where each connection is encoded by a pair of numbers: the strength of synapse K_{ij} and its phase η_{ij}. It is convenient to represent such a synapse by a complex number $c_{ij} = K_{ij}e^{i\eta_{ij}}$. Theorem 3.5.1 imposes the additional requirement that $\eta_{ij} = -\eta_{ji}$, which means that the complex-valued synaptic matrix $C = (c_{ij})$, is self-adjoint; i.e., $c_{ij} = \bar{c}_{ji}$, where $\bar{c}$ is the complex conjugate of c.

Following Hoppensteadt and Izhikevich, [30], using the representation $c_{ij} = K_{ij}e^{i\eta_{ij}}$ they derive frOm Hebb's assumptions the synaptic modification rule presented above (3.75)

$$\begin{aligned}\dot{K}_{ij} &= \epsilon\{\mu\cos(\phi_i - \phi_j - \eta_{ij}) - K_{ij}\}\\ \dot{\eta}_{ij} &= \frac{\mu\epsilon}{K_{ij}}\sin(\phi_i - \phi_j - \eta_{ij})\end{aligned}$$

If, during the learning period, the i-th and j-th oscillators are synchronized, then c_{ij} memorizes the phase difference. That is, if $\phi_i(\tau) - \phi_j(\tau) = \chi^*$, then $\eta_{ij}(\tau) \to \chi^*$ (memorization of phase information) and $K_{ij}(\tau) \to \mu\epsilon/\epsilon = \mu$ (increase of synaptic strength). On the other hand, if the oscillators are incoherent, the $\sin(\phi_i(\tau) - \phi_j(\tau) - \eta_{ij})$ and $\cos(\phi_i(\tau) - \phi_j(\tau) - \eta_{ij})$ may average to zero. In this case, η_{ij} is relatively unaffected (persistence of phase memory), but $K_{ij} \to 0$ (synaptic fading). The learning rule then establishes the relation $c_{ij} = \bar{c}_{ji}$, so that the convergence theorem 3.5.1 applies.

Therefore, the property we have described is what allows the Kuramoto model to be used as an associative memory. The result given in Theorem 3.5.1 is very important because it implies that an energy function could be found and under some conditions the network will always converge to an oscillatory phase-locked pattern. The key difference between the Hopfield and the oscillatory network is that memorized images correspond to equilibrium points (attractors) in the former and to limit cycle attractors in the latter. In general such memories have a lower capacity compared to the Hopfield model, but if we consider a simple modification of the classical Kuramoto model by adding a second-order Fourier mode,

$$H_{ij}(\phi_i - \phi_j) = K_{ij}\sin(\phi_i - \phi_j) + \frac{\bar{\epsilon}}{N}\sin(2(\phi_i - \phi_j)) \tag{3.81}$$

oscillators associative memories based on phase-locking with a Hebbian connection scheme are capable of performing as well as the Hopfield network, as it has been shown by Nishikawa et al. in [42].

3.6 From Hopfield to Kuramoto model

The question that may arise now is how the models of Hopfield and Kuramoto are linked together. From the previous discussion, we have understand how these two models can be both used for learning patterns. Thus, we need now to highlight how we can pass from the context of spikes and action potential due to Hopfield to that characterized by phases in Kuramoto.

3.6.1 Integrate and fire model

We have seen in chapter 2 that the neuronal dynamics can be conceived as a summation process (sometimes also called "integration" process) combined with a mechanism that triggers action potentials above some critical voltage. Indeed in experiments firing times are often defined as the moment when the membrane potential reaches some threshold value from below. In order to build a phenomenological model of neuronal dynamics, we describe the critical voltage for spike initiation by a formal threshold U. If the voltage $u_i(t)$ (that contains the summed effect of all inputs) reaches U from below, we say that neuron i fires a spike.
Such type of model makes use of the fact that neuronal action potentials for a given neuron always have roughly the same form. If the shape of an action potential is always the same, then the shape cannot be used to transmit information: rather information is contained in the presence or absence of a spike. Therefore action potentials are reduced to "events" that happen at a precise moment in time.
Neuron models where action potentials are described as events are called *integrate-and- fire models* (IFM). They have two separate components that are both necessary to define their dynamics: first an equation that describes the evolution of the membrane potential $u_i(t)$ and second a mechanism to generate spikes.
Following [25] and using the same reasoning of section 2.3.1, we now introduce the simplest form for integrate-and-fire models. Thus the action potential of neuron i, u_i, obeys the following equation

$$\frac{du_i}{dt} = -\frac{u_i}{\tau} + I_i + I_i^{ext} \tag{3.82}$$

The first term on the r.h.s. describes relaxation of u_i towards the resting potential with a relaxation time τ. In most models, u_i is normalized so that it acquires its maximum value at 1, that means

$$-1 \leq u_i \leq 1 \tag{3.83}$$

Then I_i is defined as the input from other neurons, i.e. the sum of the postsynaptic potentials (PSPs) that are triggered by pulses of afferent neurons at times t'. I_i^{ext} represents external

Table 3.1: Correspondence between the Integrate and Fire model and the Hopfield model

IFM	Hopfield model
u_i	u_i
τ	CR
I_i^{ext}	$(1/C)I_i$
I_i	$(1/C)\sum_j T_{ij}V_j$

input stemming from sensory neurons.
It is now immediate to understand how Hopfield model represents a generalization of the IFM with the additional assumption of a non-linear sigmoidal output $V_i = g(u_i)$. Therefore starting from equation (3.82) and comparing with the same equation for the continuous Hopfield model derived before (2.69), we make the correspondence between the two models, that is summarized in Table 3.1.

3.6.2 Lighthouse model

As we have presented before, a neuron receives inputs from other neurons in the form of their axonal spikes. At synapses these spikes are converted into dendritic currents that lead to a potential change at the axon hillock. In response to incoming signals, the neuron produces an output in the form of axonal pulses (spikes). The occurrence of (more or less) regular spikes in an axon can be considered as a periodic event. Periodic or at least rhythmic events can also be observed in electric and magnetic fields of the brain, as measured by EEG and MEG, respectively. Thus, in order to quantitatively deal with rhythms, the concept of phase is indispensable.
A possible mathematical translation of the dynamics just presented is the one proposed by Haken in [24] with the model of spikes. Following are briefly outlined the guidelines that lead to this model, which, even in the most basic version discussed here, give an anticipation of how it is possible to read the neuronal dynamics according to the phase.
To mathematically model action potentials, we use an idealization that leads to the Dirac function. Consider, in fact, the impulse modeled as a curve with a Gaussian form:

$$\frac{1}{\sqrt{\pi\sigma}}e^{-\frac{t^2}{\sigma}}$$

where the factor in front of the exponential serves for normalization. Suppose we want to temporally shorten the impulse so as to make it biologically reliable. Analytically this is achievable by sending $\sigma \to 0$: we obtain that the domain of existence of the impulse is so short that for $t \neq 0$ the spike vanishes, while for $t = 0$ it remains normalized:

$$\delta(t) = 0, \quad \text{for } t \neq 0 \tag{3.84}$$

$$\int_{-\epsilon}^{\epsilon} \delta(t)dt = 1 \tag{3.85}$$

where ϵ is arbitrarily small. Note how, in fact, this is precisely the definition of δ Dirac function, that is explicitly a time dependent function, denoted by $\phi(t)$, having the meaning of phase in the sense that we have described in section 3.2. A special case of δ function, which leads to the definition of "spiking function" used later, is the following. Consider

$$\int_{t_1}^{T} \delta(\phi(t))dt \tag{3.86}$$

subject to the following work hypothesis:

1. for $t_1 \leq t \leq T$ the unique zero of $\phi(t)$ is at $t = t_0$
2. $\frac{d\phi(t_0)}{dt} \equiv \dot{\phi}(t_0) \neq 0$
3. $\dot{\phi}(t_0) > 0$

The change of variable $s = \phi(t)$ and the assumption $\phi(t_1) < \phi(T)$ allow to obtain

$$\int_{t_1}^{T} \delta(\phi(t))dt = \begin{cases} 0 & \text{for } T < t_0 \\ \frac{1}{2\dot{\phi}(t_0)} & \text{for } T = t_0 \\ \frac{1}{\dot{\phi}(t_0)} & \text{for } T > t_0 \end{cases} \tag{3.87}$$

However, it is known that the occurrence of the factor $\dot{\phi}$ in Eq. (3.87) causes the function δ to depend on ϕ in a difficult way to treat. To overcome this difficulty it is customary to define the following function, called precisely *spiking function*

$$f(t) = \delta(\phi(t))\dot{\phi}(t) \tag{3.88}$$

which, replaced with the δ starting function, offers

$$\int_{t_0-\epsilon}^{t_0+\epsilon} f(t)dt = \int_{t_0-\epsilon}^{t_0+\epsilon} \delta(\phi(t))\dot{\phi}(t)dt = 1 \tag{3.89}$$

With the introduction of ϕ it is possible to think of the function δ, that shapes impulses, so as to have a peak in t_n

$$\phi(t_n) = 2\pi n$$

Therefore in this case a "peaked function" is normally used

$$\delta(\phi(t) - 2\pi n)\dot{\phi}(t) \tag{3.90}$$

that will be normalized in order to make the modeling of the spike more manageable and explicitly dependent on ϕ. To now represent all the spikes over time we proceed to add the various peaks, obtaining

$$Q(t) \equiv f(\phi(t)) = \sum_n \delta(\phi(t) - 2\pi n)\dot{\phi}(t_n) \tag{3.91}$$

where n ideally ranges between $-\infty$ and ∞. At this point the essential question to answer becomes how it is possible to determine mathematically the dependence of ϕ, and therefore by extension of the spikes, from the time t. The answer comes, for example, from the Naka-Rushton formula, which provides a quantitative link between the intensity of the Q-stimulus acting at the pulse-generating site and the pulse-producing rate. The formula offers

$$S(X) = \frac{rX^M}{\Theta^M + X^M} \tag{3.92}$$

where M is an approximate measure of the steepness of the curve representing the stimulus and Θ is the value for which half of the maximum rate is obtained. Absorbing any proportional factors by redefining the constant r, we arrive at the formulation of the following differential equation for the phase ϕ_i connected with neuron i

$$\dot{\phi}_i(t) = S(X_i) \tag{3.93}$$

in which the vector field represents the set of external and internal stimuli, due to interaction with other neurons and external patterns. For a complete understanding of the equation we have to determine the input X_i which leads to the spike generation. As we know, spikes are generated at the axon hillock at which the potentials due to the dendritic currents are added. We may take time delays τ' into account as well as coefficients c that convert ψ into potential contributions. Also external signals, $p_{ext}(t)$, stemming from sensory neurons must be taken into account. Thus we arrive at

$$X_i(t) = \sum_m c_{i,m}\psi_m(t-\tau') + p_{ext,i}(t-\tau'') \tag{3.94}$$

The explicit resolution of the dynamics between two interacting neurons follows, finally, from the temporal evolution of ψ: this current is caused by a spike generated in the axon of another neuron (denoted with i) at a previous time τ. The impulse can be represented by the function $Q_i(t-\tau)$, which in the simplest of cases is a Delta Dirac function, as seen in Eq. (3.91): after its generation the current ψ decays with a decay constant γ that it is supposed to be independent of m. Therefore we obtain the following dendritic equation

$$\dot{\psi}_m(t) = aQ_i(t-\tau) - \gamma\psi_m(t) \tag{3.95}$$

In conclusion, the system of equation that governs the interaction between two neurons in terms of their phase is usually given by

$$\begin{cases} \dot{\phi}_i(t) = \sum_m c_{im}\psi_m(t) + p_{ext,i}(t) \\ \dot{\psi}_m(t) = aQ_i(t-\tau) - \gamma\psi_m(t) \end{cases} \tag{3.96}$$

where we have assumed that $\tau = \tau' = 0$ and that S is a linear function, so that it can be replaced with its argument.

3.6.3 From IFM to LM

A comparison between Eq. (3.82) and the equation of the phases as derived in (3.96) reveals that these equations have the same structure except for the first term on the r.h.s of (3.82). Actually, the analogy between the models is quite close. A minor difference is this: while u_i adopts its values according to (3.83), ϕ_i runs in its first segment from 0 to 2π so that we arrive at the relationship

$$\phi_i = 2\pi\Big(\frac{u_i+1}{2}\Big) = \pi(u_i+1) \tag{3.97}$$

that connects the phase of the axonal pulses with the action potential u of the corresponding neuron. But in contrast to the integrate and fire models with (3.82), in the lighthouse model ϕ_i may increase indefinitely. To overcome this problem, we generalize the lighthouse model by taking into account a damping of the rotation speed in between two firings. This means we want to mimic the effect of the damping term that occurs on the r.h.s. of (3.82). We have to take into account the fact that u is restricted to the region between -1 and 1, or correspondingly that the relaxation dynamics of ϕ in the intervals $[n2\pi, (n+1)2\pi]$ must be self-similar. In other words, we must reduce ϕ after each rotation. This is achieved by replacing ϕ with $\phi mod 2\pi$. Therefore, the connection between the phase angle ϕ and the action potential u is thus given by

$$\phi(t) mod 2\pi = \pi(u(t) + 1) \tag{3.98}$$

The two sets of equations (3.82) and (3.96) can then be incorporated into forming a unique equation, as stated in [24].
Denote the dendritic current of dendrite m by ψ_m and the axonal pulse of axon i by Q_i. The equations for the dendritic currents read

$$\left(\frac{d}{dt} + \gamma\right)^{\beta} \psi_m(t) = \sum_i a_{mi} Q_i(t) \tag{3.99}$$

where $\beta = 1$ refers to the lighthouse model and $\beta = 2$ to the integrate and fire model. The constants and quantities in this equation have the following meaning: γ is the damping constant, a_{mi} represents the coupling coefficient (synaptic strength).
Then, if we take into consideration the phase ϕ we deduce the following equation

$$\dot{\phi}_i(t) + \gamma' \phi_i(t) mod 2\pi = S\Big(\sum_m c_{im} \psi_m(t) + p_{ext,i}(t), \Theta_i\Big) \tag{3.100}$$

where c_{im} are the coupling constants, $p_{ext,i}$ the external signal and Θ_i the threshold. Depending on $\gamma' = 0$ or $\gamma' = 1$ we are dealing with the lighthouse model in the first case and with the integrate and fire model in the second case. The sigmoidal function S in Eq. (3.100) is well established by physiological experiments, e.g. from the Naka-Rushton formula (3.92), or, in a rather good approximation, in the following form

$$S(X, \Theta) = \begin{cases} 0 & \text{for } X < X_{min} \\ X & \text{for } X_{min} \leq X \leq X_{max} \\ S_{max} & \text{for } X \geq X_{max} \end{cases} \tag{3.101}$$

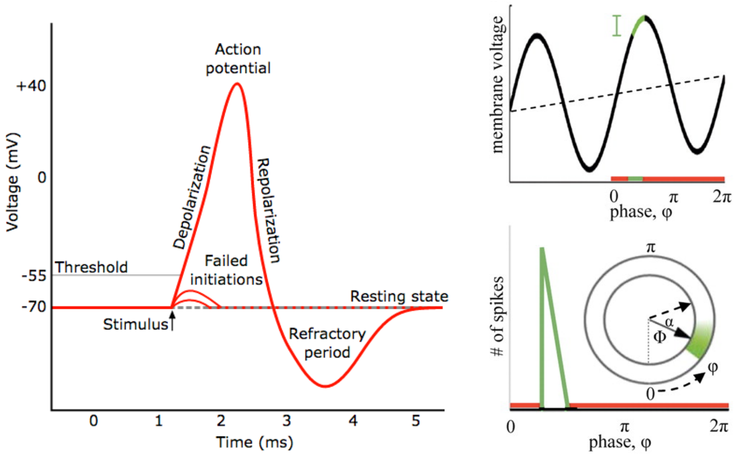
Now it is easy to understand also in this case how we can connect the lighthouse model to that of Hopfield, as summarized in table 3.2. Note that in Hopfield model we do not have the presence of the function S, but from Eq. (3.101) we have that if the argument is inside a certain interval (and this is always the case for us) we can replace the function S with it. We can think at X_{max} as the threshold value U and X_{min} the minimum value to start the depolarization process.

3.6.4 Neurons as oscillators

The last step now consists in reporting the description of Haken within the context of Kuramoto, characterized by oscillators coupled with a sinusoidal function. We have moved from a description in terms of action potentials in Hopfield to one in the phase space with the lighthouse model of Haken. We now want to create a link within the context of Kuramoto. The model of Kuramoto, as we have seen in 3.2, involves N identical oscillators interacting with each other, therefore the possibility to apply profitably the theory exposed to the dynamic system represented by the brain is based on the ability to bring the single neurons, or moreover agglomerates of them, back to oscillating elements and to interpret every element of the Eq. (3.15) in a biological key.

Table 3.2: Correspondence between the Lighthouse model and the Hopfield model.

LM	Hopfield model
ϕ_i	$\pi(u_i + 1)$
ψ	V
$p_{ext,i}$	I_i^{ext}
Θ_i	U_i
c_{im}	T_{im}



(a) Diagram time-voltage **(b)** Relation phase-spike

Figure 3.7: Figure (a) shows the different phases of the action potential. The time course of the spike is shown with depolarization and repolarization phases, refractory period and resting state labeled. There are also indicated the levels of the membrane potential and in particular the value at rest. Note that the transmembrane voltage is indicated on the ordinate and the time is indicated on the abscissa. Figure (b) represents the relation between phase and membrane voltage. In the figure above the diagram shows the action potential, highlighting the value of the associated phase. In the figure below, it is stressed the correspondence between the emission of a spike and phase equal to more or less 2π or its integer multiple.

- *Neuron-Oscillators*: from neuronal functioning it is understandable how the occurrence of more or less regular spikes can be considered as a rhythmic event. As we have discussed for the lighthouse model, in order to quantitatively and qualitatively treat this rhythm, the concept of phase $\phi(t)$ is indispensable. The rotation period T of the neuron-oscillator can be interpreted as the time interval that elapses between the instantaneous and regular discharges of two consecutive spikes, modeled as functions: whenever the phase is equal to 2π, or to an integer multiple, one spike is generated. In fact, at each value of the phase ϕ corresponds both the temporal description of the spike and, above all, a specific electrical state of the neuronal membrane. Considering that the trend of the electrical potential is not symmetrical, but follows the curve in Fig. 3.7, and that the process is instantaneous, we have that for $\phi = 2\pi$ the voltage of the membrane is such that the neuron is in the phase of depolarization, for $\phi = 2\pi + \Delta$ with small Δ it reaches the phase of hyperpolarization. The intermediate values of ϕ correspond to the decrease of the voltage inversion $\phi \in [2\pi, 2\pi + \Delta]$ due to the opening

of specific ionic channels by the sodium-potassium pump and to the stable return to the resting value and its maintenance until $\phi = 4\pi$.
The frequency ω of the neuron-oscillator corresponds, as can be guessed, to the production rate of the spikes themselves: the greater the speed of rotation, and therefore by extension the frequency, the greater the production of nerve impulses in the same period of time. This frequency is variable and is influenced by the level of electrical excitation of the nerve fiber following the reception of different external patterns.

- *Interaction between oscillators*: more delicate to treat is the question of the coupling force K, which could in fact refer both to the number of connections between neurons, i.e. trivially to the cardinality of synaptic strands, and to the electrical strength of such connections, that is the conductivity of the nerve fiber based on external inputs. This last possible interpretation, in particular, logically derives from the same mechanism of phase transition: from the direct link between the increase of K in relation to K_c and the increasing level of synchronization between the neurons-oscillators we can suppose the correspondence of K with the degree of electrical conductivity of the neuron and therefore with a greater probability of propagation. One of the most complete formulation of the coupling strength on a physiological level is that proposed by Grannan et al. (1993): here the actual coupling between the neurons-oscillators is given by $K_{ij}(r, r_0) = L(r)W(r, r_0)$ where L denotes the average level of pre- and post-synaptic cell activity, related to the electrical properties and permeability of the membrane, and W the specific architecture of the connections, whose geometry influences the propagation of the spikes in terms of longitudinal resistance and distance. Then for the interaction function H_{ij} it has been proved in section 3.5 that the *sine* function is a good approximation of the neural interaction and a function that guarantees the validity of theorem 3.5.1. Therefore, thanks to Malkin's theorem 3.3.3 we know that if we include a perturbation in our model, its properties are still valid and thus the stability is preserved.

- *Synchronization level*: in light of what we have briefly explained on the neuronal synchronization mechanism in section 3.2 it is possible to understand the biological meaning to be attributed to the incoherent state and to the synchronization state of the theoretical model. Considering the interaction between neurons stimulated by the same pattern, i.e. related to a common brain function, the *incoherent state*, characterized by the absence of synchrony, is substantially expressed in two cases. The first is the trivial one in which the interested nerve fibers are not electrically excited by external stimuli, i.e. when there is no transmission of information: in this case, in fact, the oscillators do not rotate because they cannot generate spikes and their phases ϕ remain fixed on the angular values corresponding to a stable condition of resting potential. The second occurs when the neurons emit electrical discharges, and therefore oscillate in the sense described above, with an intrinsic frequency such that the generated chains are in a non-negligible phase delay ($\approx 180^\circ \rightarrow$ anti-phase), probably due to a failure of correct electrical functioning of the excitation process of the single neurons or of a significant delay in the transmission of the signal.
On the other hand, there is a *synchronization state* when a configuration of frequencies is implemented such that the electrical impulses sent by each starting neuron i arrive at the target neuron j when its phase ϕ_j is close to 2π. Bearing in mind the correspondence between the phase and the electrical potential presented above, this means that the voltage of the membrane of each neuron i reaches the state of maximum depolarization

only a fraction of a second earlier than neuron j: consequently the spike discharged by each neuron i reaches neuron j exactly when the latter is in its state of maximum depolarization so as to increase the chances of triggering an action potential in neuron j. In fact, as K increases and therefore the signal conduction capacity, this situation occurs when the frequencies of the start and arrival neurons are out of phase by an arbitrary small factor $\tilde{\epsilon}$, which takes into account the delay due to the finite propagation speed in an electrical resistance environment.

3.6.5 From Hopfield to Kuramoto model

From section 3.6.3, we have understood how neurons can be treated as oscillators, since we can describe the spike process with a phase model. Following the contributions[52] of Hoppensteadt and Izhikevich in [30] and in [35], we will prove that Hopfield continuous model (2.69) can be transformed into Kuramoto model (3.15) using the assumptions of *weakly connections* and *quasi-periodic oscillators.*

- Brain units, i.e. neurons, are *weakly connected.* A reasonable way to characterize weakness of synaptic connections between single neurons is to consider amplitudes of postsynaptic potentials (PSPs). The PSP amplitudes may differ substantially along dendrites due to the electronic attenuation; see the illustration in Figure 3.8. A somatic PSP is the weakest, but it best characterizes the magnitude of the postsynaptic neuron response because the soma is near the axon hillock, that is the place of initiation of the action potential. In fact, the average soma PSP is smaller than 1 mV, while the action potential emanating from the hillock region is approximately 100 mV in amplitude. This assumption is thus saying that there must be many (a few hundred) presynaptic neurons firing simultaneously to make a given cell fire.

  Now we can add this hypothesis into Eq. (2.69) by multiplying the first term on the right hand side by a certain ε, small enough, that describes the (dimensionless)

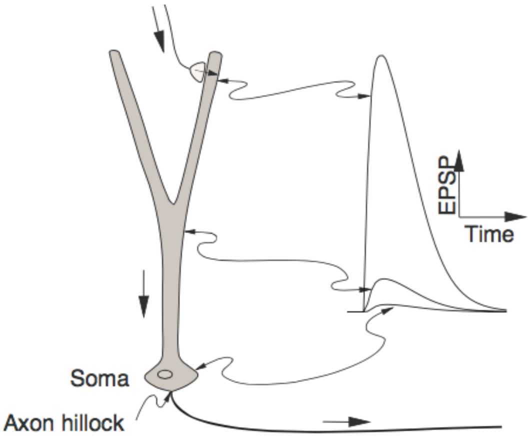


Figure 3.8: EPSP(Exitatory Post-Synaptic Potential) in three different locations. The EPSP size at the soma is much smaller than the one in the vicinity of the synapse due to electrotonic attenuation. Figure taken from [35].

[52] In the first work [30] of Hoppensteadt and Izhikevich they have discussed the case of weakly connected neural networks for limit cycle oscillators, whereas Izhikevich have then generalized it for quasi-periodic oscillators in [35].

strength of synaptic connections:

$$\frac{du_i}{dt} = \varepsilon \sum_j T_{ij} V_j - u_i + I_i = F_i(u_i) + \varepsilon G_i(u_1, \ldots, u_N, \varepsilon), \qquad \varepsilon \ll 1 \tag{3.102}$$

where we assume $C_i = 1$ and $R_i = 1$ for all $i = 1, \ldots, N$ and we define the function F and G as

$$\begin{aligned} F_i(u_i) &= -u_i + I_i, & i = 1, \ldots, N \\ G_i(u_1, \ldots, u_N, \varepsilon) &= \sum_j T_{ij} V_j, & i = 1, \ldots, N \end{aligned}$$

- The second assumption we made is that the rhythmic signal that characterizes the emission of spikes is *quasi-periodic*. EEG recordings of brain rhythmic activity rarely show periodic oscillations, even during epileptic seizures. Fourier and power spectrum analysis reveal that rhythmic activity of a local field potential is "random" with a few pronounced frequencies. The most prominent are gamma ($30 - 100$ Hz) oscillations in the cortex and theta ($4 - 8$ Hz) oscillations in the hippocampus. Thus, the next natural step in modeling the brain using the oscillatory network approach is to assume that each oscillator can exhibit multifrequency oscillations, which in the simplest case may be just a collection of periodic oscillators with different frequencies (as can be seen in Fig. 3.9).

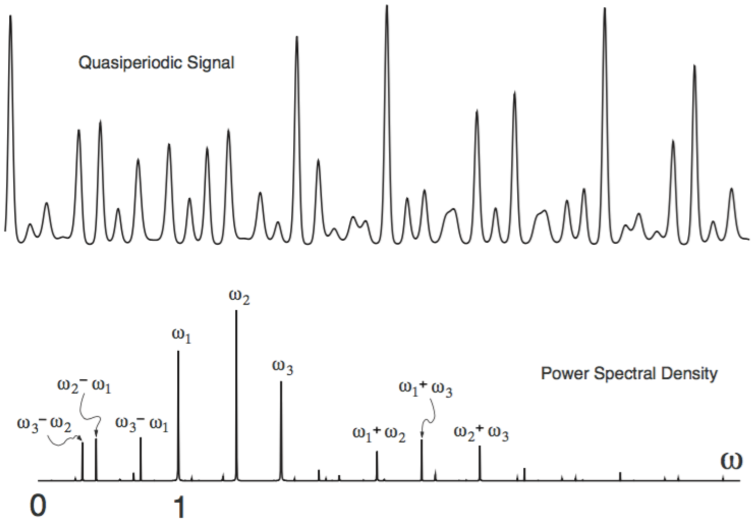


Figure 3.9: Top: An example of a quasi-periodic signal $X(t)$ with three different frequencies ω_1, ω_2 and ω_3. Bottom: its discrete power spectrum. Figure taken from [35].

When the number of incommensurable frequencies is finite, the oscillator is said to be *quasi periodic*. The signals might look chaotic or noisy, but they are not. Their power spectra are discrete with peaks corresponding to the composed frequencies and their linear combinations, as illustrated in Fig. 3.9. Let now define this concept mathematically following Izhikevich in [35].

Definition 3.6.1 (Quasi-periodic signal)**.** *A continuous rhythmic signal* $u(t)$ *is* quasi-periodic *if there is a continuous function* $q(\theta_1, \ldots, \theta_N)$*, which is* 2π*-periodic in each* *argument, such that*

$$u(t) = q(\Omega_1 t, \dots, \Omega_N t) \qquad \text{for all } t \geq 0 \tag{3.103}$$

where $\Omega = (\Omega_1, \dots, \Omega_N)^T \in \mathbb{R}^N$ *is a frequency vector.*

Let now $\mathbb{T}^N$ denote the N-torus and let $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{T}^N$ be the phase variable on the torus. Then we may rewrite the equation above in the form

$$u(t) = q(\theta(t)), \qquad \dot{\theta} = \Omega \tag{3.104}$$

which is a convenient way to represent quasi-periodic oscillators.

Therefore, Hopfield model can be seen as a weakly connected neural network that describes the dynamics of a quasi-periodic signal $u(t)$.

Definition 3.6.2. *We introduce the following notions, which we will need later on, as presented in [30], pp. 126-128.*
Consider a dynamical system $\dot{u} = F(u)$ *with* $u \in \mathbb{R}^N$.

1. *A set* $\mathcal{D} \subset \mathbb{R}^N$ *is said to be an invariant set of a dynamical system if each trajectory starting in* $\mathcal{D}$ *will remain in* $\mathcal{D}$ *for every* $t > 0$*, i.e.* $\mathcal{D}$ *is invariant if*

$$u(0) \in \mathcal{D} \quad \Longrightarrow \quad u(t) \in \mathcal{D} \text{ for all } t > 0$$

We will consider the case in which $\mathcal{D}$ *is an invariant manifold, i.e.* $\mathcal{D}$ *is not only a set but also a differentiable manifold.*
For example, a single equilibrium point is an invariant manifold: it is clearly invariant and is a trivial (zero-dimensional) manifold.

2. *Let* $\mathcal{D}$ *be a manifold and consider its tangent space* $T_u\mathcal{D}$ *and the normal space* $N_u\mathcal{D}$. *Let*

$$\Pi : T_u\mathcal{D} \to N_u\mathcal{D}$$

be the orthogonal projection to the normal space $N_u\mathcal{D}$. *We define the contractions of vectors at* $\mathcal{D}$ *as*

$$\begin{array}{rcll} v(t) & = & D_u\Phi(u,t)v(0), & v(0) \in T_u\mathcal{D} \\ w(t) & = & \Pi D_u\Phi(u,t)w(0), & w(0) \in N_u\mathcal{D} \end{array}$$

where D_u *is the differential operator and* $\Phi(u,t)$ *the flux associated to the dynamical system.*

3. *An invariant manifold* $\mathcal{D}$ *is normally hyperbolic when*

$$\lim_{t\to\infty} |w(t)| = 0 \qquad \text{exponential stability}$$

and

$$\lim_{t\to\infty} \frac{|w(t)|}{|v(t)|} = 0 \qquad \text{normal hyperbolicity}$$

for all $u \in \mathcal{D}$ *and all nonzero vectors* $w \in N_u\mathcal{D}$ *and* $v \in T_u\mathcal{D}$.

Remark. When the manifold is an equilibrium, the definition of normally hyperbolic coincides with the usual definition of hyperbolicity: the Jacobian matrix does not have eigenvalues with zero real part. So for our uncoupled system

$$\dot{u}_i = F_i(u_i) = -u_i + I_i$$

with the initial condition that the action potential at $t = 0$ is equal at the value at rest $u_0 = u_{rest}$, the solution is known to be

$$u_i(t) = e^{-t}(u_0 - I_i) + I_i$$

that when $t \to \infty$, it tends to $\mathcal{D}_i = I_i$, which represents an equilibrium point for our dynamical system (3.106). The Jacobian matrix is then represented by $J(u_i) = -1$ and thus the property is satisfied.

Following Izhikevich in [35], p. 2198, we now introduce the following theorem, that will lead us to transform such system in a phase model. Furthermore, we implicitly assume that all manifolds and functions here are as smooth as necessary for our manipulations.

Theorem 3.6.3 (Phase model for weakly connected quasi-periodic oscillators)**.** *Consider a weakly connected system of the form*

$$\dot{u}_i = F_i(u_i) + \varepsilon G_i(u_1, \dots, u_N, \varepsilon), \qquad u_i \in \mathbb{R},\ i = 1, \dots, N, \quad \varepsilon \ll 1 \tag{3.105}$$

such that each uncoupled system

$$\dot{u}_i = F_i(u_i) \tag{3.106}$$

has an exponentially stable quasi-periodic equilibrium $\mathcal{D}_i$. Then there is an $\varepsilon_0 > 0$ such that for all $\varepsilon \leq \varepsilon_0$ there is an open neighborhood $\mathcal{W}$ of $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_N$ and a continuous mapping $\mathcal{S} : \mathcal{W} \to \mathbb{T}^N$ that transforms all local solutions of (3.105) to those of

$$\dot{\theta}_i = \Omega_i + \varepsilon \mathcal{G}_i(\theta_1, \dots, \theta_N, \varepsilon), \qquad \theta_i \in \mathbb{T},\ i = 1, \dots, N \tag{3.107}$$

where each $\Omega_i \in \mathbb{R}$ is a frequency vector with $\Omega_i \neq 0$ and $\mathcal{G}_i$ is some function. Therefore, (3.107) is a local system for (3.105).

Proof. See Appendix F. □

We now made the assumption that the connection function $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_N)$ has a special form; namely we suppose that

$$\mathcal{G}_i(\theta, \varepsilon) = \sum_{j=1}^{N} \mathcal{G}_{ij}(\theta_i, \theta_j) + \mathcal{O}(\varepsilon) \qquad \text{for all } i = 1 \dots, N \tag{3.108}$$

In analogy with Hopfield model it seems to be reasonable to have this pairwise coupled form. In order to do not have additional constants and to simplify the discussion, we then require that $\mathcal{G}_{ij}$ has zero average. To gain this result we can start from the averaging theory developed by Hoppensteadt and Izhikevich in [30], pp. 262-273.
First of all, write $\theta(t)$ in the form

$$\theta = \Omega t + \phi$$

where $\phi = (\phi_1, \cdots, \phi_N)^T \in \mathbb{T}^N$ is a vector of phase deviations. Now, we highlight that $\mathcal{G}_{ij}(\theta_i, \theta_j)$ is a 2π-periodic function in θ_i and θ_j, whereas $\mathcal{G}_{ij}(\Omega_i t + \phi_i, \Omega_j t + \phi_j)$ is a *quasi-periodic* function of t. It is not periodic unless the vector of frequencies Ω is proportional to

a vector of integers.
Thus following [30], the average of $\mathcal{G}_{ij}$ is defined as

$$\bar{\mathcal{G}}_{ij}(\phi_i,\phi_j) = \lim_{t\to\infty}\frac{1}{T}\int_0^T \mathcal{G}_{ij}(\Omega_i t+\phi_i,\Omega_j t+\phi_j)dt \tag{3.109}$$

From ergodic theory it follows that this integral converges to the constant

$$\mathcal{G}^0_{ij} = \frac{1}{(2\pi)^2}\int_0^{2\pi}\mathcal{G}_{ij}(s)ds \tag{3.110}$$

Thus, we can rewrite our model

$$\dot{\theta}_i = \Omega_i + \varepsilon\sum_{j=1}^N \mathcal{G}_{ij}(\theta_i,\theta_j) + \mathcal{O}(\varepsilon^2) \tag{3.111}$$

using the realtions

$$\tilde{\mathcal{G}}_{ij}(\theta_i,\theta_j) = \mathcal{G}_{ij}(\theta_i,\theta_j) - \mathcal{G}^0_{ij} \tag{3.112}$$

and

$$\omega_i = \sum_{j=1}^N \mathcal{G}^0_{ij} \tag{3.113}$$

Thus, $\tilde{\mathcal{G}}_{ij}$ has zero average, as we want, and we will use this function instead of the previous one $\mathcal{G}_{ij}$. Hence, Eq. (3.111) is rewritten as

$$\dot{\theta}_i = \Omega_i + \varepsilon\omega_i + \varepsilon\sum_{j=1}^N \tilde{\mathcal{G}}_{ij}(\theta_i,\theta_j) + \mathcal{O}(\varepsilon^2) \tag{3.114}$$

We now have the following Theorem ([30], pp. 275-276), that will transform Eq. (3.114) into an equation close to that of Kuramoto.

Theorem 3.6.4. *Consider a weakly connected quasi-periodic oscillatory system*

$$\dot{\theta}_i = \Omega_i + \varepsilon\omega_i + \varepsilon\sum_{j=1}^N \tilde{\mathcal{G}}_{ij}(\theta_i,\theta_j) + \mathcal{O}(\varepsilon^2), \quad \theta_i \in \mathbb{S}^1 \tag{3.115}$$

where the functions $\tilde{\mathcal{G}}_{ij}$ have zero average (the average is taken into account by ω_i). There is a change of variables of the form

$$\theta = \Omega t + \phi + \varepsilon g(\phi, t)$$

that transforms Eq. (3.115) to

$$\dot{\phi}_i = \varepsilon\omega_i + \varepsilon\sum_{j=1}^N s_{ij}(\phi_i,\phi_j) + \mathcal{O}(\varepsilon^2), \qquad \phi_i \in \mathbb{S}^1 \tag{3.116}$$

where each function s_{ij} satisfies

$$s_{ij}(\phi_i,\phi_j) = \begin{cases} 0 & \text{if } \Omega_i \not\sim \Omega_j, \\ H_{ij}(k_{ij}\phi_j - m_{ij}\phi_i) & \text{if } \Omega_i \sim \Omega_j, \end{cases} \tag{3.117}$$

where k_{ij} and m_{ij} are positive relatively prime integers for which $k_{ij}\Omega_j - m_{ij}\Omega_i = 0$ and the function H_{ij} has zero average, i.e.

$$\int_0^{2\pi} H_{ij}(\chi)d\chi = 0 \tag{3.118}$$

In particular, if $\Omega_i \not\sim \Omega_j$, then ϕ_i and ϕ_j do not depend on each other (up to order ε^2).

Proof. See [30], pp. 275-276. □

Furthermore, in Theorem 3.6.4 we have used an equivalence relation $\sim$, that in [30], p. 275, is defined as follows.

Definition 3.6.5 (Commensurable). *Nonzero real numbers a and b are said to be* commensurable*, denoted by $a \sim b$, when there are positive integers k and m such that*

$$ka - mb = 0$$

that is the vector $(a, b)^T \in \mathbb{R}^2$ is resonant[53].

Remark. To be commensurable is an equivalence relation, since it satisfies the following conditions, which are easy to verify:

- *Reflixity*: $a \sim a$ for any $a \in \mathbb{R}$.
- *Symmetry*: if $a \sim b$, then $b \sim a$.
- *Transitivity*: if $a \sim b$ and $b \sim c$, then $a \sim c$.

As a consequence of this, any set of real numbers, e.g. the set $\{\Omega_1, \dots, \Omega_N\}$, can be partitioned into disjoint equivalence classes. Accordingly, a weakly connected network having natural frequencies $\Omega_1, \dots, \Omega_N$ can also be partitioned into noninteracting subnetworks, which follows from theorem 3.6.4 and is stated in the following corollary ([30], p. 277).

Corollary 3.6.6. *All oscillators can be divided into pools, or ensembles, according to their frequencies Ω_i. Oscillators from different pools have incommensurable frequencies and interactions between them are negligible.*

Notice that the corollary contains two statements:

1. The network is partitioned into pools of oscillators.
2. Interactions between oscillators from different pools are negligible.

The first statement agrees with the well-known fact from statistical mechanics ([34]) that a network of oscillators can be decomposed into coherent pools if the strength of connections is strong enough in comparison with the distribution of frequencies. The second statement is novel and was first pointed out in [30]. It claims that oscillators having incommensurable frequencies work independently from each other even when they have synaptic contacts between them. Although physiologically present and active, those synaptic connections

[53] We recall that a vector $v = (v_1, \dots, v_N)^T \in \mathbb{R}^N$ is said to have a *resonant relation* if

$$c \cdot v = c_1 v_1 + \cdots + c_N v_N = 0$$

for some nonzero integer row vector $c = (c_1, \dots, c_N) \in \mathbb{Z}^N$.

average to zero. Therefore, they are functionally insignificant and do not play any role in the dynamics of the network.
From corollary 3.6.6 it follows that a weakly connected network of quasi-periodic oscillators can be partitioned into pools of oscillators according to their natural frequencies Ω_i. Since interactions between oscillators from different pools are negligible, we study only interactions between oscillators from the same pool, as done by Hoppensteadt and Izhikevich in [30]. Thus, without loss of generality we may assume that the entire network is one such pool.
For sake of simplicity we assume that these oscillators have equal natural frequencies $\Omega_1 = \cdots = \Omega_N$. If we use slow time $\tilde{t} = \varepsilon t$ and discard high-order terms in ε, then using the following change of variables:

$$\begin{aligned} \phi_i &= \varepsilon\tilde{\phi}_i \\ \omega_i &= \varepsilon\tilde{\omega}_i \\ \tilde{t} &= \varepsilon t \end{aligned}$$

the model (3.116) can be written in the form

$$\dot{\tilde{\phi}}_i = \tilde{\omega}_i + \sum_{j=1}^{N} H_{ij}(\tilde{\phi}_j - \tilde{\phi}_i) \tag{3.119}$$

Now we have already discussed in section 3.5.1 how we can recover the Kuramoto model (3.15) using Fourier series and considering only the first term. Hence, starting from Hopfield model we have managed to get to Kuramoto's equation

$$\dot{\phi}_i = \omega_i + \sum_{j=1}^{N} \frac{K_{ij}}{N} \sin(\phi_j - \phi_i) \tag{3.120}$$

In conclusion, we have linked Hopfield's description to Kuramoto model. This is important because it highlights how these descriptions are not separated but interconnected with each other and how we can move from one to another. We have thus succeeded in integrating and unifying the two models, which describe two important and different features visible in the brain, such as the emission of spikes and oscillations. This relationship is then useful since it allows us to move from the description of single neurons, experimentally not feasible, to one that takes into account groups of neurons with the same frequencies, as seen with Kuramoto.

APPENDIX A Proof of Eq. 2.48

Statement.

$$\Delta E = -\sum_i \sum_j T_{ij} V_j \Delta V_i$$

Proof. We prove this statement by steps:

1. First, we consider the case when only one component k changes $V_k(t+1) = -V_k(t)$, whereas the others remain the same $V_j(t+1) = V_j(t),\ \forall\ j \neq k$.

$$\begin{aligned}
\Delta E = E(t+1) - E(t) = & - \frac{1}{2}\sum_i \sum_{j\neq i} T_{ij} V_i(t+1) V_j(t+1) + \\
& + \frac{1}{2}\sum_i \sum_{j\neq i} T_{ij} V_i(t) V_j(t) = \\
\text{(case } i \neq k,\ j \neq k\text{)} \qquad = & - \frac{1}{2}\sum_{i\neq k} \sum_{\substack{j\neq i \\ j\neq k}} T_{ij} V_i(t) V_j(t) + \frac{1}{2}\sum_{i\neq k} \sum_{\substack{j\neq i \\ j\neq k}} T_{ij} V_i(t) V_j(t) + \\
\text{(case } i = k,\ j \neq k\text{)} \qquad & - \frac{1}{2}\sum_{j\neq k} T_{kj} V_k(t+1) V_j(t) + \frac{1}{2}\sum_{j\neq k} T_{kj} V_k(t) V_j(t) + \\
\text{(case } i \neq k,\ j = k\text{)} \qquad & - \frac{1}{2}\sum_{i\neq k} T_{ik} V_i(t) V_k(t+1) + \frac{1}{2}\sum_{i\neq k} T_{ik} V_i(t) V_k(t) + \\
\text{(case } i = k,\ j = k\text{)} \qquad & - \frac{1}{2} T_{kk} V_k(t) V_k(t) + \frac{1}{2} T_{kk} V_k(t) V_k(t) = \\
= & - \frac{V_k(t+1) - V_k(t)}{2} \sum_{j\neq k} T_{jk} V_j(t) + \\
& - \frac{V_k(t+1) - V_k(t)}{2} \sum_{j\neq k} T_{kj} V_j(t) = \\
(T \text{ symmetric}) \qquad = & - (V_k(t+1) - V_k(t)) \sum_{j\neq k} T_{kj} V_j(t)
\end{aligned}$$

2. If we change two components k and l, the passages are the same, but in this case we have to take into account the case when the indexes of the summation are equal or different from l. At the end we gain

$$\Delta E = -\sum_{ind\in I} (V_{ind}(t+1) - V_{ind}(t)) \sum_{j\neq ind} T_{kj} V_j(t)$$

with $I = \{k, l\}$.

3. When we extend the change to all the components i with $i \in I = \{1, \ldots, N\}$, we obtain the desired result

$$\Delta E = -\sum_i \sum_j T_{ij} V_j \Delta V_i$$

where we have consider in the summation also the case $j = i$, i.e. terms that are equal zero, since $T_{ii} = 0, \ \forall i$.

□

APPENDIX B Discretization of Eq.2.69

In section 2.2.1 and in section 2.3 we have presented the discrete and continuous Hopfield model. We want to establish a link between these two models by proving that the discretization of Eq. (2.69) corresponds to the discrete dynamics equation (2.6), as done in [40].
We recall the equations for the dynamics in the two cases:

- *Discrete case*

$$\begin{aligned} h_i(t+1) &= \sum_{j=1}^{N} T_{ij}V_j(t) + I_i & \text{(B.1)} \\ V_i(t+1) &= \operatorname{sgn}(h_i(t+1)) & \text{(B.2)} \end{aligned}$$

- *Continuous case*

$$\begin{aligned} \dot{u}_i &= \sum_{j=1}^{N} T_{ij}V_j(t) + I_i - u_i & \text{(B.3)} \\ V_i &= g(u_i) & \text{(B.4)} \end{aligned}$$

where we have supposed that in Eq. (2.69) R_i and C_i are equal 1.

Remark. We note that $u_i = h_i$ in the two models and in the continuous model the *sign* function is replaced with a generic sigmoid function g.

Consider the continuous dynamic equation (B.3). By Euler's method, it can be approximated in discrete time by replacing

$$\dot{u}_i(t) \simeq \frac{u_i(t+1) - u_i(t)}{s} \qquad \text{(B.5)}$$

In the above equation, s is the sampling period and as $s \to 0$ the accuracy of the approximation increases. The substitution of (B.5) into (B.3) yields

$$u_i(t+1) - u_i(t) = s(\sum_{j=1}^{N} T_{ij}V_j(t) + I_i - u_i(t)) \qquad \text{(B.6)}$$

Hence,if consider a unitary interval dimension $s = 1$, we obtain

$$u_i(t+1) = \sum_{j=1}^{N} T_{ij}V_j(t) + I_i \qquad \text{(B.7)}$$

that corresponds to the discrete dynamics equation (B.1).

APPENDIX C Spectral method

Consider a vector field $X : \mathbb{R}^N \to \mathbb{R}^N$ with related differential equation:

$$\dot{x} = X(x) \tag{C.1}$$

Let $t \mapsto \bar{x}(t)$ be a T-periodic solution of (C.1); i.e. for every t we have

$$\dot{\bar{x}}(t) = X(\bar{x}(t)) \quad \text{and} \quad \bar{x}(t+T) = \bar{x}(t) \tag{C.2}$$

We search for solutions of (C.1) of the form:

$$x(t) := \bar{x}(t) + \Delta x(t) \tag{C.3}$$

Impose now that (C.3) solve (C.1). We obtain

$$\dot{\bar{x}}(t) + \dot{\Delta} x(t) = X(\bar{x} + \Delta x(t)) \tag{C.4}$$

Expanding in Taylor series the vector field X in the second member of (C.4) we have

$$\dot{\bar{x}}(t) + \dot{\Delta} x(t) = \underbrace{X(\bar{x}(t))}_{\dot{\bar{x}}(t)} + X'(\bar{x}(t))\Delta x(t) + R(t, \Delta x(t))$$

where $R(t, \Delta x(t))$ is $O(|\Delta x|^2)$ and T-periodic in t, hence is uniformly limited in $t \in \mathbb{R}$. (This condition is very important, because it will allow us to use the *first method of Lyapunov*, that we will present later.) In conclusion:

$$\dot{\Delta} x(t) = X'(\bar{x}(t))\Delta x(t) + R(t, \Delta x(t)) \tag{C.5}$$

Then, the linearized system of (C.1) around the T-periodic solution $t \mapsto \bar{x}(t)$ is represented by the following linear equation with T-periodic coefficient:

$$\dot{\Delta} x(t) = X'(\bar{x}(t))\Delta x(t) \tag{C.6}$$

To solve this equation (C.6) we expose a classical result following, for instance, the notes *"Sistemi dinamici meccanici"* of Franco Cardin, [12], about linear systems with periodic coefficient.

Consider a homogeneous linear differential equation with T-periodic coefficient in $\mathbb{R}^N$:

$$\dot{x} = A(t)x, \qquad A(t+T) = A(t) \tag{C.7}$$

An important result for this type of system consists in being able to represent the resolving matrix as a product of a periodic matrix and a solution of a constant coherent system. We now see how to achieve this.
Suppose that $\Phi(t)$ is the resolving matrix of (C.7) and that it satisfies the condition $\Phi(0) = \mathbb{I}$. Any other resolving matrix will be of the type $\Psi(t) = \Phi(t)Q$, with Q an arbitrary non-singular matrix ($detQ \neq 0$), in fact:

$$\Psi'(t) = \Phi'(t)Q = A(t)\Phi(t)Q = A(t)\Psi(t)$$

From the existence and uniqueness theorem we gain that $det\Phi(t) \neq 0$ and in particular that $det\Phi(T) \neq 0$.
Furthermore, $\bar{\Phi}(t) = \Phi(t+T)$ is a resolving matrix of (C.7):

$$\dot{\bar{\Phi}}(t) = \dot{\Phi}(t+T) = A(t+T)\Phi(t+T) = A(t)\Phi(t+T)$$

and hence it will be obtained from $\Phi(t)$ through the right product with an appropriate constant matrix Q

$$\Phi(t+T) = \Phi(t)Q$$

Statement.

$$Q = \Phi(T)$$

Proof. To obtain the value of Q, we consider $S(t) = \Phi(t+T)\Phi(T)^{-1}$. We have that:

$$\frac{d}{dt}S(t) = \Phi'(t+T)\Phi(T)^{-1} = A(t+T)\Phi(t+T)\Phi(T)^{-1} = A(t)S(t)$$

So $S(t)$ is a fundamental matrix and $S(0) = \mathbb{I}$, then $S(t) = \Phi(t)$ which it means that

$$\Phi(t+T) = \Phi(t)\Phi(T)$$

Therefore Q in this case is equal to $\Phi(T)$. □

Now since $\Phi(T)$ is non-singular, from the theorem of existence of the logarithmic matrix it follows that exists a *logarithmic matrix* B, i.e. such that

$$e^B = \Phi(T)$$

Now we can state the Floquet theorem:

Theorem C.0.1 (Floquet)**.**

$$\Phi(t) = \mathcal{P}(t)e^{\frac{B}{T}t} \tag{C.8}$$

where $\mathcal{P}(t)$ is a periodic matrix function.

Proof. Define $\mathcal{P}(t)$ as $\mathcal{P}(t) = \Phi(t)e^{-\frac{B}{T}t}$, we obtain.

$$\mathcal{P}(t+T) = \Phi(t+T)e^{-\frac{B}{T}(t+T)} = \Phi(t)\Phi(T)e^{-B}e^{-\frac{B}{T}t} = \Phi(t)e^{-\frac{B}{T}t} = \mathcal{P}(t)$$

Hence, $\mathcal{P}$ is T-periodic and $\Phi(t) = \mathcal{P}(t)e^{\frac{B}{T}t}$. □

Going back to (C.5), we set:

$$A(t) := X'(\bar{x}(t)), \qquad T\text{-periodic matrix}$$

and

$$z := \Delta x$$

then (C.5) becomes

$$\dot{z} = A(t)z + R(t,z) \tag{C.9}$$

Then, the linearized system

$$\dot{z} = A(t)z \tag{C.10}$$

can be solved using Floquet C.0.1: the resolving matrix is given by $\Phi(t) = \mathcal{P}(t)e^{\frac{B}{T}t}$, where

$$\begin{cases} e^B = \Phi(T) \\ \mathcal{P}(t+T) = \mathcal{P}(t) \end{cases}$$

Remark. Note that since $\Phi(t)$ is an isomorphism and $det(e^{\frac{B}{T}t}) \neq 0$, then also $\mathcal{P}(t)$ is an isomorphism.

The solution of (C.10) is therefore

$$z(t) = \mathcal{P}(t)e^{\frac{B}{T}t}z(0)$$

In order to combine (C.9) with a linear differential equation with constant coefficients, we introduce the time-dependent and T-periodic transformation in t:

$$y = \mathcal{P}^{-1}(t)z \tag{C.11}$$

A vector field $Z(t,z)$, to which is associated a differential equation

$$\dot{z} = Z(t,z), \tag{C.12}$$

is *conjugated* through (C.11) in the vector field

$$\dot{y} = Y(t,y)$$

whose solutions are correlated with the solution of (C.12) through (C.11)[54]. We now want to study the vector field Y to understand it better:

$$\begin{aligned} \dot{y}(t) &= \dot{\mathcal{P}^{-1}z}(t) = \mathcal{P}^{-1}(Az + R(t,z))\Big|_{z=\mathcal{P}y} + \dot{\mathcal{P}}^{-1}\mathcal{P}y = \\ &= \mathcal{P}^{-1}(AP - \dot{\mathcal{P}})y + \mathcal{P}^{-1}R(t,\mathcal{P}y) \end{aligned}$$

since $\dot{\mathcal{P}}^{-1}\mathcal{P} + \mathcal{P}^{-1}\dot{\mathcal{P}} = 0$.

On the other side

$$\dot{\Phi} = \Big(\dot{\mathcal{P}} + \frac{\mathcal{P}B}{T}\Big)e^{\frac{B}{T}t} = A\Phi = A\mathcal{P}e^{\frac{B}{T}t}$$

from which

$$\dot{\mathcal{P}} + \frac{\mathcal{P}B}{T} = A\mathcal{P}, \qquad \Longrightarrow \qquad A\mathcal{P} - \dot{\mathcal{P}} = \frac{\mathcal{P}B}{T}$$

Hence:

$$\dot{y}(t) = \mathcal{P}^{-1}(\frac{\mathcal{P}B}{T})y + \mathcal{P}^{-1}(t)R(t,\mathcal{P}(t)y)$$

and in conclusion the vector field Y is given by:

$$Y(t,y) := \frac{B}{T}y + \mathcal{P}^{-1}(t)R(t,\mathcal{P}(t)y) \tag{C.13}$$

where in the second factor of the right-hand side member the dependence on t is T-periodic.

Theorem C.0.2. *If $\Re(SpectB) < 0$ then $y = 0$ is asymptotically stable for (C.13).*

Proof. Starting from the transformation (C.11), we have to prove that $y = 0$ is asymptotically stable for

$$\dot{y} = \frac{B}{T}y + R(t,y) \tag{C.14}$$

[54] This means that all the solutions of an equation are all and only the solutions of the other, less than transformations of the type (C.11).

where $R(t,y)$ is T-periodic in t (so uniformly limited in t) and $R(t,y) = O(|y|^2)$.
Let $W(y)$ be a Lyapunov function for the asymptotic stability of

$$\dot{y} = \frac{B}{T} y$$

For example: $W(y) = \langle y, Ly \rangle$, where

$$L := \int_0^{+\infty} e^{\frac{B^T}{T}t} e^{\frac{B}{T}t} dt$$

We have that:

$$\begin{aligned} \frac{B^T}{T} L + L\frac{B}{T} &= \int_0^{+\infty} \Big(\frac{B^T}{T} e^{\frac{B^T}{T}t} e^{\frac{B}{T}t} + e^{\frac{B^T}{T}t} e^{\frac{B}{T}t} \frac{B}{T} \Big) dt = \\ &= \int_0^{+\infty} \frac{d}{dt} \Big(e^{\frac{B^T}{T}t} e^{\frac{B}{T}t} \Big) dt = -\mathbb{I} \end{aligned}$$

Hence

$$\frac{B^T}{T} L + L\frac{B}{T} = -\mathbb{I}$$

Now we verify that $W(y)$ is a Lyapunov function for the asymptotic stability of (C.14).

1. $W(y)$ is positive definite:

$$\langle y, Ly \rangle = \int_0^{+\infty} \langle y, e^{\frac{B^T}{T}t} e^{\frac{B}{T}t} y \rangle \, dt = \int_0^{+\infty} |e^{\frac{B}{T}t} y|^2 dt \geq 0$$

   and we have the equality $\langle y, Ly \rangle = 0$ if and only if $y = 0$.

2. We have to prove that $\mathcal{L}_Y W(y) < 0$ uniformly in t in a neighborhood of $y = 0$.

$$\begin{aligned} \dot{W}(y) &= \Big\langle \frac{B}{T} y + R(t,y), Ly \Big\rangle + \Big\langle y, L\Big(\frac{B}{T} y + R(t,y) \Big) \Big\rangle = \\ &= \Big\langle y, \Big(\frac{B^T}{T} L + L\frac{B}{T} \Big) y \Big\rangle + \langle R(t,y), Ly \rangle + \langle y, LR(t,y) \rangle = \\ &= -|y|^2 + O(t, |y|^3) < 0, \qquad \forall\ y \neq 0 \end{aligned}$$

   where $O(t, |y|^3)$ is uniformly limited in t for what we have seen for $R(t, \Delta x(t))$.

□

Remark. If, instead of using the hypothesis of the theorem C.0.2, we consider the more weakly hypthoesis $\Re(SpectB) \leq 0$, then we speak of "linear stability". In this case, the equilibrium solution is not asymptotically stable, and sometimes not even stable, for the original system.

Theorem C.0.3. *If $\Re(SpectX'(\bar{x}(t))) < 0$, then the solution $t \mapsto \bar{x}(t)$ is asymptotically stable for (C.1).*

Proof. Using the transformation (C.11), the result follows directly from what has been proved in theorem C.0.2. □

We can now summarize all the results seen in the following theorem:

Theorem C.0.4 (Spectral method or First Lyapunov Method)**.** *Given the following differential equation $\dot{x} = X(x)$ with x^* an equilibrium point, such that $X(x^*) = 0$, if*

$$Spect \frac{\partial X(x^*)}{\partial x} = \{\lambda_1, \ldots, \lambda_N\} \in \mathbb{C} \tag{C.15}$$

then the equilibrium point x^* is:

- *asymptotically stable if all the eigenvalues λ_l are such that $\Re(\lambda_l) < 0$;*
- *unstable if exists at least a λ_l such that $\Re(\lambda_l) > 0$;*
- *undefined if for $\Re(\lambda_l) \geq 0$ exists at least a λ_l such that $\Re(\lambda_l) = 0$.*

APPENDIX D Proof of $\rho v = constant$

In the limit case of $N \to \infty$ it is convenient to formulate the system (3.19) in terms of density, rather than keeping track of each individual oscillator. Thus, if we focus on the dynamics of the system in this limit, we can consider the continuous case, in which for each frequency ω we imagine the oscillators distributed continuously on the circumference.
Define $\rho(\phi, \omega, t)d\phi$ the fraction of oscillators with angular frequency ω and angular position between ϕ and $\phi + d\phi$ at time t. This function is 2π-periodic with respect to the first component and satisfies the normalization condition

$$\int_{-\pi}^{\pi} \rho(\phi, \omega, t)d\phi = 1 \tag{D.1}$$

Applying the Law of large numbers[55] to Eq. (3.19) we gain that the dynamics of the system is governed, in the continuous case, by

$$\frac{\partial \tilde{\phi}}{\partial t}(\phi^{(0)}, \omega, t) = \omega + Kr\sin(\psi - \tilde{\phi}(\phi^{(0)}, \omega, t)) \tag{D.2}$$

with the initial condition

$$\tilde{\phi}(\phi^{(0)}, \omega, 0) = \phi^{(0)} \in \mathbb{S}^1. \tag{D.3}$$

Equation (D.2) is analogous to Eq. (3.19) for the discrete case.
Set $\rho^*(\phi^{(0)}, \omega, t_{in})$ the distribution of the frequencies between $\phi^{(0)}$ and $\phi^{(0)} + d\phi^{(0)}$ at the initial instant t_{in}. Note how, despite being the analogue of the discrete function $i \to \omega_i$, it provides more information about the latter. In particular, it highlights that for different initial phases there are distributions of different frequencies and therefore different evolutions of the system.
Fixed ω, we want to impose the conservation of the density of the oscillators:
for every $\Delta\mathbb{S}^1$ measurable in $\mathbb{S}^1$ it is valid

$$\int_{\phi^{(0)} \in \Delta\mathbb{S}^1} \rho^*(\phi^{(0)}, \omega, t_{in})d\phi^{(0)} = \int_{\phi \in \tilde{\phi}(\Delta\mathbb{S}^1, \omega, t)} \rho(\phi, \omega, t)d\phi \tag{D.4}$$

A change of variable in the second member gives

$$\begin{aligned} \int_{\phi^{(0)} \in \Delta\mathbb{S}^1} \rho^*(\phi^{(0)}, \omega, t_{in})d\phi^{(0)} &= \int_{\phi \in \tilde{\phi}(\Delta\mathbb{S}^1, \omega, t)} \rho(\phi, \omega, t)d\phi \\ &= \int_{\phi^{(0)} \in \Delta\mathbb{S}^1} \underbrace{\rho(\tilde{\phi}(\phi^{(0)}, \omega, t), \omega, t)}_{\hat{\rho}(\phi^{(0)}, \omega, t)} \left| det \frac{\partial \tilde{\phi}}{\partial \phi^{(0)}} \right| d\phi^{(0)} \end{aligned} \tag{D.5}$$

[55] There exists two different versions of the law of large numbers; they are called the *strong law of large numbers* and the *weak law of large numbers.* Stated for the case where $X_1, X_2, \dots$ is an infinite sequence of i.i.d. Lebesgue integrable random variables with expected value $E(X_1) = E(X_2) = \cdots = \mu$, both versions of the law state that (with virtual certainty) the sample average

$$\bar{X}_N = \frac{1}{N}(X_1 + \cdots + X_N)$$

converges to the expected value $\bar{X}_N \to \mu$ for $N \to \infty$.

For the arbitrariness of $\Delta\mathbb{S}^1$ we gain that

$$\rho^*(\phi^{(0)},\omega,t_{in}) = \hat{\rho}(\phi^{(0)},\omega,t)\left|det\frac{\partial\tilde{\phi}}{\partial\phi^{(0)}}\right| \qquad \text{for every } \phi^{(0)}\in\mathbb{S}^1 \tag{D.6}$$

We have then that $\rho^*(\phi^{(0)},\omega,t_{in})$ is correlated with $g(\omega)$; the density of presence of ω on $\mathbb{S}^1$ at $t=0$ is obtained integrating $\rho^*(\phi^{(0)},\omega,t_{in})$ on all $\phi^{(0)}$:

$$g(\omega) = \int_{\phi^{(0)}\in\mathbb{S}^1}\rho^*(\phi^{(0)},\omega,t_{in})d\phi^{(0)}$$

Defined $\hat{\rho}(\phi^{(0)},\omega,t) \coloneqq \rho(\tilde{\phi}(\phi^{(0)},\omega,t),\omega,t)$ the density of the oscillators in Lagrangian form and $\rho^*(\phi,\omega,t)$ that in Eulerian form, from Eq. (D.6) we obtain the equation of continuity for ρ

$$\frac{\partial\rho}{\partial t} + \frac{\partial}{\partial\phi}(\rho(\phi,\omega,t)v(\phi,\omega,t)) = \frac{\partial\rho}{\partial t} + \frac{\partial}{\partial\phi}[\omega + Kr\sin(\psi-\phi)\rho] = 0 \tag{D.7}$$

where

$$v(\phi,\omega,t) = \frac{\partial\tilde{\phi}}{\partial t}(\phi^{(0)},\omega,t)\bigg|_{\phi^{(0)}=\phi^{(0)}(\phi,\omega,t)} = \omega + Kr\sin(\psi-\phi) \tag{D.8}$$

is the instantaneous velocity of an oscillator in position ϕ and having an initial frequency ω and $\phi^{(0)}(\phi,\omega,t)$ is the inverse motion.
To conclude the evolution of ρ is governed by the continuity equation:

$$\frac{\partial\rho}{\partial t} = -\frac{\partial\rho v}{\partial\phi} \tag{D.9}$$

that expresses the conservation of the oscillators with frequency ω.
Now we want to prove the following result:

Lemma D.0.1. *The density $\rho(\phi,\omega,t)$ of oscillators in $(\phi,\phi+d\phi)$ in the stationary case*

$$\frac{\partial\rho}{\partial t} = 0, \qquad \frac{\partial v}{\partial t} = 0$$

is inversely proportional to the velocity $v(\phi,\omega,t)$ in ϕ, i.e.:

$$\rho = \frac{C}{v} \tag{D.10}$$

Proof. To verify the statement, we start from Eq. (D.9). If ρ is stationary, then $\frac{\partial\rho}{\partial t}=0$. So, for (D.9)

$$\frac{\partial\rho v}{\partial\phi} = 0 \qquad \Longrightarrow \qquad \rho v \equiv \text{constant}$$

and ρ must be inversely proportional to the velocity in ϕ. □

The continuity equation (D.9) can be solved using Eq. (D.1) together with the continuous counterpart of Eq. (3.16), that is given by

$$re^{i\theta} = \int_{-\pi}^{\pi}\int_{-\infty}^{\infty}e^{i\phi}\rho(\phi,\omega,t)d\phi d\omega \tag{D.11}$$

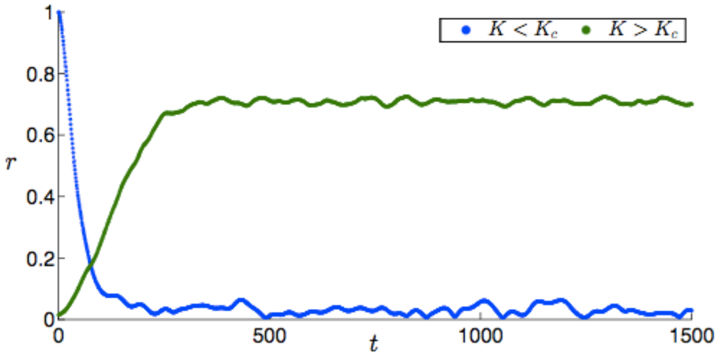


Figure D.1: Typical evolution of $r(t)$ for the model of Kuramoto. In the simulation, $N = 1000$ and g Gaussian were taken as initial data, which provided a critical value of $K_c \approx 1.16$. The blue curve corresponds to $K = 1$, the green curve to $K = 2$. Figure taken from the bachelor's thesis of Zeegers [68].

Therefore, the system of equations (D.1) - (D.7) - (D.11) admits as a trivial stationary solution

$$\rho = \frac{1}{2\pi}, \qquad r = 0 \tag{D.12}$$

independently of the choice we make for K and g, corresponding to an angular distribution of oscillators having an equal probability in the interval $[-\pi, \pi]$. In this case the oscillators have a chaotic motion and the solution is therefore called *incoherent* as for istance shown in Figure D.1.

In the extreme case of very strong *coupling strength*, $K \to \infty$, there is a *global synchronization* in which the oscillators all have the same phase and $r \to 1$. For finite values of K it is observed a lower degree of synchronization with amplitude of the order parameter between 0 and 1, as can be seen in Figure D.1.

APPENDIX E Residual method

The standard method to solve the integral

$$\int_{-\pi}^{\pi} \frac{C}{\omega - \Omega + Kr\sin(\phi)} d\phi, \qquad 0 < Kr < \omega - \Omega$$

is the residual method. Referring to the book *"Analisi Due"* by Giuseppe De Marco, [15], we recall how to use the residual method in the calculation of integrals of rational functions.

Definition E.0.1 (Residue)**.** *Let $\mathcal{U} \subseteq \mathbb{C}$ be an open set and $f : \mathcal{U} \to \mathbb{C}$ an holomorphic function having a singularity in $z_0 \in \partial\mathcal{U}$. The* residue *of f in z_0 is the coefficient a_{-1} of the Taylor expansion of f in the point z_0, that is $f(z) = \sum_{n=-\infty}^{\infty} a_n(z - z_0)^n$.*

Theorem E.0.2 (Residual Theorem)**.** *Let $\mathcal{U} \subseteq \mathbb{C}$ be an open simply connected set and be $z_1, \ldots, z_{\bar{p}} \subset \mathcal{U}$. Let it also $f : \mathcal{U} \setminus \{z_1, ..., z_{\bar{p}}\} \to \mathbb{C}$ an holomorphic function. For each closed path $\gamma \subset \mathcal{U} \setminus \{z_1, \ldots, z_{\bar{p}}\}$ it is true that:*

$$\int_\gamma f(z)dz = 2\pi i \sum_{j=1}^{\bar{p}} ind_\gamma(z_j) Res(f; z_j)$$

where $ind_\gamma(\cdot)$ is the winding number of the closed curve γ in the plane around a given point, that is an integer representing the total number of times that the curve travels counterclockwise around the point. The winding number depends on the orientation of the curve and is negative if the curve travels around the point clockwise.

Firstly, the following result is proved:

Lemma E.0.3. *Suppose that the power series*

$$\sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} \tag{E.1}$$

is convergent everywhere in $\mathbb{C} \setminus \{z_0\}$. For every closed path $\gamma \subset \mathbb{C} \setminus \{z_0\}$ we have then

$$\int_\gamma \Big(\sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n}\Big) dz = 2\pi i b_1 ind_\gamma(z_0)$$

Proof. Note that the series (E.1) is uniformly convergent on the compact sets of $\mathbb{C} \setminus \{z_0\}$ and therefore it is possible to integrate it term to term. Thus we obtain

$$\int_\gamma \frac{dz}{(z - z_0)^n} = \begin{cases} 0 & n \geq 2 \\ 2\pi i \ ind_\gamma(z_0) & n = 1 \end{cases}$$

□

Proof. (*Residual Theorem*). Indicate with

$$\sum_{n=-\infty}^{\infty} a_n^{(k)}(z-z_k)^n, \qquad k=1,\dots,\bar{p}$$

the Taylor expansion of f in z_k and set

$$p_k(z) \coloneqq \sum_{n=1}^{\infty} a_{-n}^{(k)}(z-z_k)^{-n}, \qquad k=1,\dots,\bar{p}$$

This last series converges in $\mathbb{C}\setminus\{z_k\}$ and hence its sum is an holomorphic function. Set $g(z) \coloneqq f - \sum_{k=1}^{\bar{p}} p_k(z)$, we have that g is an holomorphic function in $\mathcal{U}\setminus\{z_1,\dots,z_{\bar{p}}\}$ with a singularity that can be eliminated in each point z_k, $k=1,\dots,\bar{p}$. Note that it is possible to extend g in each of them, obtaining thus a holomorphic function in all $\mathcal{U}$. Therefore it is valid that $\int_\gamma g(z)dz = 0$, $\forall\gamma$ closed path in $\mathcal{U}$. This last relation, together with the previous lemma, allows to obtain

$$\begin{aligned} \int_\gamma f(z)dz &= \sum_{k=1}^{\bar{p}} \int_\gamma p_k(z)d(z) = 2\pi i \sum_{k=1}^{\bar{p}} a_{-1}^{(k)} ind_\gamma(z_k) \\ &= 2\pi i \sum_{k=1}^{\bar{p}} Res(f;z_k) ind_\gamma(z_k) \end{aligned} \tag{E.2}$$

□

Now we summarize the strategy for calculating the integrals used for the explicit calculation of the constant C. Assume that we want to calculate the following integral

$$I = \int_0^{2\pi} R(\cos\theta,\sin\theta)d\theta$$

where R is a real rational function in two variables defined in $\mathbb{S}^1 = \{(x,y)\in\mathbb{R}^2 \ : \ x^2+y^2=1\}$. Set $z=e^{i\theta}=\cos\theta+i\sin\theta$ we have

$$\begin{aligned} \cos\theta &= \frac{1}{2}\Big(z+\frac{1}{z}\Big) \\ \sin\theta &= \frac{1}{2i}\Big(z-\frac{1}{z}\Big) \\ dz &= d(e^{i\theta}) = ie^{i\theta}d\theta = izd\theta \qquad \Longrightarrow \qquad d\theta = \frac{dz}{iz} \end{aligned}$$

Thus substituting these relations in the integral I we gain

$$I = \int_0^{2\pi} R(\cos\theta,\sin\theta)d\theta = \int_\gamma R\Big(\frac{1}{2}\Big(z+\frac{1}{z}\Big),\frac{1}{2i}\Big(z-\frac{1}{z}\Big)\Big)\frac{dz}{iz}$$

where the argument of the integral is now a function that has no poles on γ.
Using the previous theorem E.0.2, the value of the integral is the sum of the residuals of the function on the poles with modulus strictly less than 1 multiplied by $2\pi i$. Therefore we obtain that I is equal to the sum of the residuals of the function

$$\frac{1}{iz}R\Big(\frac{1}{2}\Big(z+\frac{1}{z}\Big),\frac{1}{2i}\Big(z-\frac{1}{z}\Big)\Big)$$

corresponding to the poles contained in $B(0,1)$. It follows that

$$I = 2\pi \sum Res\Big(\frac{1}{z}R\Big(\frac{1}{2}\Big(z+\frac{1}{z}\Big), \frac{1}{2i}\Big(z-\frac{1}{z}\Big)\Big)\Big)$$

with the sum extended on the poles in $B(0,1) \in \mathbb{C}$.

APPENDIX F Proof of Theorem 3.6.3

We now prove Theorem 3.6.3, as done in [35], pp. 2217-2219.
The following theorem and lemmas are needed for the proof.

Theorem F.0.1 (Invariant Manifold Reduction for Weakly Connected Systems, Hoppensteadt and Izhikevich, [30])**.** *Suppose that a dynamical system*

$$\dot{u} = F(u), \qquad u \in \mathbb{R}^N$$

has an attractive normally hyperbolic compact invariant manifold $\mathcal{D} \in \mathbb{R}^N$. Then there is an $\varepsilon_0 > 0$ such that for all $\varepsilon \leq \varepsilon_0$ the dynamical system

$$\dot{u} = F(u) + \varepsilon G(u, \varepsilon), \qquad u \in \mathbb{R}^N \tag{F.1}$$

has a local model, which is defined on the unperturbed invariant manifold $\mathcal{D}$:

$$\dot{\tilde{u}} = F(\tilde{u}) + \varepsilon g(\tilde{u}, \varepsilon), \qquad \tilde{u} \in \mathcal{D} \tag{F.2}$$

where

$$g(\tilde{u}, 0) = G(\tilde{u}, 0) + ad_F P(\tilde{u}), \quad \textit{where} \quad ad_F P = DFP - DPF$$

is the Poisson bracket of the vector field F and the vector-valued function P can be determined from the condition

$$G(\tilde{u}, 0) + ad_F P(\tilde{u}) \in T_{\tilde{u}}\mathcal{D} \qquad \textit{for all } \tilde{u} \in \mathcal{D}$$

That is, there is an open neighborhood $\mathcal{W}$ of $\mathcal{D}$ and a continuous transformation $p_{\mathcal{W}} : \mathcal{W} \to \mathcal{D}$ that maps solutions of (F.1) to those of (F.2).

Proof. For the proof see [30], p. 138. □

Lemma F.0.2 (Samoilenko, [53])**.** *Consider a dynamical system of the form*

$$\dot{u} = F(u), \qquad u \in \mathbb{R}^N \tag{F.3}$$

having a quasi-periodic attractor $\mathcal{D}$ with the dimension $N > 1$. Let $\Omega \in \mathbb{R}^N$ be the frequency vector. Then (F.3) is conjugate to

$$\dot{\theta} = \Omega, \qquad \theta \in \mathbb{T}^N \tag{F.4}$$

on the attractor. That is, there is a homeomorphism $q : \mathbb{T}^N \to \mathcal{D}$ such that $u(t) = q(\theta(t))$ whenever $u(0) \in \mathcal{D}$.

We implicitly assume that all manifolds and functions are as smooth as necessary for our manipulations. In particular, we assume that the homeomorphism q above is a diffeomorphism (differentiable with differentiable inverse).

Lemma F.0.3 (Izhikevich, [35], p. 2218)**.** *Suppose that each uncoupled system ($\varepsilon = 0$)*

$$\dot{u}_i = F_i(u_i), \tag{F.5}$$

has a hyperbolic quasi-periodic equilibrium $\mathcal{D}_i$. Then, (F.5) considered as a system, i.e. an uncoupled dynamical system of the form (F.3), has a normally hyperbolic stable compact invariant manifold $\mathcal{D}$, which is the direct product of all $\mathcal{D}_i$.

Proof. Obviously, $\mathcal{D}$ is compact, invariant and stable. Let us prove its normal hyperbolicity. Since $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_N$, we have $T\mathcal{D} = T\mathcal{D}_1 \times \cdots \times T\mathcal{D}_N$ and $N\mathcal{D} = N\mathcal{D}_1 \times \cdots \times N\mathcal{D}_N$. Therefore, $v \in T\mathcal{D}$ can be represented as $v = v_1 + \cdots + v_N$, where $v_i \in T\mathcal{D}_i$. Similarly, $w = w_1 + \cdots + w_N$ for $w_i \in N\mathcal{D}_i$, $i = 1, \ldots, N$.
From the exponential stability of each $\mathcal{D}_i$ it follows that

$$\lim_{t\to\infty} |w_i(t)| = 0 \qquad \Longrightarrow \qquad \lim_{t\to\infty} |w(t)| = 0$$

i.e. we gain the exponential stability of $\mathcal{D}$.
To determine the dynamics of $v_i(t)$ we need to find the flow $\Phi_i(u_i, t)$ on $\mathcal{D}_i$. Using Lemma F.0.2 we see that

$$\Phi_i(X_i, t) = q_i(\Omega_i t + q_i^{-1}(u_i))$$

for some vector Ω_i and some diffeomorphism $q_i : \mathbb{T}^N \to \mathcal{D}_i$. Therefore

$$D_{u_i}\Phi(u_i, t) = Dq_i(\Omega_i t + q_i^{-1}(u_i))Dq_i^{-1}(u_i) = Dq_i(\Omega_i t + \theta_i))Dq_i(\theta_i)^{-1}$$

where $\theta_i = q_i^{-1}(u_i)$. Hence,

$$\begin{aligned} |v_i(t)| &= |D_{u_i}\Phi_i(u_i, t)v_i(0)| = |Dq_i(\Omega_i t + \theta_i)Dq_i(\theta_i)^{-1}v_i(0)| \\ &\le \min_{\vartheta_i, \theta_i \in \mathbb{T}} |Dq_i(\vartheta_i))Dq_i(\theta_i)^{-1}v_i(0)| \le \frac{\lambda_{min}(Dq_i)}{\lambda_{max}(Dq_i)} v_i(0) \end{aligned}$$

where $\lambda_{min}(Dq_i)$ and $\lambda_{max}(Dq_i)$ are the minimal and the maximal (in absolute value) eigenvalues of Dq_i on $\mathbb{T}$, respectively. They exist since $\mathbb{T}$ is compact. Moreover, $\lambda_{min}(Dqi) \neq 0$, since q_i is nonsingular on $\mathbb{T}$.
It follows that each $|v_i(t)|$ is uniformly bounded from 0 and hence $|v(t)|$ is too. Therefore,

$$\lim_{t\to\infty} \frac{|w(t)|}{|v(t)|} = 0$$

which completes the proof. □

Proof of Theorem 3.6.3. First, we apply Lemma F.0.3 to conclude that the system (3.105) for $\varepsilon = 0$ has a normally hyperbolic stable compact invariant manifold $\mathcal{D}$. Then, we use Theorem F.0.1 to determine the existence of an open neighborhood $\mathcal{W}$ of $\mathcal{D}$ and a mapping $p_{\mathcal{W}} : \mathcal{W} \to \mathcal{D}$ that projects all local solutions of (3.105) to those of the system

$$\dot{\tilde{u}}_i = F_i(\tilde{u}_i) + \varepsilon g_i(\tilde{u}_1, \ldots, \tilde{u}_N, \varepsilon), \qquad \tilde{u}_i \in \mathcal{D}_i, \quad i = 1, \ldots, N, \tag{F.6}$$

where each F_i is the same as in (3.105) and g_i are some functions.
Notice that the inverse of q_i, which we denote by $\mathcal{S}_i = q_i^{-1} : \mathcal{D}_i \to \mathbb{T}$, transforms each subsystem $\dot{\tilde{u}}_i = F_i(\tilde{u}_i)$ into the form $\dot{\theta}_i = \Omega_i$ thanks to Lemma F.0.2. Differentiating $\theta_i(t) = \mathcal{S}_i(\tilde{u}_i(t))$ with respect to t yields

$$\Omega_i = D\mathcal{S}_i(\tilde{u}_i)F_i(\tilde{u}_i)$$

for all $\tilde{u}_i \in \mathcal{D}_i$. Now we apply the diffeomorphism to the weakly connected system (F.6) to obtain

$$\dot{\theta}_i = D\mathcal{S}_i(\tilde{u}_i)[F_i(\tilde{u}_i) + \varepsilon g_i(\tilde{u}_1, \ldots, \tilde{u}_N, \varepsilon)] = \Omega_i + \varepsilon \mathcal{G}_i(\theta_1, \ldots, \theta_N, \varepsilon),$$

where

$$\mathcal{G}_i(\theta_1, \ldots, \theta_N, \varepsilon) = D\mathcal{S}_i(\tilde{u}_i)g_i(q_1(\theta_1), \ldots, q_N(\theta_N), \varepsilon)$$

The mapping p that transforms solutions of (3.105) to those of (3.107) is the superposition of $p_{\mathcal{W}} : \mathcal{W} \to \mathcal{D}$ and $(\mathcal{S}_1, \ldots, \mathcal{S}_N) : \mathcal{D} \to T^N$. □

Bibliography

[1] Alberto Abbondandolo, Olga Bernardi, and Franco Cardin. "Chain Recurrence, Chain Transitivity, Lyapunov Functions and Rigidity of Lagrangian Submanifolds of Optical Hypersurfaces:" in: *Journal of Dynamics and Differential Equations* 30.1 (2018), pp. 287–308.

[2] Juan A. Acebrón et al. "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Rev. Mod. Phys.* 77.1 (2005), pp. 137–185.

[3] S. V. B. Aiyer, M. Niranjan, and F. Fallside. "A theoretical investigation into the performance of the Hopfield model". In: *IEEE Transactions on Neural Networks* 1.2 (1990), pp. 204–215.

[4] Daniel J. Amit. *Modeling Brain Function- The world of attractor neural networks.* Cambridge University Press, 1989.

[5] Daniel J. Amit, Hanoch Gutfreund, and Haim Sompolinsky. "Information storage in neural networks with low levels of activity". In: *Phys. Rev. A* 35.5 (1987), pp. 2293–2303.

[6] Nirwan Ansari and Edwin Hou. *Computational Intelligence for Optimization.* Kluwer Academic Publishers, 1997.

[7] Michael A. Arbib and James J. Bonaiuto. *From Neuron to Cognition via Computational Neuroscience.* MIT Press, 2016.

[8] R.C. Atkinson and R.M. Shiffrin. "Human Memory: A Proposed System and its Control Processes". In: vol. 2. Psychology of Learning and Motivation. Academic Press, 1968, pp. 89–195.

[9] Yoram Baram. *Orthogonal Patterns in Binary Neural Networks.* NASA technical memorandum. National Aeronautics and Space Administration, Ames Research Center, 1988.

[10] Michael Breakspear, Stewart Heitmann, and Andreas Daffertshofer. "Generative Models of Cortical Oscillations: Neurobiological Implications of the Kuramoto Model". In: *Front. Hum. Neurosci.* 2010.

[11] Dean V. Buonomano and Michael Merzenich. "Cortical plasticity: from synapses to maps." In: *Annual review of neuroscience* 21 (1998), pp. 149–86.

[12] Franco Cardin. *Sistemi dinamici meccanici.* Cleup, 2018.

[13] Hayato Chiba. "A proof of the Kuramoto conjecture for a bifurcation structure of the infinite-dimensional Kuramoto model". In: *Ergodic Theory and Dynamical Systems* 35.3 (2013), pp. 762–834.

[14] M. A. Cohen and S. Grossberg. "Absolute stability of global pattern formation and parallel memory storage by competitive neural networks". In: *IEEE Transactions on Systems, Man and Cybernetics* SMC-13.5 (1983), pp. 815–826.

[15] G. De Marco. *Analisi due. Teoria ed esercizi.* Collana di matematica. Testi e manuali. Zanichelli, 1999.

[16] B. Derrida, E. Gardner, and A. Zippelius. "An Exactly Solvable Asymmetric Neural Network Model". In: *EPL (Europhysics Letters)* 4.2 (1987), pp. 167–173.

[17] M.V. Feigelman and Lev Ioffe. "The Augmented Models of Associative Memory Asymmetric Interaction and Hierarchy of Patterns". In: *International Journal of Modern Physics B - IJMPB* 1 (1987), pp. 51–68.

[18] J. L. Garciá-Palacios. *Introduction to the theory of stochastic processes and Brownian motion problems*. Universidad de Zaragoza. 2004.

[19] E Gardner, Stephan Mertens, and Annette Zippelius. "Retrieval properties of a neural network with an asymmetric learning rule". In: *Journal of Physics A: Mathematical and General* 22 (1989), pp. 2009–2018.

[20] Wulfram Gerstner and Leo van Hemmen. "Associative memory in a network of 'spiking' neurons". In: *Network Computation in Neural Systems* 3 (1992), pp. 139–164.

[21] Wulfram Gerstner and Werner Kistler. *Spiking Neuron Models: An Introduction*. New York, NY, USA: Cambridge University Press, 2002.

[22] W. Gerstner et al. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

[23] Peter Giesl and Sigurdur Hafstein. "Local Lyapunov functions for periodic and finite-time ODEs". In: *Recent Trends in Dynamical Systems*. Ed. by Andreas Johann et al. Springer Proceedings in Mathematics & Statistics. Springer, 2013, pp. 125–152.

[24] Hermann Haken. *Brain Dynamics: Synchronization and Activity Patterns in Pulse-Coupled Neural Nets with Delays and Noise (Springer Series in Synergetics)*. Springer-Verlag, 2006.

[25] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1998.

[26] D.O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Taylor & Francis, 2002.

[27] John Hertz, Richard G. Palmer, and Anders S. Krogh. *Introduction to the Theory of Neural Computation*. Perseus Publishing, 1991.

[28] John J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proc. Natl. Acad. Sci. USA* 79.8 (1982), pp. 2554–2558.

[29] John J. Hopfield. "Neurons with graded response have collective computational properties like those of two-state neurons". In: *Proceedings of the National Academy of Sciences* 81.10 (1984), pp. 3088–3092.

[30] F.C. Hoppensteadt and E.M. Izhikevich. *Weakly Connected Neural Networks*. Applied Mathematical Sciences. Springer New York, 1997.

[31] I. Kanter and H. Sompolinsky. "Associative recall of memory without errors". In: *American Physical Society* 35 (1987), pp. 380–392.

[32] Raymond P. Kesner and Edmund T. Rolls. "A computational theory of hippocampal function, and tests of the theory: New developments". In: *Neuroscience and Biobehavioral Reviews* 48 (2015), pp. 92–147.

[33] R. Klinke et al. *Fisiologia*. EdiSES, 2012.

[34] Y. Kuramoto. *Chemical Oscillations, Waves, and Turbulence*. Springer Series in Synergetics. Springer Berlin Heidelberg, 2012.

[35] Eugene M. Izhikevich. "Weakly Connected Quasi-periodic Oscillators, FM Interactions, and Multiplexing in the Brain". In: 59 (1999), pp. 2193–2223.

[36] E. A. Martens et al. “Exact results for the Kuramoto model with a bimodal frequency distribution”. In: *Phys. Rev. E* 79.2 (2009), pp. 026204–026216.

[37] A. D. Maruani, R. C. Chevallier, and G. Y. Sirat. “Information retrieval in neural networks. -I. Eigenproblems in neural networks”. In: *Revue de Physique Appliquee* 22.10 (1987), pp. 1321–1325.

[38] Warren Mcculloch and Walter Pitts. “A Logical Calculus of Ideas Immanent in Nervous Activity”. In: *Bulletin of Mathematical Biophysics* 5 (1943), pp. 127–147.

[39] R. McEliece et al. “The capacity of the Hopfield associative memory”. In: *IEEE Transactions on Information Theory* 33.4 (1987), pp. 461–482.

[40] A. N. Michel, J. A. Farrell, and H. F. Sun. “Analysis and synthesis techniques for Hopfield type synchronous discrete time neural networks with application to associative memory”. In: *IEEE Transactions on Circuits and Systems* 37.11 (1990), pp. 1356–1366.

[41] H. Mori, G.C. Paquette, and Y. Kuramoto. *Dissipative Structures and Chaos.* Springer Berlin Heidelberg, 2013.

[42] Takashi Nishikawa, Ying-Cheng Lai, and Frank C. Hoppensteadt. “Capacity of Oscillatory Associative-Memory Networks with Error-Free Retrieval”. In: *Phys. Rev. Lett.* 92.10 (2004), pp. 108101.1–108101.4.

[43] Emin Orhan. *The Hopfield model.* NYU. 2014.

[44] G Parisi. “Asymmetric neural networks and the process of learning”. In: *Journal of Physics A: Mathematical and General* 19.11 (1986), pp. L675–L680.

[45] Alan J Parkin. “Human memory: The hippocampus is the key”. In: *Current Biology* 6.12 (1996), pp. 1583–1585.

[46] Ulises Pereira and Nicolas Brunel. “Attractor Dynamics in Networks with Learning Rules Inferred from In Vivo Data.” In: *Neuron* 99.1 (2018), pp. 227–238.

[47] L. Personnaz, I. Guyon, and G. Dreyfus. “Information storage and retrieval in spin-glass like neural networks”. In: *Journal de Physique Lettres* 46.8 (1985), pp. 359–365.

[48] “Pharmacology And Nerve Endings: First Dixon Memorial Lecture By Sir Henry Dale”. In: *The British Medical Journal* 2.3859 (1934), pp. 1161–1163.

[49] A. Pikovsky et al. *Synchronization: A Universal Concept in Nonlinear Sciences.* Cambridge Nonlinear Science Series. Cambridge University Press, 2003.

[50] Rodrigo Quian Quiroga. *Borges e la memoria: Viaggio nel cervello umano da Funes al neurone Jennifer Aniston.* Collana: Il cervello e le idee. Erickson, 2018.

[51] Björn Rasch and Jan Born. “About sleep’s role in memory.” In: *Physiological reviews* 93.2 (2013), pp. 681–766.

[52] H. Risken and T. Frank. *The Fokker-Planck Equation: Methods of Solution and Applications.* Springer Series in Synergetics. Springer Berlin Heidelberg, 1996.

[53] A.M. Samoilenko. *Elements of the Mathematical Theory of Multi-Frequency Oscillations.* Mathematics and its Applications. Springer Netherlands, 2012.

[54] S.N. Sharma and H.G. Patel. *The Fokker-Planck Equation.* INTECH Open Access Publisher, 2010.

[55] H. Sompolinsky. “Neural networks with nonlinear synapses and a static noise”. In: *Phys. Rev. A* 34.3 (1986), pp. 2571–2574.

[56] H. Sompolinsky, A. Crisanti, and H. J. Sommers. “Chaos in Random Neural Networks”. In: *Phys. Rev. Lett.* 61.3 (1988), pp. 259–262.

[57] Amos Storkey and Romain Valabregue. “The Basins of Attraction of a new Hopfield Learning Rule”. In: *Neural networks : the official journal of the International Neural Network Society* 12.6 (1999), pp. 869–876.

[58] Steven H. Strogatz. “From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators”. In: *Physica D: Nonlinear Phenomena* 143.1 (2000), pp. 1–20.

[59] Humayun Karim Sulehria and Ye Zhang. “Study on the Capacity of Hopfield Neural Networks”. In: *Information Technology Journal* 7 (2008), pp. 684–688.

[60] D. W. Tank and J. J. Hopfield. “Neural computation by concentrating information in time”. In: *Proceedings of the National Academy of Sciences* 84.7 (1987), pp. 1896–1900.

[61] Leo Timms and Lars Q. English. “Synchronization in phase-coupled Kuramoto oscillator networks with axonal delay and synaptic plasticity.” In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 89.3 (2014), pp. 032906.1–032906.9.

[62] Alessandro Treves. “Computational constraints between retrieving the past and predicting the future, and the CA3-CA1 differentiation.” In: *Hippocampus* 14 (2004), pp. 539–556.

[63] M. V. Tsodyks and M. V. Feigel’man. “The Enhanced Storage Capacity in Neural Networks with Low Activity Level”. In: *EPL (Europhysics Letters)* 6.2 (1988), pp. 101–105.

[64] S. S. Venkatesh and D. Psaltis. “Linear and logarithmic capacities in associative neural networks”. In: *IEEE Transactions on Information Theory* 35.3 (1989), pp. 558–568.

[65] Jin-Liang Wang et al. *Analysis and Control of Coupled Neural Networks with Reaction-Diffusion Terms.* Springer, 2018.

[66] A.T. Winfree. *The Geometry of Biological Time.* Interdisciplinary Applied Mathematics. Springer New York, 2001.

[67] Han Yan et al. “Nonequilibrium landscape theory of neural networks.” In: *Proceedings of the National Academy of Sciences of the United States of America* 110.45 (2013), E4185–E4194.

[68] B.P. Zeegers. *Spontaneous Synchronization in Complex Networks.* 2015.

[69] Feng Zhang et al. “The potential and flux landscape theory of evolution.” In: *The Journal of chemical physics* 137.6 (2012), pp. 065102.1–065102.19.

[70] Lei Zhang et al. “Activity Invariant Sets and Exponentially Stable Attractors of Linear Threshold Discrete-Time Recurrent Neural Networks”. In: *IEEE Transactions on Automatic Control* 54.6 (2009), pp. 1341–1347.