Head Office: Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione - DEI

Ph.D. COURSE IN: Information Engineering
CURRICULUM: Information Science and Technologies
SERIES: 38th

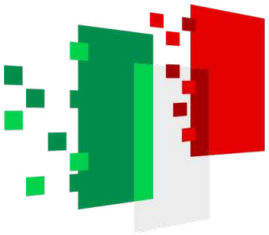
Associative Memory and Recurrent Neural Networks: A Dynamical Systems Approach

Thesis written with the financial contribution of the European Union - Next Generation EU

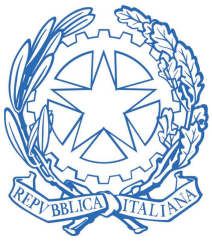
Coordinator: Prof. Fabio Vandin
Supervisor: Prof. Sandro Zampieri
Co-Supervisors: Prof. Giacomo Baggio
Prof. Francesco Bullo

Ph.D. student: Simone Betteti

Finanziato
dall'Unione europea
NextGenerationEU



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Ministero
dell'Università
e della Ricerca

Preamble

Have you ever thought of a distant friend, only to find yourself recalling a place, a shared meal, or a movie? It's strange—after all, the movie exists independently of your friendship, yet in your mind they are inseparably linked. Somewhere between recollection and meaning, memories seem to weave themselves together.

To explore this phenomenon, I dusted off a peculiar lens—figuratively speaking. It is no ordinary tool: it does not magnify everything equally, but selectively illuminates details that often seem inconsequential. Frustratingly partial, occasionally unreliable, yet persistently suggestive. When focused on memory, this lens revealed not the full picture, but a small, revealing detail: perception.

From there, the journey began. Psychology tells us that perception is not passive reception, but an active construction from sensory experience. Neuroscience adds precision: ensembles of neurons communicate through electrical and chemical synapses to encode, process, and retrieve information. But how do these microscopic interactions scale into coherent memories—or guide transitions between them?

This is where modeling becomes essential. It starts with Hopfield's seminal work on associative memory—a mathematical framework that, elegant as it is, answered some questions while raising many more. It showed how patterns can be recalled, but not how one memory leads to another, or how sequences of thought unfold over time.

Through this imperfect lens, I began to trace a thread—woven from biology, cognition, and formalism—twisting together in an attempt to capture something elusive. Each new insight added another filament, yet also increased the structure's complexity. At times, the lens failed entirely. Perhaps, as a small engraving on its frame reminded me: "This is not a magic orb. It reveals only what you have prepared yourself to see."

In this thesis, I aim to contribute a few additional filaments to the growing thread of theoretical neuroscience—small but, I hope, illuminating pieces in the ongoing effort to understand how memories form, connect, and guide thought.

Acknowledgments

I would like to express my deepest and most sincere gratitude to all those who have supported me throughout this journey.

First and foremost, I owe my deepest gratitude to my advisor, Prof. Sandro Zampieri, and to my co-advisors, Prof. Giacomo Baggio and Prof. Francesco Bullo. Their guidance, patience, and encouragement have shaped not only the course of this thesis but also my way of thinking as a researcher. To *Sandro*, I am grateful for believing in a stubborn student with an unconventional academic path. From you, I learned the value of meticulousness—the careful way you approach, investigate, and ultimately resolve difficult scientific questions has been an inspiring example to follow. To *Giacomo*, I am thankful for welcoming me into a new academic environment with generosity and trust. Your precision and systematic approach to work, together with the openness with which you accommodate diverging perspectives, have taught me both rigor and balance in research and collaboration. To *Francesco*, I am indebted for showing me how ideas can inspire and resonate. From you, I learned the art of striving for simplicity, of delivering messages that are clear, precise, and powerful enough to reach wide scientific audiences. Finally, to all three of you, I extend my heartfelt thanks for your unwavering support—intellectual, financial, and logistical—throughout these years, and especially during my stay in the beautiful city of Santa Barbara, CA.

I would also like to thank Prof. Fabio Fagnani and Prof. Giovanni Russo for their thoughtful feedback, constructive suggestions, and support during the evaluation process.

I am deeply grateful to my collaborators and colleagues in Padova, who have accompanied me through three remarkable years of research, dialogue, and shared adventures. Together, we transformed the challenges of daily work into moments of discovery, laughter, and camaraderie, making this journey far more personal and memorable. Equally, I am indebted to the wonderful colleagues I met in Santa Barbara, who welcomed me into their lives with generosity, warmth, and an openness that I will always treasure. I have learned so much from our conversations—about science, about life, and about the curiosity and perseverance that drive both. Beyond the lab, every beach volleyball match, ski trip, festival, and spontaneous outing became a thread in the tapestry of an unforgettable experience. I feel truly blessed to have been surrounded by such extraordinary people, whose friendship and inspiration have enriched this journey in ways that extend far beyond academia.

I am grateful to the administrative and technical staff at University of Padova for their assistance and for ensuring that the research environment was smooth and welcoming.

On a personal note, I would like to thank my family and friends for their unwavering support and encouragement. Above all, my deepest gratitude goes to my parents, who have walked beside me throughout this entire academic journey and who never ceased to believe in my ambitions.

Their love, patience, and faith have been my foundation. This thesis is dedicated to them, as a small testament to the boundless care they have always given me. I am equally grateful to my close friends, who kept me grounded and reminded me of the world beyond academia, while always remaining supportive of my passions. Their presence has been a constant source of balance, joy, and perspective. Finally, my heartfelt thanks go to Cecilia, who has accompanied me through the ups and downs, across distances and time zones, with grace and patience. You brought lightness and laughter into the sometimes overly serious life of an engineer, offering me countless reasons to smile, and to love.

Finally, I acknowledge the funding sources that supported this work, including the Ministry of Education and Research, through the project *NextGenerationEU C96E22000350007*, without which this research would not have been possible.

To all of you, I extend my heartfelt thanks.

Contents

List of Figures

Symbol	Description	Page
$\mathbb{N}$	Set of natural numbers $\{0, 1, 2, 3, \dots\}$	
$\mathbb{R}$	Set of real numbers	
$\mathbb{R}_{\geq 0}$	Set of non-negatvie real numbers	
$\mathbb{R}_{>0}$	Set of positive real numbers	
$\mathbb{R}^n$	Set of vectors of dimension n with real entries	
$\mathbb{0}_n$	Denotes the origin in $\mathbb{R}^n$	
$\mathbb{1}_n$	Denotes the vector of all ones in $\mathbb{R}^n$	
$\mathbb{R}^{n\times m}$	Set of $n \times m$ matrices with real entries	
$B_r(x) \subset \mathbb{R}^n$	Ball of radius $r > 0$ centred at $x \in \mathbb{R}^n$	
$B_r \subset \mathbb{R}^n$	Ball of radius $r > 0$ centred at $\mathbb{0}_n$	
$\mathcal{I}_n$	Identity matrix of order n (subscript omitted when clear from context)	
$\otimes$	Denotes the Kronecker product	
$\odot$	Denotes the Hadamard entrywise product	
$[x]_i$ or x_i	Denotes the i-th entry of the vector x	
A_{ij}	Denotes the (i, j)-th entry of the matrix A	
$A^\top$ $(x^\top)$	Denotes the transpose of matrix A (vector x)	
$\lambda_{min}(A)$	Denotes the minimum real eigenvalue of the symmetric matrix A	
$\lambda_{max}(A)$	Denotes the maximum real eigenvalue of the symmetric matrix A	
$\mathbf{X}$	Bold uppercase letters denote stochastic processes	
$C^k(\mathcal{X}; \mathcal{Y})$	Space of functions $f : \mathcal{X} \to \mathcal{Y}$ of class C^k for $k \geq 0$	
∇ (∇_x)	Denotes the gradient operator	
$\triangle$ $(\triangle_x)$	Denotes the Laplacian operator	
$\nabla \cdot$ $(\nabla_x \cdot)$	Denotes the divergence operator	
J_x	Denotes the Jacobian operator	
$\|\cdot\|_{\mathcal{X}}$	Denotes the $\mathcal{X}$-norm	
$\mathbb{P}$	Denotes the probability mass of discrete random variables	
$\mathbb{E}$	Denotes the expected value operator	
$T_x\mathcal{M}$	Denotes the tangent space at $x \in \mathcal{M}$, for $\mathcal{M}$ smooth manifold	
δ_{ij}	Denotes the Kronecher delta, with $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $j \neq i$	

1

Neurons, networks, and memory

1.1 The biology of the brain

The human brain remains one of the most complex and least understood dynamical systems in nature. It is composed of an estimated 10^{11} neurons, each establishing, on average, around 10^4 synaptic contacts [1], thereby generating a network of extraordinary richness and diversity. This dense connectivity does not emerge randomly but instead exhibits properties of a small-world topology, balancing the efficiency of local clustering with the advantages of long-range shortcuts. Such an organization is thought to support both specialized processing and global integration, two requirements that are central to cognition and behavior.

Despite occupying a volume of only about $1400 : \mathrm{cm}^3$, the human brain demonstrates a remarkable degree of structural specialization. Its convoluted cortical sheet, spread across two hemispheres, is conventionally divided into four major lobes—frontal, parietal, temporal, and occipital—each contributing to distinct aspects of sensorimotor and cognitive processing. Beneath the cortex lies the white matter, a massive bundle of myelinated axonal tracts that interconnect cortical and subcortical areas. The predominance of this connective tissue underlines the fact that brain function cannot be understood solely by examining isolated neurons; instead, it emerges from patterns of interaction across spatially distributed regions.

At the microscopic scale, the cortex is organized into a laminated structure, typically described in terms of six layers. These strata emerge during embryonic development through a carefully orchestrated process of neuronal migration and differentiation. Each layer is populated by distinct neuronal types, such as pyramidal and granular cells, and contributes to characteristic patterns of input–output connectivity. For example, layer IV serves as the main recipient of thalamic sensory input, while layers V and VI project to subcortical structures and the white matter. This laminar architecture not only provides anatomical compartmentalization but also underlies canonical circuit motifs that recur across cortical areas, hinting at a unifying computational principle.

Crucially, the development of the cortical network involves dynamic phases of synaptic over-

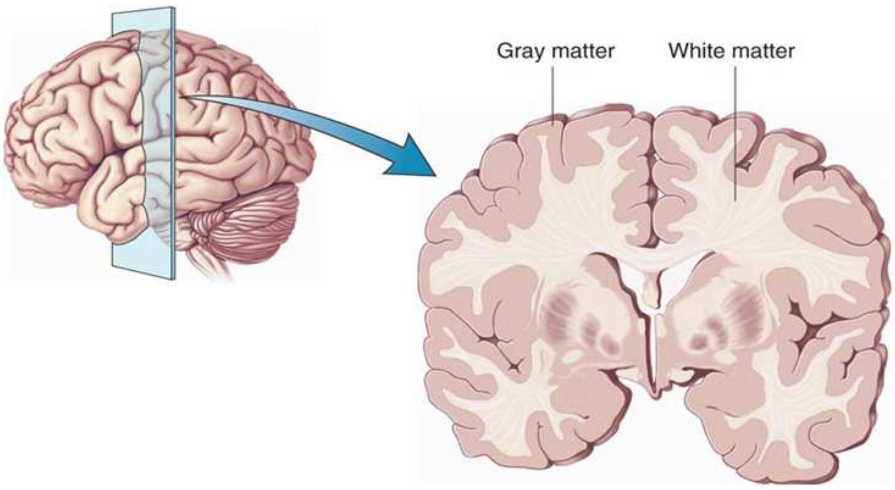


Figure 1.1: **The macro-organization of the human brain.** Visualization of the difference between gray and white matter [source: https://operativeneurosurgery.com/doku.php]. White matter appears as a thin, evenly distributed layer beneath the cortical surface and corresponds to the long-range axonal projections of neurons. In contrast, gray matter comprises the majority of brain volume and reflects the dense network of synaptic interconnections among neurons spanning different cortical regions.

production followed by activity-dependent pruning [2]. During early life, exuberant connectivity ensures that a wide repertoire of potential circuits can form. Subsequent pruning refines this architecture, stabilizing only those connections that are functionally relevant. This process of structural plasticity, tightly coupled to environmental input, is thought to be essential for the establishment of efficient neural codes. In humans, this developmental trajectory spans roughly the first 18 months of life, though more subtle refinements continue throughout adolescence. Such considerations highlight that the adult brain is not a fixed system but the outcome of a prolonged self-organizing process—a fact that complicates but also enriches any attempt at mathematical modeling.

The neuron and the action potential Within this intricate network, neurons act as the primary computational units. Alongside them, glial cells provide metabolic support, regulate ion concentrations, and participate in immune functions, outnumbering neurons by a factor of 2–10 [3]. While historically excluded from models of information processing, emerging evidence suggests that glia actively modulate synaptic transmission and may play a nontrivial role in higher cognition [4]. Nevertheless, in the present work we restrict our attention to neurons, which remain the dominant contributors to electrical signaling in the cortex.

A neuron's morphology is highly specialized for signal integration and transmission. Its soma houses the nucleus and essential metabolic machinery, while dendritic arbors receive

synaptic input. Inputs are integrated according to spatial and temporal summation, allowing the neuron to act as a nonlinear integrator of information. The axon, often myelinated to accelerate conduction, propagates the neuron's output toward thousands of synaptic terminals. At these terminals, transmission may occur through electrical gap junctions or more commonly via chemical neurotransmission, with diverse temporal dynamics and plasticity mechanisms.

The hallmark of neuronal function is the generation of the action potential, a stereotyped electrical pulse initiated at the axon hillock. This process relies on the neuron's excitable membrane, which maintains a resting potential of approximately $-70 : \mathrm{mV}$ through the differential distribution of ions—primarily K^+ inside and Na^+, Cl^-, and Ca^{2+} outside the cell. Excitatory input currents depolarize the membrane, and once a threshold (around $-45 : \mathrm{mV}$) is reached, voltage-gated sodium channels open, leading to a rapid influx of Na^+. This event triggers the rising phase of the action potential, which is soon counterbalanced by the delayed efflux of K^+ through potassium channels. The combined action of these ionic currents, together with the sodium–potassium pump, restores the resting state and prepares the membrane for subsequent excitations [6].

The all-or-none nature of this process was first characterized by Hodgkin and Huxley in their pioneering work on the squid giant axon, which established the foundation of modern computational neuroscience. Their model described how ionic conductances can be captured mathematically to predict the temporal evolution of the action potential—a framework that remains a cornerstone for theoretical studies of neuronal excitability.

The all-or-none excitability of neurons renders them efficient computational units, capable of transmitting information through rich spatio–temporal patterns of activity. Yet, despite their functional versatility, single neurons remain limited in explanatory power: they cannot, in isolation, account for the emergence of higher cognitive functions. It is only through the coordinated activity of large neuronal populations that complex phenomena such as perception, learning, and memory arise. In this sense, neuronal networks constitute the minimal ensembles in which abstract cognitive processes can be meaningfully studied. In the remainder of this chapter, our attention will therefore shift from the single neuron to networks of neurons, with a particular focus on those implicated in the formation and retrieval of memories.

1.1.1 The limbic system and memory

The limbic system—sometimes referred to as the paleomammalian brain—is not a single, homogeneous structure but rather a set of interconnected regions in the brain that collectively subserve functions central to survival and adaptation. These include emotional regulation, motivational drive, olfaction, social behavior, and, most prominently for the present discussion, learning and memory [7]–[9]. Despite its centrality in neuroscience, the limbic system is not defined

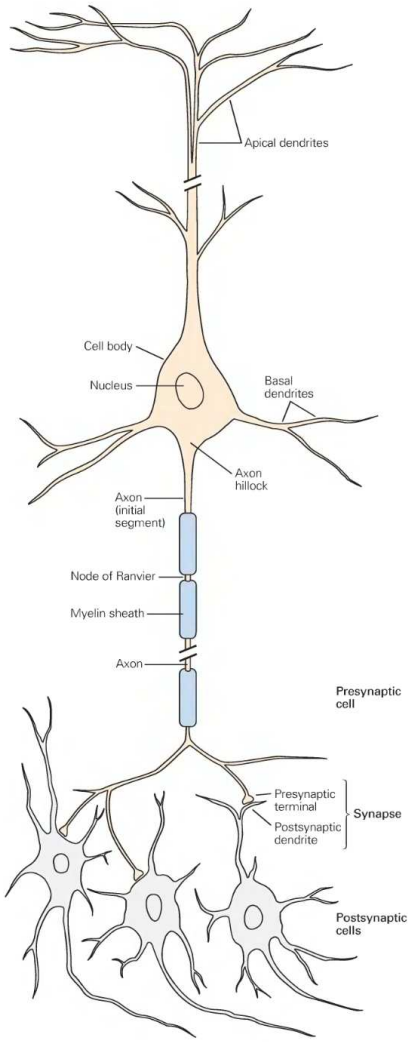


Figure 1.2: **The microscopic structure of the neuron.** Structure of a neuron [source: [5]]. Inputs from other neurons are collected through the apical and basal dendrites, which converge onto the soma containing the nucleus. Electrical integration occurs at the axon hillock, where action potentials are initiated and propagated along the axon. The axon is insulated by myelin sheaths, interrupted by nodes of Ranvier that facilitate rapid saltatory conduction. The axon terminates in presynaptic terminals that release neurotransmitters onto postsynaptic dendrites of connected neurons, forming synapses that mediate intercellular communication. This organization underlies the neuron's dual role as an integrative and transmitting unit within neural circuits.

by a universally accepted anatomical boundary: depending on the source, it may encompass the hippocampus, amygdala, hypothalamus, mammillary bodies, fornix, septal nuclei, cingulate gyrus, entorhinal cortex, and olfactory bulbs, among others (see Fig. 1.4). This lack of consensus

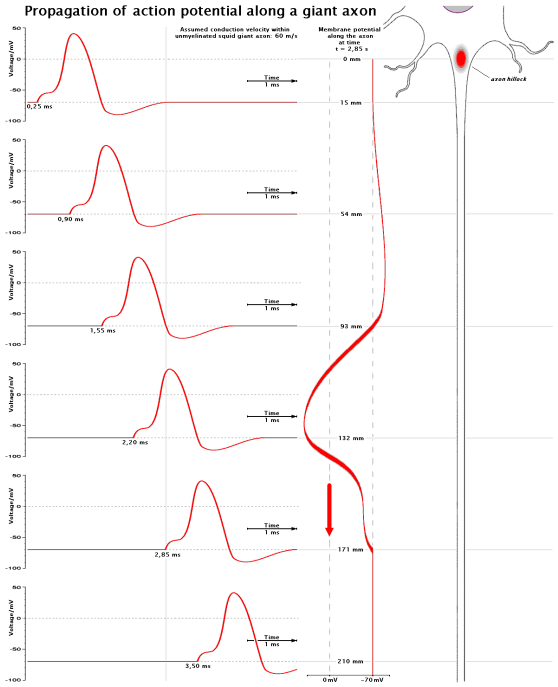


Figure 1.3: **The propagation of the action potential.** Propagation of an action potential in the squid giant axon [source: https://commons.wikimedia.org/wiki/File:Propagation_of_action_potential_along_a_giant_axon_en.png].

reflects the functional heterogeneity of the constituent structures, as well as the ongoing debate on whether the limbic system should be considered an integrated "system" at all or rather a functional label applied to partially overlapping networks.

A defining characteristic of the limbic system is its contribution to memory. Classical lesion studies, modern neuroimaging, and invasive recordings converge in identifying three regions as particularly critical: the amygdala, the hippocampal formation, and the entorhinal cortex. While other structures such as the cingulate cortex and mammillary bodies also contribute to specific memory functions, these three regions form the backbone of what is sometimes called the medial temporal lobe (MTL) memory system [10], [11].

The amygdala The amygdala consists of two almond-shaped clusters of nuclei buried deep within the medial temporal lobes. Its functional repertoire extends beyond its well-known role in fear processing: the amygdala participates in evaluating the emotional salience of stimuli, guiding decision-making, and modulating memory consolidation [12], [13]. Through reciprocal connections with the hippocampus and prefrontal cortex, the amygdala can enhance the encoding and persistence of emotionally arousing events. This modulation provides an adaptive advantage,

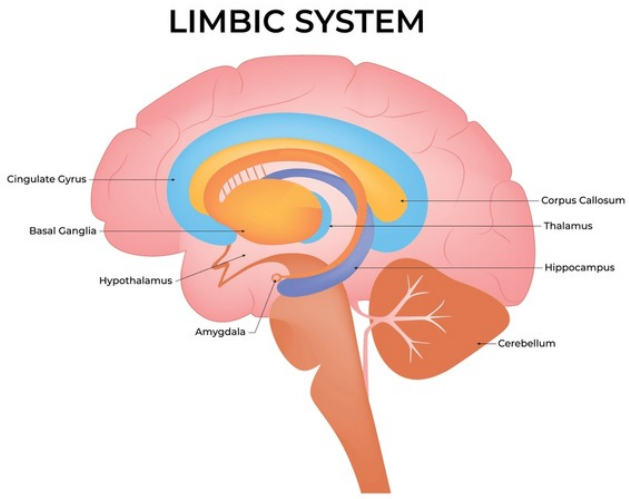


Figure 1.4: **The limbic system.** Schematic of the human limbic system, highlighting the hippocampus, amygdala, thalamus, hypothalamus, cingulate gyrus, basal ganglia, and corpus callosum. The hippocampus plays a critical role in episodic and spatial memory, while the amygdala modulates emotional memory. The thalamus and hypothalamus integrate sensory and homeostatic signals, influencing memory consolidation. Together, these interconnected regions form the neural substrate supporting encoding, retrieval, and emotional modulation of memories. [source: https://www.shutterstock.com/image-vector/limbic-system-science-design-vector-600nw-2404944453.jpg]

ensuring that experiences with high survival value—whether threatening or rewarding—are prioritized in memory storage.

The hippocampus The hippocampus has been a cornerstone of memory research since the case of patient H.M., whose bilateral hippocampal lesions led to profound anterograde amnesia [14]. Structurally, the hippocampal formation includes the hippocampus proper (Cornu Ammonis, CA fields), the dentate gyrus (DG), and the subiculum, with some definitions extending to the entorhinal cortex [15].

The hippocampus proper is subdivided into CA1–CA4, each with distinct connectivity and putative computational roles. For instance, CA3, with its dense recurrent collaterals, has been hypothesized to support autoassociative memory and pattern completion, a function reminiscent of attractor dynamics in artificial neural networks [16], [17]. By contrast, CA1 is thought to act as a comparator, detecting mismatches between predicted and incoming inputs [18].

The dentate gyrus is notable for its role in pattern separation, orthogonalizing overlapping inputs from the entorhinal cortex to minimize interference between similar memories [19]. Remarkably, it is also one of the few brain regions where adult neurogenesis occurs, a process

hypothesized to contribute to the encoding of novel experiences.

Taken together, the hippocampal formation plays a pivotal role in the consolidation of episodic and spatial memory, enabling the flexible representation of events across both time and space.

The entorhinal cortex The entorhinal cortex (EC), located adjacent to the hippocampus in the medial temporal lobe, functions as the principal gateway between the hippocampus and neocortex. It is divided into lateral and medial subdivisions, each with distinct roles. The medial EC is particularly famous for containing grid cells, neurons that fire in a periodic hexagonal pattern across space, providing a metric for spatial navigation [20], [21]. This discovery established the entorhinal–hippocampal network as a canonical system for spatial coding, later generalized to other domains such as conceptual and temporal spaces.

Functionally, the EC–hippocampus loop is central to declarative memory—encompassing episodic and semantic knowledge—and participates in memory formation, consolidation during sleep, and retrieval [22]. Disruptions of entorhinal connectivity are among the earliest hallmarks of Alzheimer's disease, underscoring its fundamental importance [23].

The above discussion highlights that memory is not localized to a single structure but emerges from the dynamic interaction of multiple limbic regions. The amygdala modulates the salience of encoded material, the hippocampus provides mechanisms for associativity, pattern separation, and sequence encoding, while the entorhinal cortex acts as a hub integrating cortical input and distributing hippocampal output.

From a computational standpoint, this distributed architecture resonates with models of attractor networks, associative memory, and state-space representations, which attempt to capture how stable yet flexible representations of past experiences can be maintained and retrieved. Yet, despite decades of research, fundamental questions remain open: How do the hippocampus and entorhinal cortex cooperate to balance pattern separation and pattern completion? To what extent can theories of attractor dynamics explain the stability of episodic memories? And how does emotional modulation reshape the attractor landscape? These open problems directly motivate the modeling perspective pursued in this thesis.

Notation We identify with $C^k(\mathcal{X};\mathcal{Y})$ the class of k-differentiable functions from $\mathcal{X}$ into $\mathcal{Y}$. Let $f \in C^1(\mathbb{R}^d;\mathbb{R})$ and denote with $\nabla f(x) \in \mathbb{R}^d$ its gradient. We identify the partial derivative of f with respect to x_i as $\partial_{x_i} f(x)$. Let $g \in C^1(\mathbb{R}^d;\mathbb{R}^d)$ and denote with $J_x g(x) \in \mathbb{R}^{d\times d}$ its Jacobian. We identify the identity matrix of dimension $d \in \mathbb{N}$ as $\mathcal{I}_d \in \mathbb{R}^{d\times d}$. We refer to a positive definite matrix $A \in \mathbb{R}^{d\times d}$ as $A \succ 0$. Let $B \in \mathbb{R}^{d\times d}$ and we denote its trace operator as $\mathrm{Tr}\,(B)$. Let $f \in C^k(\mathbb{R}^d;\mathbb{R})$ with $k > 2$; then f is strictly convex if $J_x(\nabla f(x)) \succ 0$. We identify with $\mathbb{P}(X)$ the probability mass of a discrete random variable X. We identify with $\mathbb{E}[h(x)]$ the expected value of a measurable function h where x is a generic random varibale (either discrete

or continuous). When specified as $\mathbb{E}_x[\cdot]$, the expected value is taken with respect to the measure associated to the events of the random variable x. We denote with $\mathbf{B}_t$ the standard n-dimensional Browniam motion process at time $t \geq 0$.

1.2 Modeling memory circuits as dynamical systems

The theoretical neuroscience literature offers an abundant choice of models capturing neural activity, varying in their level of details and their degree of biological realism. Perhaps the most known model- for its historical prestige -is the Hodgkin-Huxley model [24], which captures the propagation of electrical signals in the neurons' axon. The fine granularity of details captured by the Hodgkin-Huxley model makes it an ideal choice to study single neuron conductance, but unsuitable for analytical and computational studies of large numbers of neurons. Another popular class of models known for their biological plausibility is the class of Integrate-and-Fire models (IFs). The models belonging to the IFs class efficiently couple a mechanism for the integration of the inputs and a switch that produces a threshold dependent binary output. Despite being known for their computational efficiency, the IFs models present limitations in their analytical treatability due to their switching nature. Neural network models based on dynamical systems present a sound compromise that preserves both analytical treatability and computational efficiency. In the following subsection, I will present the two interesting models for the study of neural network properties that are based on dynamical systems.

1.2.1 The *voltage* model

The *voltage* model is defined by an autonomous ordinary differential equation (O.D.E.)

$$\begin{cases} \dot{x}_{\mathrm{H}}(t) = -x_{\mathrm{H}}(t) + W\Psi(x_{\mathrm{H}}(t)) + u(t), \\ x_{\mathrm{H}}(0) \in \mathbb{R}^N \end{cases} \tag{V}$$

where $x_{\mathrm{H}}(t)$ is an N-dimensional, time-dependent vector containing voltages of neuronal populations, $\dot{x}_{\mathrm{H}}$ is its time derivative and $x_{\mathrm{H}}(0) \in \mathbb{R}^N$ is the initial condition. In addition, we have that $W \in \mathbb{R}^{N\times N}$ is the synaptic matrix, $\Psi\colon \mathbb{R}^N \to \mathbb{R}^N$ denotes the activation function, and $u \in \mathbb{R}^N$ is a non-necessarily constant external input. The model describes the evolution of voltages in state-space based on the interaction among neurons, which could be either symmetric or asymmetric. The following assumptions specialize the treatment of *voltage* models to what are known as attractor networks - that is, network where the state evolves towards the fixed points for the dynamics.

Assumption 1.2.1 (Symmetry of W). The synaptic matrix $W \in \mathbb{R}^{N\times N}$ is symmetric, i.e. $W = W^\top$.

Assumption 1.2.2 (Monotonicity of the activation function). There exists a (strictly) convex function $\mathcal{L} : \mathbb{R}^N \to \mathbb{R}$, $\mathcal{L} \in C^2(\mathbb{R}^N)$ such that $\nabla\mathcal{L}(x) = \Psi(x)$.

Assumption 1.2.3 (Fixed input). The external input is constant, i.e. $u(t) \equiv u$.

Convergence to any of the fixed points for the dynamics (V) depends on the initial condition, i.e. the trajectory of the system will converge to the fixed point closest to the initial condition according to some distance. Proving convergence to the fixed point relies on the existence of an Energy (Lyapunov) function that has negative total time derivative along the solutions of (V).

Definition 1.2.4 (Energy of the *voltage* model). The *voltage* Energy function is a function $\mathrm{E}_H : \mathbb{R}^N \to \mathbb{R}$, $\mathrm{E}_H \in C^2(\mathbb{R}^N)$ defined as

$$\mathrm{E}_H(x) = -\frac{1}{2}\Psi(x)^\top W\Psi(x) + (x-u)^\top\Psi(x) + \mathcal{L}(x). \tag{1.1}$$

Proposition 1.2.5 (Convergence to fixed points). *Let $\mathrm{E}_H : \mathbb{R}^N \to \mathbb{R}$ be the Energy function as defined in (1.1) and $\mathcal{D} = \{x \in \mathbb{R}^N : -x + W\Psi(x) + u = \mathbb{0}_N\}$. Then along any trajectory of the system (V) we have that*

$$\frac{d}{dt}\mathrm{E}_H(x_H(t)) < 0 \qquad \forall x_H(t) \notin \mathcal{D}, \tag{1.2}$$

$$\frac{d}{dt}\mathrm{E}_H(x_H(t)) = 0 \iff x_H(t) \in \mathcal{D}. \tag{1.3}$$

The proof, presented in [25], is quite straightforward and will be reported for completeness.

Proof. Using assumption 1.2.1 on the symmetry of $W \in \mathbb{R}^{N\times N}$, we have that the total time derivative of (1.1) along the flow associated to (V) is

$$\begin{aligned}\frac{d}{dt}\mathrm{E}_H(x_\mathrm{H}) &= -\Psi(x_\mathrm{H})^\top W J_x\Psi(x_\mathrm{H})\dot{x}_\mathrm{H} + (x_\mathrm{H}-u)^\top J_x\Psi(x_\mathrm{H})\dot{x}_\mathrm{H} + \cancel{\Psi(x_\mathrm{H})^\top\dot{x}_\mathrm{H}} - \cancel{\nabla\mathcal{L}(x_\mathrm{H})^\top\dot{x}_\mathrm{H}} \\ &= (-W\Psi(x_\mathrm{H}) + x_\mathrm{H} - u)^\top J_x\Psi(x_\mathrm{H})\dot{x}_\mathrm{H} \\ &= -\dot{x}_\mathrm{H}^\top J_x\Psi(x_\mathrm{H})\dot{x}_\mathrm{H}\end{aligned} \tag{1.4}$$

where from the first to the second passage we have used assumptions 1.2.2 and 1.2.3. Notice now that the Jacobian $J_x\Psi(x) = J_x\nabla\mathcal{L}(x)$, and since $\mathcal{L}$ is convex and of class C^2, $J_x\Psi(x)$ is also a positive definite matrix $J_x\Psi(x) \succ 0$ for all $x \in \mathbb{R}^N$. Therefore

$$\frac{d}{dt}\mathrm{E}_H(x_\mathrm{H}) \leq 0 \qquad \forall x_\mathrm{H} \in \mathbb{R}^N \tag{1.5}$$

and in particular $\frac{d}{dt}\mathrm{E}_H(x_\mathrm{H}) = 0$ if and only if $\dot{x}_\mathrm{H} = 0$. □

Example 1.2.6 (**The Energy of the classic Hopfield model**). The most known instance of the voltage model is the Hopfield model. The original continuous time Hopfield model [26] is a specific instance of the general *voltage* model (V). In particular, in its pioneering work Hopfield drew inspiration from the theory of spin glass models (see Fig. 1.5 for a pictorial representation) and treated neural voltages as spins from an all to all lattice. Thus, the activation function was chosen to be diagonal, homogeneus, and taking values in $[-1,1]^N$. Specifically, it means that $\Psi_i(x) = \psi(x_i)$ for a function $\psi : \mathbb{R} \to [-1,1]$ and all $i = 1,\dots,N$, e.g. $\psi(x) = \tanh(x)$. Thereby, it is easy to observe that for $\mathcal{L}(x_\mathrm{H}) = \sum_{i=1}^N \int_0^{x_{\mathrm{H}i}} \psi(z)\, dz$ the associated Energy function (1.1) becomes

$$\begin{aligned}
\mathrm{E}_H(x_\mathrm{H}) &= -\frac{1}{2}\Psi(x_\mathrm{H})^\top W \Psi(x_\mathrm{H}) + (x_\mathrm{H} - u)^\top \Psi(x) + \sum_{i=1}^N \int_0^{x_{\mathrm{H}i}} \psi(z)\, dz \\
&= -\frac{1}{2}\Psi(x_\mathrm{H})^\top W \Psi(x_\mathrm{H}) + \sum_{i=1}^N \int_0^{x_{\mathrm{H}i}} \frac{d}{dz}\left(z\psi(z)\right) - \psi(z)\, dz - u^\top \Psi(x_\mathrm{H}) \\
&= -\frac{1}{2}\Psi(x_\mathrm{H})^\top W \Psi(x_\mathrm{H}) + \sum_{i=1}^N \int_0^{x_{\mathrm{H}i}} z\psi'(z)\, dz - u^\top \Psi(x_\mathrm{H}) \\
&= -\frac{1}{2}\Psi(x_\mathrm{H})^\top W \Psi(x_\mathrm{H}) + \sum_{i=1}^N \int_0^{\psi(x_{\mathrm{H}i})} \psi^{-1}(s)\, ds - u^\top \Psi(x_\mathrm{H})
\end{aligned} \tag{1.6}$$

where in the second to third passage we have used integration by parts and in the last passage we have performed the change of variable inside the integral $\psi(z) = s$. This is exactly the Energy provided in [26].

Memory assignment and retrieval The memories of an associative memory model are vectors stored in the model parameters and retrieved as fixed points for the model dynamics. The mapping of memories into model parameters can be either explicit, obtain from a known analytical rule, or implicit, hence resulting from a training procedure. Formally, we start by defining a generic set of memories.

Definition 1.2.7 (**Memories**). Define the set of memories of the associative memory model as a set

$$\Sigma := \{\xi^1, \dots, \xi^P\} \qquad P \in \mathbb{N} \tag{1.7}$$

where $\xi^\mu \in \mathbb{R}^N$, $\mu = 1,\dots,P$ are the desired memory vectors.

Once the memory vectors are defined, it is necessary to define the proper way of storing such memories into the model.

Definition 1.2.8 (Memory mapping). Let $\Sigma \subset \mathbb{R}^N$ be the set of memories defined in 1.2.7. The memories are stored in the model parameter by means of a function $\mathcal{W} : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ such that

$$\Sigma \mapsto \mathcal{W}(\Sigma) = W \tag{1.8}$$

where $W \in \mathbb{R}^{N\times N}$ is the usual synaptic matrix.

Once the desired memories have been encoded in the model via the map $\mathcal{W}$, we are interested in studying the set of vectors that are stable fixed points for the dynamics (V).

$$\Sigma^\star := \{\xi^1_\star, \dots, \xi^P_\star\} \tag{1.9}$$

$$\mathbb{0}_N = -\xi^\mu_\star + W\Psi(\xi^\mu_\star) \qquad \forall \xi^\mu_\star \in \Sigma^\star. \tag{1.10}$$

The vectors $\xi^\mu_\star \in \Sigma^\star$ are the patterns that the associative memory model is actually able to retrieve as stable equilibria for the dynamics, and are related to the respective $\xi^\mu \in \Sigma$ by the activation function and the synaptic matrix W. It now becomes important to understand what is the maximal cardinality of $\Sigma \subset \mathbb{R}^N$, hence the maximal number of memories $P \in \mathbb{N}$, that can be stored via $\mathcal{W}$ and succesfully retrieved by the model. This problem is known as the storage capacity problem, and is somewhat ill-defined, as the definition of storage capacity is usually model-specific (it depends on the activation function of the model and on the set of memories Σ). In what follows, we will provide a general enough definition that can then be adapted to the specific cases.

Definition 1.2.9 (Storage capacity). Let $\delta \geq 0$ and define $d : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ distance function. Let $\Sigma \subset \mathbb{R}^N$ be the set of memories defined in 1.2.7. The storage capacity is defined as the maximal integer $P_{\text{max}} \in \mathbb{N}$ such that

$$\Sigma^\star := \{\xi^1_\star, \dots, \xi^{P_{\text{max}}}_\star\} \subset \mathbb{R}^N \tag{1.11}$$

such that for all $\mu = 1, \dots, P_{\text{max}}$ the following conditions are satisfied:

(i) $\mathbb{0}_N = -\xi^\mu_\star + W\Psi(\xi^\mu_\star)$ is a stable equilibrium.

(ii) there exists a map $\mathcal{D}^\star : \mathbb{R}^N \to \mathbb{R}^N$ such that for each $\mu = 1, \dots, P_{\text{max}}$ we have

$$d(\xi^\mu, \mathcal{D}^\star(\xi^\mu_\star)) \leq \delta. \tag{1.12}$$

(iii) for all $\nu = 1, \dots, P_{\text{max}}$ such that $\nu \neq \mu$

$$d(\xi^\mu, \mathcal{D}^\star(\xi^\nu_\star)) > \delta \tag{1.13}$$

Remark 1.2.10 (**Trivial memory condition**). The third condition on $\mathcal{D}^\star$, for which it must map all equilibria to different points of $\mathbb{R}^N$, is necessary to avoid strange situations. For example, it allows to avoid the choice of a point $x^\star \in \mathbb{R}^N$ that is equidistant from all memories in Σ_H and a function $\mathcal{D}^\star$ mapping $\mathbb{R}^N$ to $x^\star$.

Finally, we point out to the reader that, depending on the specifics of the model and of the distance function, the threshold $\delta \geq 0$ can have different meanings. As we will see, the original discrete-time Hopfield model [27] uses binary $\{-1, 1\}$ memories, the $\psi(x) = \text{sign}(x)$ activation function and the Hamming distance [28]. In this context, the parameter $\delta \geq 0$ regulates the bits of error that are allowed in the retrieval process. Instead, in the classic continuous time Hopfield model [26] the memories are still binary vectors with entries $\{-1, 1\}$, but the activation function is taken to be $\psi(x) = \tanh(x)$ and the distance is a given $\|\cdot\|_p$-norm (usually $p = 2$ or $p = \infty$).

The storage capacity of the classic Hopfield model The storage capacity problem has been a cornerstone of theoretical neuroscience investigation since the seminal papers by Hopfield [26], [27]. The original investigation of the storage capacity [27] exploited the binary definition of the memories, a simplification of the activation function as $\psi(x) = \text{sign}(x)$ and the discretization of the dynamics. We now begin to specialize the original definition of memory vectors 1.2.7 and adapt it to the voltage model.

Definition 1.2.11 (**Hopfield's prototypical memories**)**.** We define the set of Hopfield prototypical memories as

$$\Sigma_{\text{H}} := \{\xi^1, \dots, \xi^P\} \tag{1.14}$$

where $\xi^\mu \in \{-1, 1\}^N$, $\mu = 1, \dots, P$ for $P \in \mathbb{N}$.

In its current definition, the set of Hopfield's prototypical memories is quite general, and is a subset of the set $\{-1, 1\}^N$. Intuitively, among the vectors of the set $\{-1, 1\}^N$ there is no preference for a specific subset. Instead, once the desired cardinality $P \in \mathbb{N}$ of the set Σ_H is fixed, the vectors are randomly sampled from the set $\{-1, 1\}^N$, or a subset of it that satisfies some additional constraints (see later in the chapter discussion on orthogonality and probabilistic constraints).

Remark 1.2.12. Throughout this thesis, we will refer to **prototypical memories** as the vectors in Σ that are stored in the synaptic matrix and that **we would like to retrieve** from the associative memory dynamics. Instead, we will refer to **retrievable memories** as the vectors in $\Sigma^\star$ that the network **is able to retrieve** as fixed points of the associative memory dynamics.

We now define the classic synaptic matric for symmetric voltage models as prescribed in [27].

Definition 1.2.13 (Hopfield's synaptic matrix). Consider the prototypical memory vectors in Σ_{H} defined in (1.14). The synaptic matrix $W \in \mathbb{R}^{N\times N}$ for the Hopfield model is defined as

$$W = \mathcal{W}(\Sigma_{\mathrm{H}}) = \frac{1}{N}\sum_{\mu=1}^{P} \xi^{\mu}\xi^{\mu\top}. \tag{1.15}$$

Notice that the synaptic matrix construction (1.15) is considered an instance of Hebbian learning [29], a biologically plausible learning mechanism that exploits only local information. Indeed, each of the synaptic weights has form $w_{ij} = N^{-1}\sum_{\mu=1}^{P}\xi_i^{\mu}\xi_j^{\mu}$, using only information available at the synapse between neuron i and neuron j.

Example 1.2.14 (**The case of orthogonal memories**). The simplest definition of prototypical memory vectors is the one where all the prototypical memories $\xi^{\mu} \in \Sigma_H$, for $\mu = 1, \dots, P$ are orthogonal to each other and live in the corners of the hypercube $\{-1, 1\}^N$. The constraint that needs to be satisfied is

$$\xi^{\mu\top}\xi^{\nu} = N\delta_{\mu\nu} \qquad \forall \mu, \nu = 1, \dots, P \tag{1.16}$$

The construction of matrices $\bar{H} \in \mathbb{R}^{N\times M}$ containing orthogonal rows and columns of $\{-1, 1\}$ entries, known as Hadamard matrix problem [30], is well known and still posing relevant open questions. In the special case in which there exists $n \in \mathbb{N}$ such that we can express $N = 2^n$, there is an explicit algorithmic procedure to build exactly $N = M$ ortoghonal vectors in $\{-1, 1\}^N$. This is achieved by defining the matrix

$$H = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{1.17}$$

The Kronecher product of H with itself is

$$H \otimes H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}. \tag{1.18}$$

Generalizing to

$$\bar{H} = \underbrace{H \otimes \cdots \otimes H}_{n} \tag{1.19}$$

we obtain exactly an $N \times N$ matrix with orthogonal columns (and rows) in $\{-1, 1\}^N$. Consequently, the maximal storage capacity of the model in this special case is $P_{\max} = N$.

Notice how under the orthogonality assumption, the prototypical memories in Σ_{H} become

eigenvectors of the synaptic matrix W, all with associated eigenvalue $\lambda = 1$.

$$\begin{aligned} W\xi^\nu &= \frac{1}{N}\sum_{\mu=1}^{P} \xi^\mu \xi^{\mu\top} \xi^\nu \\ &= \sum_{\mu=1}^{P} \xi^\mu \frac{N\delta_{\mu\nu}}{N} = \xi^\nu \end{aligned} \tag{1.20}$$

It is then easy to observe that the fixed point condition (1.10) is satisfied for any $\xi^\nu \in \Sigma$.

$$\begin{aligned} \mathbb{0}_N &= -\xi^\nu + W\,\mathrm{sign}(\xi^\nu) \\ &= -\xi^\nu + W\xi^\nu \\ &\overset{(1.20)}{=} -\xi^\nu + \xi^\nu \end{aligned} \tag{1.21}$$

Therefore, we have that prototypical and retrievable memories coincide $\Sigma = \Sigma^\star$ and for any distance it holds that $d(\xi^\nu, \xi^\nu_\star) = 0$.

Remark 1.2.15 (**Fixed points for general activation function**). The generalized activation function of classic Hopfield network is odd about the origin, diagonal, and homogeneous, i.e. $\Psi(x)_i = \psi(x_i)$ with ψ odd. Consequently, for all $\xi^\nu \in \Sigma_{\mathrm{H}}$ we have that $\Psi(\xi^\nu) = \psi(1)\xi^\nu$. Thus, if there exists $\gamma > 0$ such that $\gamma = \psi(\gamma)$, then the set of retrievable memories is $\Sigma^\star_H := \{\gamma\xi^1, \dots, \gamma\xi^P\} = \gamma\Sigma_{\mathrm{H}}$. Indeed, exploiting (1.20) we have for all $\xi^\nu \in \Sigma_{\mathrm{H}}$

$$\begin{aligned} -\gamma\xi^\nu + W\Psi(\gamma\xi^\nu) &= -\gamma\xi^\nu + W\xi^\nu\psi(\gamma) \\ &\overset{\gamma=\psi(\gamma)}{=} -\gamma\xi^\nu + \gamma\xi^\nu = \mathbb{0}_N. \end{aligned} \tag{1.22}$$

Example 1.2.16 (**The case of random memories**). The coefficients of the memory vectors can be randomly sampled from a Bernoulli distribution. Drawing from the statistical physics literature, the original definition of the memories' coefficients was

$$\mathbb{P}(\xi_i^\mu = 1) = \mathbb{P}(\xi_i^\mu = -1) = \frac{1}{2} \qquad \mu = 1, \dots, P,\ i = 1, \dots, N. \tag{1.23}$$

The coefficients of the memory vectors are thus initialized as independently identically distributed (i.i.d.) random variables. Notice that this realization makes the set of memories almost orthogonal to each other, and this will be exploited in the computation of the storage capacity. In addition,

notice that for all $\mu = 1, \dots, P$ and all $i = 1, \dots, N$ we have that

$$\mathbb{E}[\xi_i^\mu] = 1 \cdot \mathbb{P}(\xi_i^\mu = 1) + (-1)\,\mathbb{P}(\xi_i^\mu = -1) = 0 \tag{1.24}$$

$$\mathbb{E}[(\xi_i^\mu)^2] = (1)^2 \cdot \mathbb{P}(\xi_i^\mu = 1) + (-1)^2\,\mathbb{P}(\xi_i^\mu = -1) = 1. \tag{1.25}$$

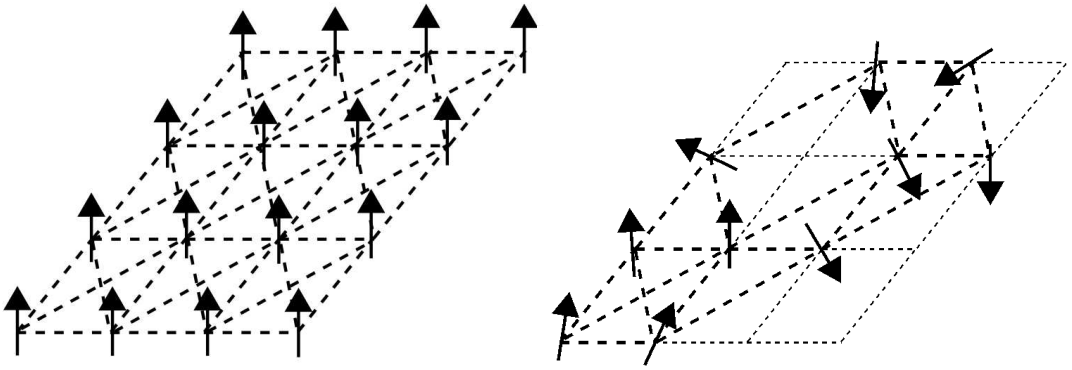


Figure 1.5: **The spin-glass model.** Graphical representation of two spin lattices, one in a ordered state (left) and one in a disordered state (right) [source: Zureks - CC0, https://commons.wikimedia.org/w/index.php?curid=10198960]. Spins (nodes) interact through couplings of heterogeneous sign and magnitude. The resulting frustrated interactions generate a complex energy landscape with numerous local minima. The statistical physics framework provides a natural analogy to associative memory models, where attractor states correspond to stored memories embedded in a high-dimensional space.

The original treatment by Hopfield [27] considered the discontinuous activation function $\psi(x_i) = \text{sign}(x_i)$ and the discretized dynamics

$$x_{\text{H}}(t+1) = \text{sign}\left(W x_{\text{H}}(t)\right). \tag{1.26}$$

Focusing on the equilibrium condition at one of the memory patterns, we observe that each coefficient ξ_i^ν must satisfy

$$\begin{aligned} \xi_i^\nu &= \text{sign}\left([W\xi^\nu]_i\right) \\ &= \text{sign}\left(\xi_i^\nu + \frac{1}{N}\sum_{\mu\neq\nu}^{P} \xi_i^\mu \xi^{\mu\top} \xi^\nu\right). \end{aligned} \tag{1.27}$$

Since $\xi_i^\nu \in \{-1, 1\}$[1], we can multiply both sides of the equation by the same coefficient and

[1]Notice that given $s = \pm 1$, for any $x \in \mathbb{R}$ it holds that $s \cdot \text{sign}(x) = \text{sign}(s \cdot x)$.

obtain the equation

$$1 = \text{sign}\left(1 + \frac{1}{N}\sum_{\mu\neq\nu}\xi_i^\nu \xi_i^\mu \xi^{\mu\top}\xi^\nu\right). \tag{1.28}$$

Therefore, we need $C_i^\nu = \frac{1}{N}\sum_{\mu\neq\nu}\xi_i^\nu\xi_i^\mu\xi^{\mu\top}\xi^\nu > -1$. The original treatment considers C_i^ν as a noise term with simple first and second order moments.

In this example, we will follow the same procedure exploited by Hopfield for the original estimate on the model capacity. Exploiting the fact that the prototypical memories have i.i.d. coefficients, it is easily found that the mean is

$$\begin{aligned}
\mathbb{E}[C_i^\nu] &= \frac{1}{N}\mathbb{E}\left[\sum_{\mu\neq\nu}\xi_i^\nu\xi_i^\mu\sum_{j=1}^N \xi_j^\mu\xi_j^\nu\right] \\
&= \frac{1}{N}\mathbb{E}\left[\sum_{\mu\neq\nu}(\xi_i^\nu)^2(\xi_i^\mu)^2 + \sum_{\mu\neq\nu}\xi_i^\nu\xi_i^\mu\sum_{\substack{j=1\\ j\neq i}}^N \xi_j^\mu\xi_j^\nu\right] \\
&= \frac{P-1}{N} + \frac{1}{N}\sum_{\mu\neq\nu}\sum_{j\neq i}\mathbb{E}[\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu] \\
&\overset{\text{i.i.d.}}{=} \frac{P-1}{N} + \frac{1}{N}\sum_{\mu\neq\nu}\sum_{j\neq i}\mathbb{E}[\xi_i^\nu]\,\mathbb{E}[\xi_i^\mu]\,\mathbb{E}[\xi_j^\mu]\,\mathbb{E}[\xi_j^\nu] = \frac{P-1}{N}.
\end{aligned} \tag{1.29}$$

The second order moment is instead given by

$$\begin{aligned}
\mathbb{E}[(C_i^\nu)^2] &= \frac{1}{N^2}\mathbb{E}\left[\left(\sum_{\mu\neq\nu}\sum_{j=1}^N \xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu\right)^2\right] \\
&= \frac{1}{N^2}\mathbb{E}\left[\sum_{\mu\neq\nu}\sum_{\beta\neq\nu}\sum_{j=1}^N\sum_{l=1}^N \xi_i^\mu\xi_i^\beta\xi_j^\mu\xi_l^\beta\xi_j^\nu\xi_l^\nu\right]
\end{aligned} \tag{1.30}$$

and we address different terms in the sum separately. Consider first the case $\mu = \beta$

$$\begin{aligned}
\mathbb{E}\left[\sum_{\mu\neq\nu}\sum_{j,l=1}^N(\xi_i^\mu)^2\xi_j^\mu\xi_l^\mu\xi_j^\nu\xi_l^\nu\right] &= \mathbb{E}\left[\sum_{\mu\neq\nu}(\xi_j^\mu)^2(\xi_j^\nu)^2 + \sum_{\mu\neq\nu}\sum_{\substack{j=1\\ j\neq l}}^N \xi_j^\mu\xi_l^\mu\xi_j^\nu\xi_l^\nu\right] \\
&\overset{\text{i.i.d.}}{=} N(P-1) + \sum_{\mu\neq\nu}\sum_{j\neq l}^N \mathbb{E}[\xi_j^\mu]\,\mathbb{E}[\xi_l^\mu]\,\mathbb{E}[\xi_j^\nu]\,\mathbb{E}[\xi_l^\nu] \\
&= N(P-1).
\end{aligned} \tag{1.31}$$

Instead, addressing now the case $\mu \neq \beta$

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\substack{\mu\neq\nu}}\sum_{\substack{\beta\neq\nu\\ \beta\neq\mu}}\sum_{j,l=1}^{N}\xi_i^{\mu}\xi_i^{\beta}\xi_j^{\mu}\xi_l^{\beta}\xi_j^{\nu}\xi_l^{\nu}\right]\\
&=\sum_{\substack{\mu\neq\nu}}\sum_{\substack{\beta\neq\nu\\ \beta\neq\mu}}\mathbb{E}\left[(\xi_i^{\mu})^2(\xi_i^{\beta})^2(\xi_i^{\nu})^2+\sum_{j\neq i\ \vee\ l\neq i}\xi_i^{\mu}\xi_i^{\beta}\xi_j^{\mu}\xi_l^{\beta}\xi_j^{\nu}\xi_l^{\nu}\right]\\
&\overset{\text{i.i.d.}}{=}(P-1)(P-2)+\sum_{\substack{\mu\neq\nu}}\sum_{\substack{\beta\neq\nu\\ \beta\neq\mu}}\sum_{j\neq i\ \vee\ l\neq i}\mathbb{E}[\xi_i^{\mu}]\,\mathbb{E}[\xi_i^{\beta}]\,\mathbb{E}[\xi_j^{\mu}]\,\mathbb{E}[\xi_l^{\beta}]\,\mathbb{E}[\xi_j^{\nu}]\,\mathbb{E}[\xi_l^{\nu}]\\
&=(P-1)(P-2)
\end{aligned}\tag{1.32}$$

Since the only term that does not cancel in expectation is the one for the case $j = l = i$. Thus, we finally have that the second order moment is

$$\begin{aligned}
\mathbb{E}[(C_i^{\nu})^2]&=\frac{1}{N^2}\left(\mathbb{E}\left[\sum_{\mu\neq\nu}\sum_{j,l=1}^{N}(\xi_i^{\mu})^2\xi_j^{\mu}\xi_l^{\mu}\xi_j^{\nu}\xi_l^{\nu}\right]+\mathbb{E}\left[\sum_{\substack{\mu\neq\nu}}\sum_{\substack{\beta\neq\nu\\ \beta\neq\mu}}\sum_{j,l=1}^{N}\xi_i^{\mu}\xi_i^{\beta}\xi_j^{\mu}\xi_l^{\beta}\xi_j^{\nu}\xi_l^{\nu}\right]\right)\\
&=\frac{P-1}{N}+\frac{(P-1)(P-2)}{N^2}
\end{aligned}\tag{1.33}$$

Under the assumption of working with very large networks, leading to the following separation of scales $N \mapsto P(N) \gg 1$, we approximate the moments of the noise term as

$$\mathbb{E}\left[C_i^{\nu}\right]\approx\frac{P}{N}=m\quad\mathbb{E}\left[(C_i^{\nu})^2\right]-\mathbb{E}\left[C_i^{\nu}\right]^2\approx\frac{P}{N}=\sigma^2.\tag{1.34}$$

Leveraging the central limit theorem [31, Section 7.4], the noise term can be approximated as

$$C_i^{\nu}=\frac{1}{N}\sum_{j=1}^{N}\underbrace{\sum_{\mu\neq\nu}\xi_i^{\nu}\xi_i^{\mu}\xi_j^{\mu}\xi_j^{\nu}}_{C_{i,j}^{\nu}}\rightharpoonup\omega\sim\mathcal{N}(m,\sigma^2)\tag{1.35}$$

Gaussian variable with mean m and variance σ^2. Thus, the probability of an error in the memory

retrieval process can be approximated as

$$\begin{aligned} p_{err} &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{-1} e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx \\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{1}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx \\ &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}\sigma}\int_{0}^{1} e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx \\ &= \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_{0}^{\sqrt{\frac{N}{P}}-\sqrt{\frac{P}{N}}} e^{-\frac{z^2}{2}}\,dz \end{aligned} \tag{1.36}$$

Fixing then p_{err} to some arbitrary small value, it is possible to use the right-hand side to estimate the maximal $P \in \mathbb{N}$ yielding the desired error. The discussed estimate characterizes a first, approximated answer to the question of the associative memory network storage capacity. Hopfield's initial estimate was for a storage capacity that grows linearly in the number of neurons as $P_{\max} \approx 0.15N$.

Later works [32], [33] grounded the inquiry in probability theory, rigorously proving that $P_{\max} = \frac{1}{4}\frac{N}{\log(N)}$. In the second chapter of this thesis, we will use analogous techniques to study the storage capacity while relaxing the assumptions on the activation function. A complete, self-contained treatment of the *voltage* model can be found in [34].

1.2.2 The *firing rate* model

The *firing rate* model - sometimes presented as neural mass model for whole brain studies - is defined as a continuous-time autonomous O.D.E.

$$\begin{cases} \dot{x}_{\mathrm{F}}(t) = -x_{\mathrm{F}}(t) + \Phi(Wx_{\mathrm{F}}(t) + u(t)), \\ x_{\mathrm{F}}(0) \in \mathbb{R}^N \end{cases} \tag{FR}$$

where the N-dimensional state vector x_{F} contains the firing rates of the neuronal populations, $\dot{x}_{\mathrm{F}}$ denotes its time derivative, $x_{\mathrm{F}}(0) \in \mathbb{R}^N$ is the initial condition for the system and $u(t) \in \mathbb{R}^N$ is the external input. The activation function $\Phi(\cdot)$ is typically assumed to be diagonal and homogeneous, meaning $\Phi(x_{\mathrm{F}})_i = \phi(x_{\mathrm{F}i})$, where $\phi\colon \mathbb{R} \to \mathbb{R}$ is a scalar, weakly increasing function. In *firing rate* systems, the activation function is assumed to be non-negative, meaning $\phi(x_{\mathrm{F}}) \geq 0$ for all $x_{\mathrm{F}} \in \mathbb{R}$. The non-negativity of $\phi(\cdot)$ implies that (FR) is a *positive system*, meaning that the trajectories of (FR) do not leave the positive orthant for non-negative initial conditions: if each entry of the initial condition satisfies $x_{\mathrm{F}i}(0) \geq 0$, then each entry remains non-negative for all time. The positivity of the *firing rate* system has relevant implications, as it allows to study associative memory problems in terms of rate of activity, and makes the model more biologically

plausible than its *voltage* counterpart.

In addition to the positivity of the system, *firing rate* models allow to account for another relevant biological observation: low rates of activation [35], [36]. The *voltage* model as defined by Hopfield has the undesirable requirement of almost orthogonality of the memories. The requirement forces memory vectors to have almost half of the coefficients in the active state 1 and the remaining in the inactive state -1. Observed rates of activation in real neuronal networks are significantly lower ($\sim 0.22N$) [36] and consequently an alternative definition of memories is needed. We specialize the general definition of memories 1.2.7 to the set $\{0,1\}^N$ of positive rates, building the structure of the new memory set through some overlap constraints.

Definition 1.2.17 (Firing rate's prototypical memories). Define the prototypical memories of the *firing rate* model as

$$\Sigma_{\mathrm{F}} := \{\zeta^1, \dots, \zeta^P\} \subseteq \{0,1\}^N \tag{1.37}$$

Remark 1.2.18 (**Structured and random memories**). Similarly to the *voltage* model, the prototypical memories of the *firing rate* model can be defined either deterministically or randomly. Starting from the deterministic case, define an activation parameter $p \in (0,1)$ such that the prototypical memories satisfy the following constraint:

$$\zeta^{\mu\top}\zeta^{\nu} = pN\delta_{\mu\nu} + (1-\delta_{\mu\nu})p^2N. \tag{1.38}$$

In plain terms, the constraint implies that each prototypical memory has exactly pN elements that are equal to 1, and all the other are 0. Similarly, each prototypical memory shares with all the other in the set Σ_{F} exactly p^2N ones. Perhaps the most trivial method to build a set Σ_{F} satisfying constraint (1.38) is to fix $N \in \mathbb{N}$ the size of the network and $P \in \mathbb{N}$ the number of memories to be stored. Then setting the activation parameter as $p = 1/(P-1)$ and defining the matrix $H \in \mathbb{R}^{N\times P}$ having the prototypical memories as column as

$$H = \begin{bmatrix} \mathbb{1}_{p^2N}\mathbb{1}_P^\top \\ I_P \\ \vdots \\ I_P \end{bmatrix}$$

with the identity matrices being repeated $p(1-p)N$ times. Then the columns of H provide a set of $P = \lfloor 1 + 1/p \rfloor$ memories satisfying the desired constraints. Greater number of memories $P = P(p)$ can be generated using more advanced techniques. The general problem of determining the maximum number of sets of equal size that share the same number of elements pairwise is

known as the *Balanced Incomplete Block Design (BIBD)* [37].

In the literature, prototypical memories for *firing rate* systems have been defined as vectors ζ^μ, for $\mu = 1, \dots, P$, with i.i.d. entries

$$\mathbb{P}(\zeta_i^\mu = 1) = 1 - \mathbb{P}(\zeta_i^\mu = 0) = p \tag{1.39}$$

where $p \in (0, 1)$ is the activation paramater. Notice that from this definition we have that the constraint (1.38) is satisfied in expectation.

$$\mathbb{E}\left[\zeta^{\mu\top}\zeta^\nu\right] = pN\delta_{\mu\nu} + (1 - \delta_{\mu\nu})p^2N. \tag{1.40}$$

Given the definition of the new prototypical memory vectors, it is possible to vary the parameter p to enforce the desired rate of activity to study memory retrieval in a more biologically plausible context. To accommodate the new definition of prototypical memory vectors, new forms of the synaptic matrix $W \in \mathbb{R}^{N\times N}$ have been proposed [38], [39]. The new synaptic matrix builds on the outer product rule in (1.15) and adds both bias and compensatory terms.

Definition 1.2.19 (Firing rate synaptic matrix). Consider the prototypical memory vectors in Σ_F defined in (1.37) and the activation parameter $p \in (0, 1)$. The synaptic matrix $W \in \mathbb{R}^{N\times N}$ for the *firing rate* model is

$$W = \mathcal{W}(\Sigma_F) = \frac{1}{p(1-p)N}\sum_{\mu=1}^{P}\left(\zeta^\mu - p\mathbb{1}_N\right)\left(\zeta^\mu - p\mathbb{1}_N\right)^\top - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top \tag{1.41}$$

Remark 1.2.20. Notice that we have an additional term $-\frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top$, whose purpose is to stabilize the network activity around the expected rate p. Indeed, at any time the contribution of that term is

$$-\mathbb{1}_N\left(\frac{\mathbb{1}_N^\top x_F(t)}{N}\right) \tag{1.42}$$

which results in the uniform subtraction of the average network activity at time t from the synaptic contribution.

Differently from the *voltage* model, not many formal properties of the *firing rate* model have been studied so far. As a matter of fact, a general characterization of the set of retrievable memories $\Sigma_{FR}^\star$ is missing in the literature. Most studies focus on the statistical mechanical derivation of the overlap parameters, i.e. how much the state of the *firing rate* model aligns with the prototypical memory vectors. These estimates are considerably valuable, as they allow to ascertain that the model is indeed able to retrieve P memories, where $P \in \mathbb{N}$ is the usual storage capacity. However, the techniques of statistical mechanics rely on the large $N \to \infty$ limit and

other approximations (expansion of the activation function, saddle point approximation) and are therefore unable to provide exact answers to classic questions in dynamical system theory. In the third chapter of this thesis, we will address the gaps in the literature of *firing rate* models by providing rigorous answers to the issues of fixed points condition- deriving a closed formula for the retrievable memories - and studying their local stability and global stability.

1.2.3 Bridging *voltage* and *firing rate* models

The relationship between the *voltage* and *firing rate* models has been investigated in prior work [40], [41]. These studies demonstrate that the two models[2] can be made equivalent through appropriate transformations of state and input. The authors show that when W is invertible, there exists a bijective map that allows to go from the *voltage* to *firing rate* formulation of the model, and viceversa. Let

$$x_{\mathrm{H}} = W x_{\mathrm{F}} + u \qquad \dot{u} + u = I \tag{1.43}$$

where $I \in \mathbb{R}^N$ is the input to the *voltage* model. Then taking the time derivative of (1.43) we obtain

$$\begin{aligned} \dot{x}_{\mathrm{H}} &= W\dot{x}_{\mathrm{F}} + \dot{u} \\ &= W\left(-x_{\mathrm{F}} + \Phi(W x_{\mathrm{F}} + u)\right) + I - u \\ &= -(W x_{\mathrm{F}} + u) + W\Phi(W x_{\mathrm{F}} + u) + I \\ &= -x_{\mathrm{H}} + W\Phi(x_{\mathrm{H}}) + I. \end{aligned} \tag{1.44}$$

It is equally easy to derive the *firing rate* dynamics from the *voltage* dynamics.

$$\begin{aligned} \dot{x}_{\mathrm{F}} &= W^{-1}(\dot{x}_{\mathrm{H}} - \dot{u}) \\ &= W^{-1}(-x_{\mathrm{H}} + W\Phi(x_{\mathrm{H}}) + I - I + u) \\ &= -W^{-1}(x_{\mathrm{H}} - u) + \Phi(x_{\mathrm{H}}) \\ &= -x_{\mathrm{F}} + \Phi(W x_{\mathrm{F}} + u) \end{aligned} \tag{1.45}$$

However, when W is low-rank, as is typical in associative memory contexts, the relationship between the models becomes less direct and more nuanced. Specifically, the equivalence is derived by projecting the *firing rate* dynamics onto the subspace spanned by the columns of W, a constraint absent in the original dynamics. Moreover, the authors do not address the challenge of designing the synaptic matrix W to achieve retrievable memory patterns as stable equilibria for

[2] In these works, the *voltage* model is presented with dynamical equations $\dot{v} = -v + W\Phi(v)$, while the *firing rate* model is expressed as $\dot{r} = -r + \Phi(Wr)$. Both models are grouped under the umbrella of *firing rate* systems, assuming they share a common positive activation function $\Phi()$ and the same synaptic matrix W.

the *firing rate* system. Finally, bridging the dynamics of the two models introduces a time-varying external input, complicating the use of standard Lyapunov methods to establish convergence to equilibria. Despite the highlighted differences, it is still common to find literature that considers the two models equivalent even when they are not to justify claims about their properties. The independent investigation of the *firing rate* model proposed in the third chapter of this thesis aims at providing additional validity and rigour to support such claims.

1.3 Associative memory models and generative AI

Grossberg and Hopfield's seminal works [26], [42] introduced an energy-based recurrent network capable of storing binary patterns as attractors in a dynamical landscape: the *voltage* model. Retrieval occurred through convergence to a stored attractor, offering a simple yet powerful metaphor for memory recall. However, classical Hopfield networks are constrained by limited storage capacity and are difficult to train via classic Hebbian learning [29]. Consequently, they have been limited in applications and often presented as an historical note in many machine learning courses. Nonetheless, associative memory models established the central idea that memories are stable, retrievable states of a dynamical system—a perspective that continues to influence both machine learning and neuroscience.

Subsequent generalizations, often referred to as dense associative memories (DAM) [43], enriched the energy function by replacing quadratic interactions with higher-order or exponential forms. This modification dramatically increased the storage capacity, reaching exponential scaling in the dimension of the state space. Importantly, these networks blurred the line between memory and computation, as the retrieval dynamics resembled attention-like selection over stored representations typical of Transfomers architecture [44]. Tranformers model have been the dominant technology behind much of the progress in artificial intelligence and machine learning in recent years, and have finally allowed to deploy Ai solutions in many domestic and commercial contexts (e.g. ChatGPT). Recent work by Ramsauer et al. [45] established a formal equivalence between the update rule of continuous-state modern Hopfield networks and the key–value attention mechanism of transformers. This result reframes transformer layers as associative memories: queries correspond to partial patterns, keys represent stored patterns, and attention weights implement a soft retrieval rule that converges to fixed points of the energy landscape. Consider the Energy function for the dense Hopfield model

$$\mathrm{E}_D(x_{\mathrm{H}}) = \frac{1}{2}\|x_{\mathrm{H}}\|^2 - \frac{1}{\beta}\log\left(\sum_{\mu=1}^{P} e^{\beta(M^\top x_{\mathrm{H}})_\mu}\right) \tag{1.46}$$

where $\beta > 0$ is called temperature parameter and regulates the steepness of the exponential. The

matrix $M \in \mathbb{R}^{N\times P}$ instead contains the vectors to encode and retrieve, i.e. the memories, and has form $M = [\xi^1, \dots, \xi^P]$[3]. The associative memory dynamics associated the Energy E_D are

$$\begin{aligned} \dot{x}_{\mathrm{H}} &= -\nabla \mathrm{E}_D(x_{\mathrm{H}}) \\ &= -x_{\mathrm{H}} + M\mathrm{softmax}_\beta(M^\top x_{\mathrm{H}}) \end{aligned} \tag{1.47}$$

where $\mathrm{softmax}_\beta(x)_\nu = \frac{e^{\beta(M^\top x)_\nu}}{\sum_{\mu=1}^{P} e^{\beta(M^\top x)_\mu}}$. Passing to the discretized dynamics, it is immediate to observe that the dense Hopfield model instantiate a self-attention mechanism.

$$x_{\mathrm{H}}(t+1) = M\mathrm{softmax}_\beta(M^\top x_{\mathrm{H}}). \tag{1.48}$$

Depending on the separation of prototypical memories, the system can converge to single memories, averages over subsets, or global averages. Specifically, if prototypical memory patterns are similar enough, hence not well separated, then the filtering $\mathrm{softmax}_\beta(M^\top x_{\mathrm{H}})$ will return a probability vector assigning (almost) equal probabilities to all of them. Consequently, the vector field (1.47) will point towards the weighted average of the similar memory patterns. Finally, the connection between associative memory models and transformers model has been further explored in [46], where the energy is expressed as a function of the standard Key K and Query Q matrices regulating the input-output relationship. This perspective provides a unifying framework for deep learning architectures. Hopfield layers, introduced as memory-equipped components, can replace pooling, recurrent, or attention layers, enabling explicit memory retrieval within standard pipelines. They have been shown to improve performance in domains ranging from multiple instance learning to fluid-dynamics [47], demonstrating that associative memory is not merely a historical curiosity but a modern computational primitive.

Building on this foundation, the rapid success of generative diffusion models, transformers, and hybrid architectures reveals that generative AI itself can be understood as an instantiation of associative memory at scale. Rather than treating recall and creativity as separate processes, these models suggest they are facets of the same underlying mechanism—convergence within an energy landscape defined by learned associations. Ambrogioni [48] demonstrated that the energy function governing diffusion models is asymptotically equivalent to that of modern Hopfield networks. In this framework, the forward diffusion process scatters memories across a noisy latent space, associating them to the moments of standard Gaussian distributions. Instead, the reverse process reconstitutes them by descending the associative energy landscape, generating samples similar to the desired memory (see Fig. 1.6). Generation thus becomes probabilistic memory recall, enriched by stochasticity that allows reconstructive and imaginative outputs

[3] Notice that for $M \in \{-1, 1\}^{N\times P}$ encoding the memories of the original Hopfield model, the synaptic matrix $W \in \mathbb{R}^{N\times N}$ can be written as $W = \frac{1}{N} MM^\top$

rather than mere rote retrieval. This resonates with psychological theories of memory as a generative, reconstructive process, where recall integrates stored traces with contextual information. Biological plausibility adds another dimension to this synthesis. Kozachkova, Slotine, and

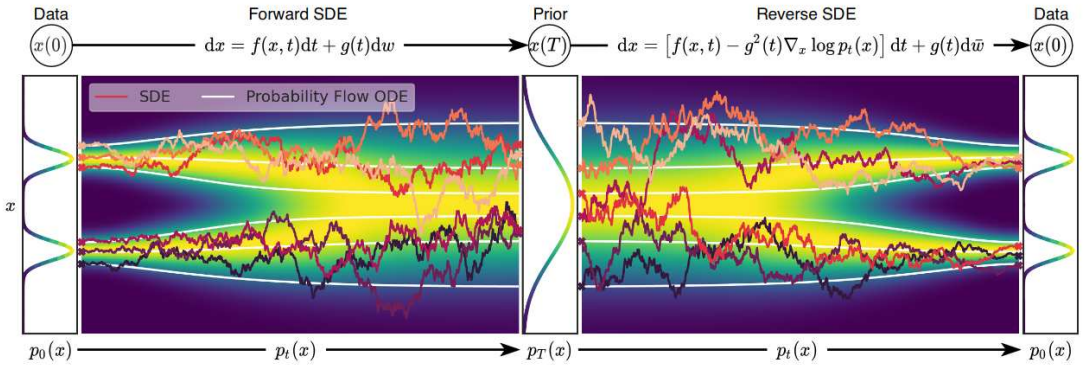


Figure 1.6: **The diffusion process in generative AI.** Schematic representation of the generative dynamics of a diffusion model (adapted from [49]). The forward diffusion process progressively perturbs an input prompt until it matches a simple, well-characterized distribution such as Gaussian noise. The generative phase then reverses this trajectory: starting from the terminal distribution, the model iteratively denoises and refines the sample, converging to an output that belongs to the same class as the initial prompt. This bidirectional view emphasizes how diffusion models unify stochastic noise processes with structured pattern generation.

Krotov [4] proposed neuron–astrocyte associative networks, showing that astrocytic processes substantially increase memory capacity and naturally implement dense associative dynamics. In essence, the authors defined a system of neurons $x \in \mathbb{R}^N$, synapses $s \in \mathbb{R}^{N\times N}$, and astrocytes $p \in \mathbb{R}^{N\times N}$, whose dynamics depend on a unique energy function. The composite energy $\mathrm{E} : \mathbb{R}^N \times \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times N} \to \mathbb{R}$ function has form

$$\mathrm{E}(x,s,p) = \mathrm{E}^p(p) + \mathrm{E}^{xs}(x,s) + \mathrm{E}^{sp}(s,p) \tag{1.49}$$

where $\mathrm{E}^{xs}(x,s)$ is the energy associated to the neural dynamics, $\mathrm{E}^{xs}(x,s) + \mathrm{E}^{sp}(s,p)$ the energy associated to the synaptic dynamics, and $\mathrm{E}^p(p) + \mathrm{E}^{sp}(s,p)$ the energy associated to the astrocyte process. The dynamics of the neuron-astrocyte architecture are then defined as

$$\tau_x \dot{x} \propto \frac{\partial \mathrm{E}}{\partial x}(x,s,p) \tag{1.50}$$

$$\tau_s \dot{s} \propto \frac{\partial \mathrm{E}}{\partial s}(x,s,p) \tag{1.51}$$

$$\tau_p \dot{p} \propto \frac{\partial \mathrm{E}}{\partial p}(x,s,p) \tag{1.52}$$

These networks can interpolate between transformer-like architectures and dense associative memories, suggesting that the computational strategies discovered in AI echo structures long embedded in brain physiology (see Fig. 1.7 for a graphic intuition of the neuron-astrocyte interaction). Importantly, astrocytes communicate across vastly different timescales and spatial domains, offering a plausible substrate for the integration of long-term stability with rapid context-dependent retrieval. Bringing these perspectives together, generative AI can be reframed as the

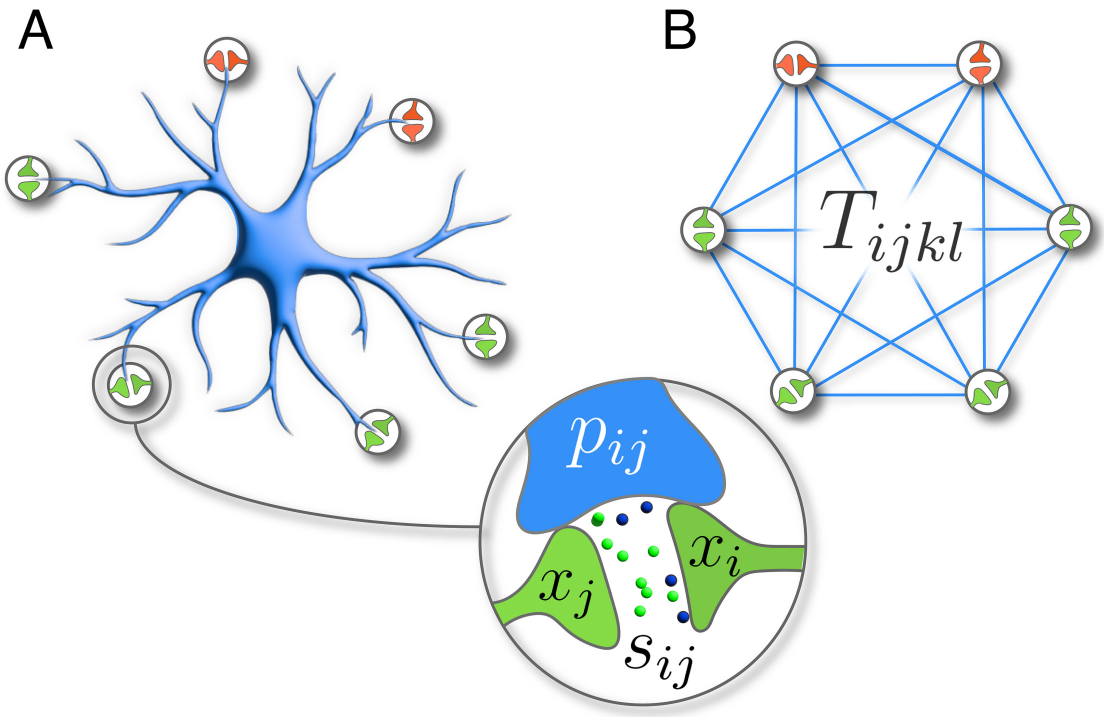


Figure 1.7: **The neuron-astrocyte model.** Schematic of the neuron–astrocyte associative memory model (adapted from [4]). The figure illustrates how astrocytes, forming tripartite synapses with neurons, modulate synaptic efficacy and thereby influence information storage and retrieval. Neural firing dynamics operate on millisecond timescales, synaptic states evolve more slowly, and astrocytic calcium waves unfold on yet longer timescales. The interaction of these processes offers a biologically plausible mechanism for the simultaneous implementation of short-term memory (mediated by neural activity) and long-term memory (stabilized through astrocyte-mediated modulation) within the same circuit.

computational continuation of associative memory. Transformers highlight the attention–memory equivalence, diffusion models extend this equivalence to stochastic generative dynamics, and neuron–astrocyte systems anchor it in biology. The synthesis points toward a unifying principle: generative processes in artificial systems, like imagination in humans, emerge from the dynamics of memory systems.

The implications are manifold. Architecturally, energy-based associative layers could serve

as fundamental components for future generative models, providing controllable retrieval, flexible generalization, and interpretability. Conceptually, this view unifies memory and creativity, dissolving the historical divide between storage and generation in AI. Biologically, it underscores the relevance of glial networks, long overlooked in neuroscience, as computational resources shaping the brain's generative capacities. And theoretically, it suggests that progress in generative AI may depend less on designing new algorithms ex nihilo than on deepening our understanding of associative dynamics across artificial and biological substrates. In summary, the integration of diffusion equivalence, astrocyte-inspired architectures, and modern Hopfield formulations reveals generative AI as an emergent consequence of associative memory principles. This perspective not only advances machine learning but also forges tighter links to cognitive science and neuroscience, pointing to a future in which memory and imagination are recognized as two sides of the same computational coin.

2

On the storage capacity of the *voltage* model

In the introduction of this thesis we highlighted a central theme in the study of associative memory models: understanding their storage capacity. Up to that point, however, the discussion remained intentionally focused on a narrow class of systems, namely discrete-time networks endowed with the $\mathrm{sign}(\cdot)$ activation function. This choice allowed us to present the classical perspective with clarity, but it leaves open the broader landscape of models that extend beyond this discrete-time formulation and step activation function.

Hopfield originally proposed two complementary descriptions of the same retrieval mechanism: the discrete-time nonlinear system [27] emphasised in the introductory chapter, and the continuous-time formulation [26], whose dynamics evolve according to an underlying energy function. The discrete-time version has traditionally served as the main framework for analysing storage capacity P, understood as the maximal number of patterns that can be reliably stored and retrieved [33], [50], [51]. The continuous-time counterpart, by contrast, has been approached primarily through stability analysis of its energy landscape, leveraging the negative time derivative of the energy along trajectories to establish convergence to equilibria. Crucially, despite their methodological differences, both formulations satisfy the same fixed-point condition. This structural equivalence allows us to translate insights between them and motivates a broader examination of associative memory models beyond the specialized setting presented earlier in the thesis.

This shared foundation naturally leads to a central and enduring question: How many memories can the *voltage* model store, and to what extent does this depend on the choice of activation function? As discussed in the previous chapter, Hopfield's original central limit theorem argument suggested a capacity of $P \approx 0.15N$ for binary patterns in $-1, +1$, allowing retrieval with small errors [27]. However, subsequent rigorous work [33] demonstrated that this estimate fails for both perfect and imperfect retrieval. The capacity question has recently gained renewed relevance, as modern *voltage* architectures [25], [43] have reestablished dynamical systems as a foundation for machine learning, achieving exponential capacities [52] and revealing conceptual

links to transformers [45] and diffusion models [48].

Yet, despite the abundance of results for the classical $\mathrm{sign}(\cdot)$ activation, the analysis of storage capacity has not been systematically extended to broader classes of activation functions. As we will show in this chapter, pursuing such generality comes at a cost: it substantially complicates the extension of existing capacity arguments and requires new analytical approaches.

The chapter proceeds as follows. Section 2.1 reviews the continuous-time *voltage* model and summarises existing theoretical results. Section 2.2 extends capacity analysis to the case of activation functions that saturate to finite asymptotic values. Section 2.3 further generalises the discussion to arbitrary activation functions, and numerical experiments reveal the possible existence of a phase transition in the model's dynamics.

Notation. The symbol sign means the sign function that maps negative reals into -1, zero into zero and positive reals into $+1$. We allow this map to be applied to vectors in a component-wise manner. The symbol $\mathbb{1}$ means the vector with all ones, while I means the identity matrix. For two real vectors x, y of the same dimension, $x \geq y$ means that $x_i \geq y_i$ for all i. If x is a vector, then $\mathrm{diag}(x)$ stands for the diagonal matrix having the entries of x on the diagonal. Given a matrix $A \in \mathbb{R}^{n\times n}$, we denote with $A^\top$ its transpose. In case A is symmetric, $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ denote its minimum and its maximum eigenvalue. The symbol $\|\cdot\|_\infty$ means the infinity norm of a vector or the induced infinity norm for a matrix. The symbol $\mathbb{P}[\omega]$ denotes the probability of the event ω while the symbol $\mathbb{E}[X]$ denotes the expected value of the random variable X.

2.1 The continuous-time *voltage* model

Consider the continuous-time *voltage* model (V) described by the nonlinear dynamical system

$$\tau \dot{x}_{\mathrm{H}}(t) = -x_{\mathrm{H}}(t) + W\Psi(x_{\mathrm{H}}(t)), \qquad x_{\mathrm{H}}(0) \in \mathbb{R}^N, \tag{2.1}$$

and its discrete-time counterpart

$$x_{\mathrm{H}}(t+1) = W\Psi(x_{\mathrm{H}}(t)) \tag{2.2}$$

where $x_{\mathrm{H}}(t) \in \mathbb{R}^N$ denotes the neuron state vector at time $t \in \mathbb{R}$, $\tau > 0$ is the time constant, $W \in \mathbb{R}^{N\times N}$ is the synaptic matrix, and $\Psi : \mathbb{R}^N \to \mathbb{R}^N$ is the activation function. The initial condition $x_{\mathrm{H}}(0)$ determines which stored memory is retrieved [27]. In addition, the synaptic matrix W satisfies the symmetry assumption 1.2.1

Assumption 2.1.1 (Diagonality, homogeneity and asymptotics of the activation function). The activation function Ψ is diagonal, i.e. its i-th component $i = 1, ..., N$, depends only on the i-th component of x, $\Psi_i(x_{\mathrm{H}}) = \Psi_i([x_{\mathrm{H}}]_i)$. Furthermore, assume that $\exists\psi : \mathbb{R} \to \mathbb{R}$ such that

$\Psi_i(z) \equiv \psi(z)$ for all $i = 1, \dots, N$. Finally, the map ψ is assumed to be continuous (or C^k, for $k \geq 1$ when specified), with $\lim_{z \to \pm\infty} \psi(z) = \pm 1$ and such that $\psi(z) = -\psi(-z)$ for all $z \in \mathbb{R}$.

Under assumptions 1.2.1, we have seen that there exists an energy function [26] for (2.1), namely a function

$$\mathrm{E}_H(x_{\mathrm{H}}) = -\frac{1}{2}\Psi(x_{\mathrm{H}})^\top W \Psi(x_{\mathrm{H}}) + x_{\mathrm{H}}^\top \Psi(x_{\mathrm{H}}) - \sum_{i=1}^{N} \int_0^{[x_{\mathrm{H}}]_i} \Psi_i(z)\, dz \tag{2.3}$$

of the state whose total time derivative is monotonically decreasing along the system trajectories, namely

$$\frac{d}{dt}\mathrm{E}_H(x_{\mathrm{H}}(t)) < 0. \tag{2.4}$$

Since the energy function is monotonically decreasing along the trajectories of the system (2.1), the behavior of this system is quite regular, that is, its solutions converge to equilibria each with its respective basin of attraction.

In this chapter the energy function will be used to prove the existence of stable equilibria within well-defined forward invariant regions of state space. To this aim, we begin with the characterization of the synaptic matrix.

Assumption 2.1.2 (Structure of the synaptic matrix). Define the set of memories

$$\Sigma_H := \{\xi^1, \dots, \xi^P\} \tag{2.5}$$

as in (1.23), hence with random i.i.d. components satisfying

$$\mathbb{P}[\xi_i^\mu = 1] = \mathbb{P}[\xi_i^\mu = -1] = \frac{1}{2} \qquad \forall \mu = 1, \dots, P,\ \forall i = 1, \dots, N. \tag{2.6}$$

We define the Hopfield-type synaptic matrix 1.15 as

$$W = \frac{1}{N} \sum_{\mu=1}^{P} \xi^\mu \xi^{\mu\top} \tag{2.7}$$

We now provide a definition for the capacity of a continuous *voltage* model, specializing the general definition 1.2.9 of Chapter 1.

Definition 2.1.3 (Capacity for retrieval without errors). Under the same notation of definition 1.2.9, let $\mathcal{D}^\star : \mathbb{R}^N \to \mathbb{R}^N$ be such that $\mathcal{D}_i^\star(x) = \mathrm{sign}(x_i)$ for all $i = 1, \dots, N$, and let $d : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be any distance in $\mathbb{R}^N$. Let $\Sigma_H := \{\xi^1, \dots, \xi^P\}$ be as in Assumption 2.1.2. We say that the continuous-time *voltage* model (2.1) correctly stores the memories in Σ_H if for

all $\xi^\mu \in \Sigma_H$ there exists a stable equilibrium $\xi^\mu_\star$ for (2.1) such that

$$d(\xi^\mu, \mathcal{D}^\star(\xi^\mu_\star)) = 0 \tag{2.8}$$

i.e., $\xi^\mu_\star$ and ξ^μ have the same sign pattern.

In the following, we will rigorously investigate the problem of the *voltage* model capacity for different choices of activation functions. Specifically, we will first extend well-known results [33] to a generalization of the sign function, and then proceed to further generalize the result to the class of smoothly saturating functions of Assumption 2.1.1.

2.2 Capacity for s-saturated activation functions

In this section, we show that the capacity analysis for the discrete-time *voltage* model can be quite easily extended to the continuous-time case if we impose a further restriction on the activation function.

Definition 2.2.1 (s-saturated activation function)**.** We say that that ψ in Assumption 2.1.1 is a **s-saturated** function if it is continuous and there exists $s > 0$ such that $\psi(\pm z) = \pm 1$ for all $z \geq s$.

From this point onward, we will suppose that Assumptions 1.2.1, 2.1.1, 2.1.2, and 2.1.1, hold and we introduce some preliminary concepts. We define for all $\nu = 1, \dots, P$ the event

$$\begin{aligned}\mathcal{A}(N, P, \nu) := \{\omega \in \Omega \mid \exists \text{ a stable equilibrium } \xi^\nu_\star \text{ of (2.1)} \\ \text{such that } \operatorname{sign}(\xi^\nu_\star) = \xi^\nu\},\end{aligned} \tag{2.9}$$

that is the event "the model correctly stores the memory ξ^ν", and the event

$$\mathcal{A}(N, P) = \bigcap_{\nu=1}^{P} \mathcal{A}(N, P, \nu) \tag{2.10}$$

that coincides with the event in which the model correctly stores all the memories.

Theorem 2.2.2 (Storage capacity for s-saturated activation functions). *Consider the continuous-time* voltage *model* (2.1)*. Assume moreover that the activation function ψ is s-saturated with $s < 1$, where s is the parameter appearing in Definition 2.2.1. If*

$$P \leq \frac{(1-s)^2}{4} \frac{N}{\log(N)} \tag{2.11}$$

then

$$\lim_{N\to\infty} \mathbb{P}[\mathcal{A}(N,P)] = 1. \tag{2.12}$$

Proof. For all $\nu = 1,\dots,P$, let $\xi^\nu_\star := W\xi^\nu$. Observe now that, if $\mathrm{diag}(\xi^\nu)\xi^\nu_\star \geq s\mathbb{1}$, then $\mathrm{sign}(\xi^\nu_\star) = \xi^\nu$ and $\Psi(\xi^\nu_\star) = \xi^\nu$ and hence $-\xi^\nu_\star + W\Psi(\xi^\nu_\star) = -\xi^\nu_\star + W\xi^\nu = -\xi^\nu_\star + \xi^\nu_\star = 0$, which implies that $\xi^\nu_\star$ is an equilibrium. Moreover, observe that, since $\psi'(\pm z) = 0$ for all $z \geq s$, then $\psi'([\xi^\nu_\star]_i) = 0$ and hence the Jacobian of the dynamics in the equilibrium is $-I$, so that we can argue that the equilibrium $\xi^\nu_\star$ is locally stable (attractive). In this way we proved that

$$\mathbb{P}[\mathcal{A}(N,P)] \geq \mathbb{P}[\mathrm{diag}(\xi^\nu)\xi^\nu_\star \geq s\mathbb{1},\ \forall\nu = 1,\dots,P]. \tag{2.13}$$

It remains to prove that this last probability tends to 1. By the union bound [33] and applying Lemma 2.4.1 in the Appendix, we can argue that

$$\begin{aligned}
&\mathbb{P}[\mathrm{diag}(\xi^\nu)\bar{x}^\nu \geq s\mathbb{1},\ \forall\nu = 1,\dots,P] \\
&= \mathbb{P}[\cap_{\nu=1}^P \{\omega \mid \mathrm{diag}(\xi^\nu)\xi^\nu_\star \geq s\}] \\
&= \mathbb{P}[\cap_{\nu=1}^P \cap_{i=1}^N \{\omega \mid [\mathrm{diag}(\xi^\nu)\xi^\nu_\star]_i \geq s\}] \\
&= 1 - \mathbb{P}[\cup_{\nu=1}^P \cup_{i=1}^N \{\omega \mid [\mathrm{diag}(\xi^\nu)W\xi^\nu]_i < s\}] \\
&\geq 1 - \sum_{\nu=1}^P \sum_{i=1}^N \mathbb{P}[[\mathrm{diag}(\xi^\nu)W\xi^\nu]_i < s] \\
&\geq 1 - PN\exp\left(-(1-s)^2\frac{N}{2P}\right).
\end{aligned} \tag{2.14}$$

Using (2.11) we can further bound the previous equation as follows

$$\begin{aligned}
&\mathbb{P}[\mathrm{diag}(\xi^\nu)\xi^\nu_\star \geq s\mathbb{1},\ \forall\nu = 1,\dots,P] \\
&\geq 1 - \frac{(1-s)^2}{4}\frac{N^2}{\log N}\exp\left(-\frac{N}{2}\frac{4}{(1-s)^2}\frac{\log N}{N}(1-s)^2\right) \\
&= 1 - \frac{(1-s)^2}{4}\frac{1}{\log N},
\end{aligned} \tag{2.15}$$

which yields the thesis. □

2.3 Capacity for general activation functions

In this section we will try to extend the analysis presented in Sec. 2.2 to the case where the activation function is not s-saturated. This extension is more intricate than expected, and hints at the existence of a phase transition, as shown by the following numerical experiments.

We selected $\psi(z) = \tanh(az)$, where a is a positive real parameter. Then, for any choice of this parameter a and of the model dimension N, we generated 25 different matrices W randomly according to Assumption 2.1.2 with $P = \frac{N}{4\log(N)}$. Then we evaluated the fraction of the memories that are correctly stored by the model averaged over the 25 instances. This can be considered as an estimate of the probability that a memory is correctly stored by the model as a function of the parameters a and N. The plot in Figure 2.1 refers to the choice of $N \in [100, 2900]$ with intervals of 200 and $a \in [1.1\ , 2]$ with intervals of 0.1. We notice that, for small a the probability decreases with N, while for large a, namely when the associated activation function approaches the sign function, the probability increases with N and tends to one, as evident from Figure 2.2. What is more interesting is the fact that for large N there seems to be a sharp transition from small probabilities to probability one as a increases. As it can be observed from Figure 2.3, there exists a threshold around $a = 1.4$ such that above this value the retrieval probability grows with N. We now try to capture this exact phenomenology using analytical techniques.

To prove the existence of the previously defined stable equilibria $\{\xi^\nu_\star\}_\nu$, we need to show that for any memory ξ^ν we can find a forward invariant set contained in the orthant

$$\begin{aligned} Q^\nu &:= \{x_H \in \mathbb{R}^N \ : \ \text{sign}(x_H) = \xi^\nu\} \\ &= \{x_H \in \mathbb{R}^N \ : \ x_H = \text{diag}(\xi^\nu) z, \ z \in \mathbb{R}^N_{\geq 0}\}. \end{aligned} \tag{2.16}$$

Let

$$\xi^\nu_\star := \beta W \xi^\nu, \tag{2.17}$$

$$e^\nu := W\xi^\nu - \xi^\nu, \tag{2.18}$$

where $\beta > 0$ is such that $\psi(\beta) = \beta$. Observe that

$$\xi^\nu_\star = \beta\xi^\nu + \beta e^\nu. \tag{2.19}$$

Moreover, for any $r > 0$, define

$$\begin{aligned} Q^\nu_r &:= \{x_H \in \mathbb{R}^N \ : \ x_H = \xi^\nu_\star - r\xi^\nu + z, \ z \in Q^\nu\} \\ &= \{x_H \in \mathbb{R}^N \ : \ x_H = \xi^\nu_\star - r\xi^\nu + \text{diag}(\xi^\nu) z, \ z \in \mathbb{R}^N_{\geq 0}\}. \end{aligned} \tag{2.20}$$

We want to estimate the probability that there exists $r > 0$ such that Q^ν_r is forward invariant.

Lemma 2.3.1 (Forward-invariance). *If* $r := P(1 - \psi(s))$ *with* $0 < s < \beta$, *then*

$$\mathbb{P}\left[\{Q^\nu_r \textit{ forward invariant}\}\right] \geq \mathbb{P}\left[\text{diag}(\xi^\nu) e^\nu \geq -\frac{\beta - s - r}{\beta} \mathbb{1}\right]. \tag{2.21}$$

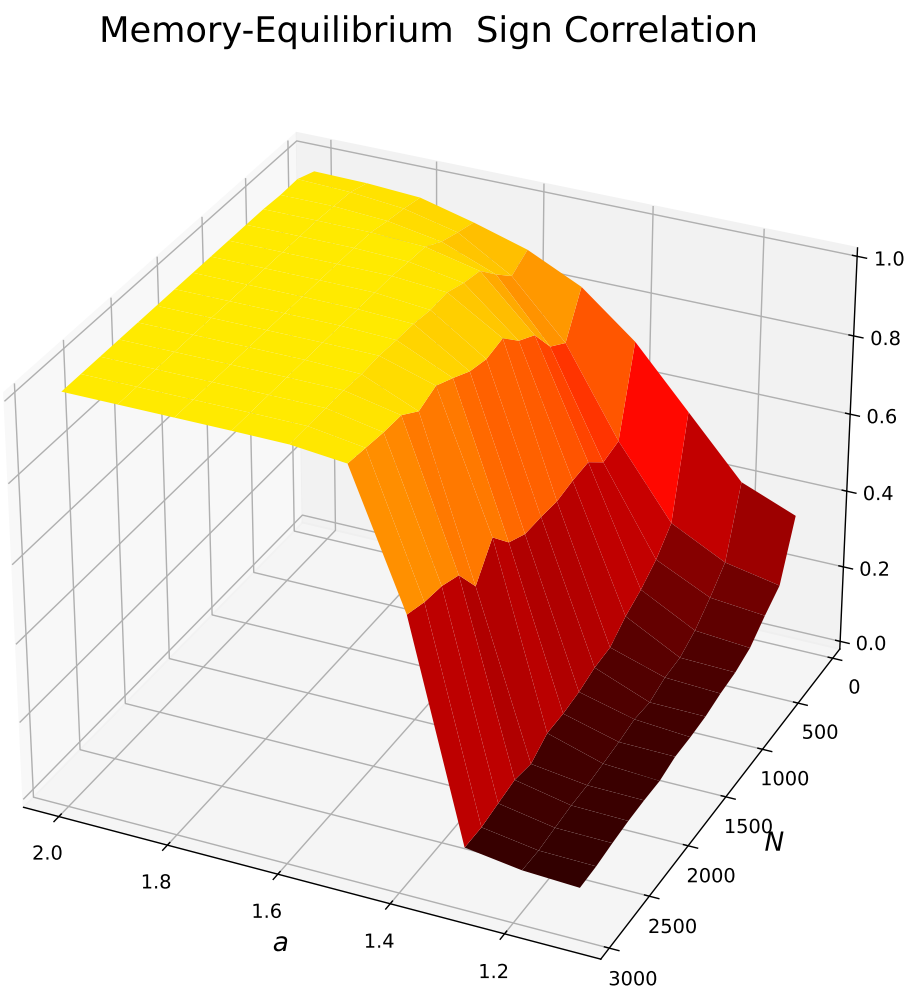


Figure 2.1: **Fraction of successful retrievals as the slope a and number of units N vary.** This figure shows the fraction of the memories that are correctly stored by the model as a function of model dimension N and of the parameter a (activation function $\psi(z) = \tanh(az)$). Then for any choice of $a \in \{1.1 + 0.1 * k\},\ k = 0, \dots, 9$ and $N \in \{100 + 200 * w\},\ w = 0, \dots, 15$ we generated randomly 25 different W according to Assumption 2.1.2 with $P = \frac{N}{4\log(N)}$. The surface displays the average fraction of correctly stored memories.

Proof. To prove the thesis it is enough to prove that, if $r = P(1 - \psi(s))$ and

$$\operatorname{diag}(\xi^{\nu})e^{\nu} \geq -\frac{\beta - s - r}{\beta}\mathbb{1}, \tag{2.22}$$

then Q_r^{ν} is forward invariant. By Nagumo's theorem [53] it is enough to prove that, if $r = P(1 - \psi(s))$ and if (2.22) holds, then for any x_{H} belonging to the boundary of Q_r^{ν}, the vector

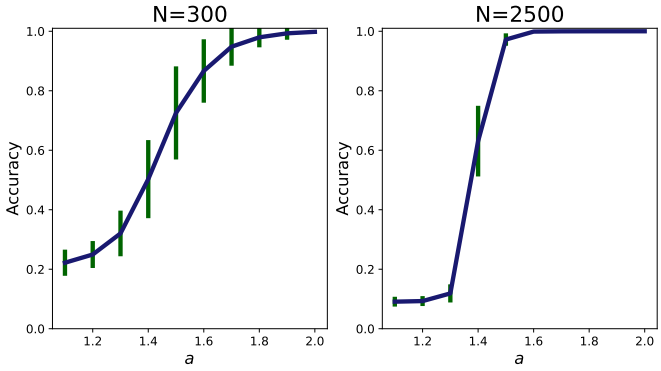


Figure 2.2: **Fraction of successful retrievals at fixed N.** In this figure, the blue line shows the fraction of the memories that are correctly stored by the model as a function of the parameter a appearing in the activation function $\psi(z) = \tanh(az)$ when $N = 300$ and $N = 2500$. The green bars represent the standard deviation across the 25 trials.

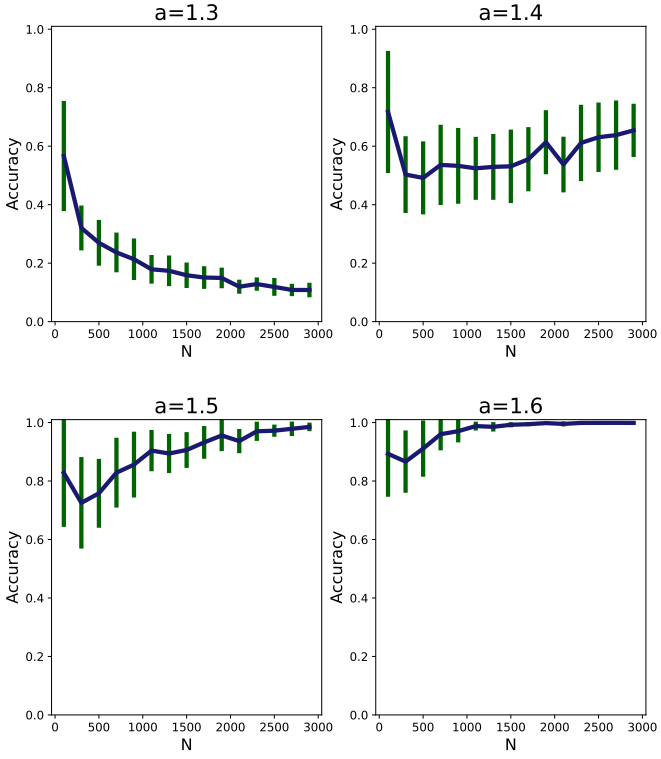


Figure 2.3: **Fraction of succssesfull retrievals at fixed slope a.** In this figure, the blue line shows the fraction of the memories that are correctly stored by the model as a function of the model dimension N when $a = 1.3$, $a = 1.4$, $a = 1.5$ and $a = 1.6$. The green bars represent the standard deviation across the 25 trials.

field $f(x_{\mathrm{H}}) = -x_{\mathrm{H}} + W\Psi(x_{\mathrm{H}})$ of (2.1) points inside the set Q_r^{ν} or, more precisely, that for any

$$x_{\mathrm{H}} = \xi_\star^\nu - r\xi^\nu + \mathrm{diag}(\xi^\nu)z \tag{2.23}$$

with $z \in \mathbb{R}^N_{\geq 0}$, $z_i = 0$, we have that $\xi_i^\nu f_i(x) \geq 0$. Observe that, if (2.22) holds, then

$$\begin{aligned}\mathrm{diag}(\xi^\nu)x_{\mathrm{H}} &= \mathrm{diag}(\xi^\nu)[\beta\xi^\nu + \beta e^\nu - r\xi^\nu + \mathrm{diag}(\xi^\nu)z] \\ &= \beta\mathbb{1} + \beta\mathrm{diag}(\xi^\nu)e^\nu - r\mathbb{1} + z \\ &\geq \beta\mathrm{diag}(\xi^\nu)e^\nu + (\beta - s - r)\mathbb{1} + s\mathbb{1} \geq s\mathbb{1}. \end{aligned} \tag{2.24}$$

This implies that $\Psi(\mathrm{diag}(\xi^\nu)x_{\mathrm{H}}) \in [\psi(s), 1]^N$ and hence $\Psi(\mathrm{diag}(\xi^\nu)x_{\mathrm{H}}) - \Psi(\beta\mathbb{1}) \in [-(\psi(\beta) - \psi(s)), 1 - \psi(\beta)]^N \subseteq [-(1 - \psi(s)), 1 - \psi(s)]^N$. This implies that $u := \Psi(x_{\mathrm{H}}) - \Psi(\beta\xi^\nu) \in [-(1 - \psi(s)), 1 - \psi(s)]^N$ and hence

$$\begin{aligned}\mathrm{diag}(\xi^\nu)f(x_{\mathrm{H}}) &= \mathrm{diag}(\xi^\nu)[-\xi_\star^\nu + r\xi^\nu - \mathrm{diag}(\xi^\nu)z + W\Psi(x_{\mathrm{H}})] &(2.25)\\ &= \mathrm{diag}(\xi^\nu)[-\xi_\star^\nu + r\xi^\nu - \mathrm{diag}(\xi^\nu)z + W(\beta\xi^\nu + u)] \\ &= \mathrm{diag}(\xi^\nu)[r\xi^\nu - \mathrm{diag}(\xi^\nu)z + Wu] \\ &= r\mathbb{1} - z + \mathrm{diag}(\xi^\nu)Wu. &(2.26)\end{aligned}$$

Since we know that $z_i = 0$, and observing that

$$\begin{aligned}\|W\|_\infty &= \max_{i=1,\dots,N} \frac{1}{N}\sum_{\mu=1}^{P}\sum_{j=1}^{N}\underbrace{|\xi_i^\mu \xi_j^\mu|}_{\equiv 1} \\ &= \frac{PN}{N} = P \end{aligned} \tag{2.27}$$

then

$$\begin{aligned}\xi_i^\nu f_i(x_{\mathrm{H}}) &= r + (\mathrm{diag}(\xi^\nu)Wu)_i \geq r - \|W\|_\infty\|u\|_\infty \\ &\geq r - P(1 - \psi(s)) = 0. \end{aligned} \tag{2.28}$$

This concludes the proof. □

Notice that if (2.22) holds, from the previous lemma we have that $\mathrm{diag}(\xi^\nu)x_{\mathrm{H}} \geq s\mathbb{1}$, and consequently $Q_r^\nu \subseteq Q^\nu$. Additionally, from the previous lemma we also have that Q_r^ν is forward invariant.

$$\mathrm{diag}(\xi^\nu)e^\nu \geq -\frac{\beta - s - r}{\beta}\mathbb{1} \quad \Rightarrow \quad Q_r^\nu \subseteq Q^\nu \text{ and } Q_r^\nu \text{ forward invariant} \tag{2.29}$$

The next result provides a bound on the number of correctly stored memories for the class of activation functions in Assumption 2.1.1, where now we require at least first order differentiability, i.e. $\psi \in C^k(\mathbb{R})$ for $k \geq 1$.

Theorem 2.3.2 (Storage capacity for non-saturated activation functions). *Consider the continuous-time* voltage *model* (2.1). *Assume there exists* $\beta > 0$ *such that* $\psi(\beta) = \beta$ *and fix any* $0 < s < \beta$ *such that* $\psi(s) < 1$. *If*

$$P < \frac{\beta - s}{1 - \psi(s)} \tag{2.30}$$

then

$$\lim_{N\to\infty} \mathbb{P}[\mathcal{A}(N, P)] = 1. \tag{2.31}$$

Proof. Let $r = P(1 - \psi(s))$ and

$$\epsilon := \frac{\beta - s - r}{\beta}. \tag{2.32}$$

Observe that by (2.30) we can argue that $\epsilon > 0$. Moreover, using Lemma 2.3.1 we have that if

$$\operatorname{diag}(\xi^\nu) e^\nu \geq -\epsilon \mathbb{1} \tag{2.33}$$

then Q_r^ν is forward invariant and $Q_r^\nu \subseteq Q^\nu$ (2.29). This holds if and only if

$$\operatorname{diag}(\xi^\nu) W \xi^\nu \geq (1 - \epsilon)\, \mathbb{1}. \tag{2.34}$$

This implies that for all $\nu = 1, \ldots, P$ the following set inequality holds

$$\{Q_r^\nu \text{ forw. inv. and } Q_r^\nu \subseteq Q^\nu\} \supseteq \{\operatorname{diag}(\xi^\nu) W \xi^\nu \geq (1 - \epsilon)\, \mathbb{1}\}. \tag{2.35}$$

Since there exists an energy function $\mathrm{E}_H(x_\mathrm{H})$ for (2.1) whose total time derivative is monotonically decreasing along the system trajectories and since it can be proved that this energy is radially unbounded, we can argue that under the realization of the forward invariance events of Q_r^ν there exists a stable equilibrium in Q_r^ν. In view of the latter fact and from the set relation (2.29), it follows that

$$\mathcal{A}(N, P, \nu) \supseteq \{Q_r^\nu \text{ forw. inv. and } Q_r^\nu \subseteq Q^\nu\}. \tag{2.36}$$

From (2.35), (2.36) and Lemma 2.4.1 in the Appendix we have

$$\begin{aligned} \mathbb{P}[\mathcal{A}(N, P, \nu)] &\geq \mathbb{P}[\{\operatorname{diag}(\xi^\nu) W \xi^\nu \geq (1 - \epsilon)\, \mathbb{1}\}] \\ &\geq 1 - N \exp\left(-\epsilon^2 \frac{N}{2P}\right). \end{aligned} \tag{2.37}$$

Finally, from the fact that $\mathcal{A}(N, P) = \cap_\nu\, \mathcal{A}(N, P, \nu) \supseteq \cap_\nu \{Q_r^\nu \text{ forw. inv. and } Q_r^\nu \subseteq Q^\nu\}$ and

the union bound

$$\begin{aligned}\mathbb{P}[\mathcal{A}(N,P)] &\geq \mathbb{P}[\cap_{\nu}\{Q_r^{\nu} \text{ forw. inv. and } Q_r^{\nu} \subseteq Q^{\nu}\}] \\ &\geq 1 - NP\exp\left(-\epsilon^2\frac{N}{2P}\right). \end{aligned} \tag{2.38}$$

For any fixed $P \in \mathbb{N}$, the previous inequality implies

$$\lim_{N\to\infty} \mathbb{P}[\mathcal{A}(N,P)] = 1. \tag{2.39}$$

This concludes the proof. □

Some comments on Theorem 2.3.2 are in order. First, it is important to notice that the more saturating is the activation function $\psi(\cdot)$, the larger is the bound on the number P of correctly stored memories in condition (2.30). More precisely, for any fixed s such that $0 < s < \beta$, the right-hand side of (2.30) becomes larger as $\psi(s)$ approaches one, which is the saturation level. Second, when P grows with the network size N, Theorem 2.3.2 is able to capture only the sub-threshold behavior illustrated in Figure 2.1. In fact, it can be shown that the bound on P in (2.30) cannot grow as $N/\log(N)$ if the activation function is fixed. Thus, the supra-threshold behaviour of Figure 2.1, where the retrieval accuracy grows as $N/\log(N)$, cannot be captured by Theorem 2.3.2.

2.4 Conclusion

In this chapter, we have undertaken an initial exploration of the storage capacity of the continuous-time *voltage* model, building upon insights from prior studies on its discrete-time counterpart. We have shown that the rigorous results established for the discrete-time case extend naturally to the continuous-time setting when the activation functions are of the saturated type.

To go beyond this restricted class, we investigated smoothly saturating activation functions through numerical experiments. These simulations revealed evidence for a critical slope threshold in the activation function, beyond which a qualitative change in retrieval performance emerges. Specifically, our results suggest the existence of a phase transition in the large-network limit $N \to +\infty$: for slopes above the threshold, retrieval performance scales positively with N, while for slopes below it, the capacity drops sharply.

Our theoretical analysis in the general case remains incomplete, as it is restricted to a fixed memory load P. Extending the results to the regime where P grows with N forces the admissible activation functions back to the saturated case—an outcome at odds with the supra-threshold behaviour observed in simulations. This discrepancy between theoretical predictions

and numerical evidence calls for the development of new analytical tools capable of capturing the mechanisms underlying the phase transition.

This chapter is based on the conference proceeding [54] for the *IEEE* 63^{rd} *Conference on Decision and Control (CDC)*. The following chapter takes a different perspective, turning to the biologically inspired *firing rate* model. By moving beyond the *voltage* formulation, we explore how principled synaptic design can enforce fixed-point existence, and how stability analysis—both local and global—can shed light on mechanisms that strengthen robustness in memory retrieval.

Appendix In this paragraph we present an instrumental result which is used in the chapter.

Lemma 2.4.1 (Bound on the equilibrium error). *Assume that we have random memory vectors* $\xi^1, \dots, \xi^p \in \{+1, -1\}^n$. *Assume moreover that* $\mathcal{G}$ *is the deterministic graph as in Example 1 and that* $W_{\mathcal{G}}$ *is a matrix obtained from* $\mathcal{G}$ *as in* (2.7). *Then, for any* $\epsilon \geq 0$,

$$\mathbb{P}[\mathrm{diag}(\xi^\nu) W \xi^\nu \geq (1-\epsilon)\mathbb{1}] \geq 1 - N \exp\left(-\epsilon^2 \frac{N}{2P}\right).$$

Proof. We define first the event $B(N, P, \nu)$ as follows

$$B(N, P, \nu) := \{\omega \in \Omega \mid \mathrm{diag}(\xi^\nu) W \xi^\nu \geq (1-\epsilon)\mathbb{1}\}. \tag{2.40}$$

We give now an estimate of the probability of $B(N, P, \nu)$ following the arguments presented in [33]. First, observe that $B(N, P, \nu) = \cap_{i=1}^N B(N, P, \nu, i)$ where

$$B(N, P, \nu, i) := \left\{ \omega \in \Omega \;\middle|\; \frac{1}{N} \xi_i^\nu \sum_{\mu=1}^P \xi_i^\mu \sum_{j=1}^N \xi_j^\mu \xi_j^\nu \geq 1 - \epsilon \right\}$$

The probability of its complement, that is $\overline{B}(N, P, \nu, i) := \{\omega \in \Omega \mid \xi_i^\nu \sum_{\mu=1}^P \xi_i^\mu \sum_{j=1}^N \xi_j^\mu \xi_j^\nu < N(1-\epsilon)\}$, can be bounded as follows

$$\begin{aligned} \mathbb{P}\left[\overline{B}(N, P, \nu, i)\right] &= \mathbb{P}\left[\sum_{\mu=1}^P \sum_{j=1}^N \xi_i^\nu \xi_i^\mu \xi_j^\mu \xi_j^\nu < N(1-\epsilon)\right] \\ &= \mathbb{P}\left[\sum_{\mu \neq \nu} \sum_{j \neq i} \xi_i^\nu \xi_i^\mu \xi_j^\mu \xi_j^\nu + N + P - 1 < N(1-\epsilon)\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left[\exp\left(-t\sum_{\mu\neq\nu}\sum_{j\neq i}\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu\right) > \exp(t(N\epsilon+P-1))\right] \\
&\leq \frac{\mathbb{E}\left[\exp\left(-t\sum_{\mu\neq\nu}\sum_{j\neq i}\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu\right)\right]}{\exp(t(N\epsilon+P-1))},
\end{aligned}
\tag{2.41}
$$

where the equality holds for any $t \geq 0$ and where the last inequality follows from the Markov inequality [33]. We now estimate the expected value in the previous equation. Observe that

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(-t\sum_{\mu\neq\nu}\sum_{j\neq i}\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu\right)\right] \\
&= \mathbb{E}\left[\prod_{\mu\neq\nu}\prod_{j\neq i}\exp(-t\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{\mu\neq\nu}\prod_{j\neq i}\exp(-t\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu)\,\middle|\,\xi_j^\mu \text{ with } \mu=\nu \text{ or } j=i\right]\right] \\
&= \mathbb{E}\left[\prod_{\mu\neq\nu}\prod_{j\neq i}\mathbb{E}\left[\exp(-t\xi_i^\nu\xi_i^\mu\xi_j^\mu\xi_j^\nu)\,\middle|\,\xi_j^\mu \text{ with } \mu=\nu \text{ or } j=i\right]\right] \\
&= \mathbb{E}\left[\prod_{\mu\neq\nu}\prod_{j\neq i}\cosh(t\xi_i^\nu\xi_i^\mu\xi_j^\nu)\,\middle|\,\xi_j^\mu \text{ with } \mu=\nu \text{ or } j=i\right] \\
&= \mathbb{E}\left[\prod_{\mu\neq\nu}\prod_{j\neq i}\cosh(t\xi_i^\nu\xi_i^\mu\xi_j^\nu)\right].
\end{aligned}
$$

Notice now that $\xi_j^\nu, \xi_i^\mu, \xi_j^\mu \in \{\pm 1\}$ and, since the function $\cosh(\cdot)$ is even, then $\cosh(t\xi_j^\nu\xi_i^\mu\xi_j^\mu) = \cosh(t)$. This implies that

$$
\begin{aligned}
\mathbb{E}\left[\prod_{\mu\neq\nu}\prod_{j=1}^{N}\cosh(t\xi_i^\nu\xi_i^\mu\xi_j^\nu)\right] &= (\cosh(t))^{(N-1)(P-1)} \\
&\leq \left(\exp(t^2/2)\right)^{(N-1)(P-1)} \\
&= \exp((N-1)(P-1)t^2/2)),
\end{aligned}
$$

where we used the fact that $\cosh(t) \leq \exp(t^2/2)$. In this way we obtain that for all $t \geq 0$

$$
\mathbb{P}\left[\overline{B}(N,P,\nu,i)\right] \leq \exp\left((N-1)(P-1)\frac{t^2}{2} - t(N\epsilon+P-1)\right).
$$

Taking the infimum with respect to the parameter $t \geq 0$, which is attained at

$$t = -\frac{N\epsilon + P - 1}{(N-1)(P-1)},$$

we obtain that

$$\mathbb{P}\left[\overline{B}(N, P, \boldsymbol{\nu}, i)\right] \leq \exp\left(-\frac{(N\epsilon + P - 1)^2}{2(N-1)(P-1)}\right).$$

Hence, we have that

$$\begin{aligned}
\mathbb{P}[\overline{B}(N, P, \boldsymbol{\nu})] &= \mathbb{P}[\cup_{i=1}^{N} \overline{B}(N, P, \boldsymbol{\nu}, i)] \\
&\leq \sum_{i=1}^{N} \mathbb{P}[\overline{B}(N, P, \boldsymbol{\nu}, i)] \\
&\leq N \exp\left(-\frac{(N\epsilon + P - 1)^2}{2(N-1)(P-1)}\right).
\end{aligned}$$

Since

$$\frac{(N\epsilon + P - 1)^2}{2(N-1)(P-1)} \leq \epsilon^2 \frac{N}{2P}$$

then we have the thesis. □

3

Synaptic design and stability in the *firing rate* model

The *firing rate* model has long been regarded as a more biologically grounded description of neural computation, thanks to its non-negativity and its capacity to produce sparse patterns of activity. Although closely related to the voltage-based Hopfield model, its mathematical structure is considerably more delicate: the interplay between the activation function and the synaptic matrix complicates both analytical treatment and the explicit synthesis of networks that store prescribed memories as stable equilibria. As a consequence, much of the existing work on *firing rate* models relies on heuristics or approximate constructions, and formal results on their stability properties remain comparatively scarce.

The development of associative memory models in theoretical neuroscience traces back to the early dynamical formulations of Amari [55], [56], Grossberg [42], and the constructive energy-based framework of Hopfield [26], [27]. These models framed memory retrieval as a dynamical process governed by ODEs, with stored patterns realized as attracting equilibria. A Lyapunov function ensured convergence, and the use of binary $-1, +1$ encodings allowed for explicit synaptic constructions tailored to specific memories. Subsequent work generalised this setting to positive activation functions [57] and $0, 1$ encodings [38], [58], [59], maintaining the dynamical perspective while moving closer to biologically observable firing profiles. Yet, in these models the link between voltages and firing rates remained indirect and dependent on model parameters, prompting a shift towards descriptions in which firing rates themselves form the core dynamical variables.

Alongside these developments, detailed biophysical models of cortical circuits grew increasingly influential, driven by their ability to generate synthetic EEG-like signals and thereby illuminate the mechanisms underlying experimental observations. Since the seminal work of Hodgkin and Huxley [24], dynamical systems theory has provided a unifying language for describing neuronal processes at multiple scales. Early contributions focused on microscopic

biophysics, such as ion channel dynamics and membrane conductances, yielding celebrated models like FitzHugh–Nagumo [60] and Morris–Lecar [61]. Their analytical richness, however, was counterbalanced by considerable mathematical and computational complexity, restricting early studies to small networks. This limitation motivated the development of simplified yet expressive neuron models, most notably the family of Integrate-and-Fire (IF) models [62]–[65], which enabled both large-scale simulations and partial analytical insight.

A few pioneering works [66], [67] extended the IF formalism to include covariance-based synaptic mechanisms, facilitating the use of statistical mechanics tools in the study of memory retrieval. Still, the hybrid continuous-discrete nature of IF dynamics makes a rigorous stability analysis challenging, keeping the theory of associative memory in these models relatively underdeveloped.

This naturally led to the adoption of time-averaged descriptions of neural activity. Averaging spike trains over suitable temporal windows yields firing rates [36], [68], which can be driven directly by non-negative activation functions rather than derived indirectly from membrane voltages. The resulting *firing rate* model captures population-level dynamics in a tractable and biologically meaningful way. Yet its mathematical foundations remain thin: the systematic design of *firing rate* networks whose synaptic matrices and activation functions ensure that given patterns are locally stable equilibria is still an open challenge [39, Section 7.4].

While several classical studies explored associative memory with positive activation functions [38], [57]–[59], the majority analysed networks storing random memories and relied heavily on statistical mechanics. In contrast, the present chapter develops a dynamical systems perspective on deterministic memories, drawing on the framework outlined by Dayan and Abbott [39, Section 7.4]. This approach allows us to move beyond heuristic constructions and establish formal results on both the existence and the stability of equilibria in *firing rate* networks.

The contributions of this chapter are twofold. First, we introduce a constructive method for designing synaptic matrices in *firing rate* models that encode memories - specifically, rescaled versions of a given set of prototypical patterns - as equilibrium points, applicable to arbitrary activation functions. These prototypical memories are assumed to be equally sparse and equally correlated binary vectors, sharing the same number of active components and overlapping entries. We prove a necessary and sufficient condition for the existence of such equilibria, and interpret the resulting synaptic structure in terms of well-established biological mechanisms: long-term potentiation (LTP), long-term depression (LTD), and homeostatic regulation. Notably, we show that the classical construction of Dayan & Abbott emerges as a special case of our framework. We further demonstrate that the phenomenon of "anti-memories" can occur only in degenerate cases reducible to *voltage*-type synaptic matrices, and we analyse the possible emergence of spurious equilibria, with special attention to homogeneous states.

Second, we establish sufficient conditions for the local asymptotic stability of the stored memories. Leveraging classical results by Grossberg [42] and Hopfield [26], we construct an energy function to investigate the global behaviour of trajectories. Numerical experiments assess the tightness of these stability conditions and illustrate the corresponding energy landscapes for representative examples. These simulations reveal that negative homeostatic strength parameters lead to larger stability regions in parameter space, consistent with canonical choices in the literature.

3.1 Equilibria assignment through synaptic weights

Consider the *firing rate* model (FR).

$$\begin{cases} \dot{x}_{\mathrm{F}} = -x_{\mathrm{F}} + \Phi(W x_{\mathrm{F}} + u(t)) \\ x_{\mathrm{F}}(0) \in \mathbb{R}^N \end{cases} \tag{FR}$$

with $u(t) \equiv \mathbb{0}_N$. From now on we will assume that the activation function satisfies the following hypothesis.

Assumption 3.1.1 (Positive and monotonic activation functions)**.** The activation function $\Phi(\cdot)$ in (FR) is diagonal and homogeneous, i.e., there exists $\phi\colon \mathbb{R} \to \mathbb{R}$ such that $\Phi(x)_i = \phi(x_i)$ for all $x \in \mathbb{R}^N$. Moreover, the function $\phi(\cdot)$ is continuous, non-negative, and weakly increasing.

The prototypical memories Given a set of $\{0,1\}$-valued vectors in $\mathbb{R}^N$

$$\{\zeta^\mu\}_{\mu=1}^P \tag{3.1}$$

corresponding to *prototypical memory* patterns with either active or inactive units, we are interested in designing biologically plausible synaptic matrices W such that an appropriately scaled version of these vectors are equilibria of the *firing rate* dynamics (FR). We consider prototypical memories $\{\zeta^\mu\}_{\mu=1}^P$ satisfying the following hypothesis.

Assumption 3.1.2 (Equally sparse and correlated memories). Given an *average activity parameter* $p \in (0,1)$, a correlation parameter $r \in (0,1)$ the prototypical memories $\{\zeta^\mu\}_{\mu=1}^P$, $\zeta^\mu \in \{0,1\}^N$ satisfy

$$\text{(equal sparsity):} \qquad \mathbb{1}_N^\top \zeta^\mu = pN, \quad \mu = 1, \dots, P, \tag{3.2a}$$

$$\text{(equal correlation):} \qquad {\zeta^\mu}^\top \zeta^\nu = prN, \quad \forall\, \mu \neq \nu. \tag{3.2b}$$

The conditions in Assumption 3.1.2 for $p = r$ are inspired by the choice of prototypical memories in the classic *voltage* [69], [70] and *firing rate* models [39, Section 7.4]. Theoretical results are first established in full generality and then specialized to the case $p = r$, a scenario of historical importance in the literature and particular relevance for biological interpretation. In probabilistic terms, the parameter p is related to the average activity of the network in each memory pattern assuming that these activities are uncorrelated. The equal sparsity (3.2a) and equal correlation (3.2b) constraints refer to the simplest statistical structure that memory patterns can have. As a matter of fact, if we assume that the entries of prototypical memories $\zeta_h^\mu \in \{0, 1\}$, $\mu = 1, \dots, P$, $h = 1, \dots, N$ are i.i.d. random binary variables such that the probability that $\zeta_h^\mu = 1$ is equal to p for all h, μ, then the conditions (3.2a) and (3.2b) are satisfied in expectation. Importantly, unlike the case of random memories, Assumption 3.1.2 protects the *firing rate* model from the emergence of endogenous noise and improves its analytical tractability.

From prototypical memories and selected firing rates to retrievable memories. Select low and high firing rates $x_{\mathrm{F}}^0 < x_{\mathrm{F}}^1 \in \mathbb{R}$ both belonging to the range of the activation function ϕ, and associate to each prototypical memory $\zeta^\mu \in \mathbb{R}^N$ a *retrievable memory* $\bar{\zeta}^\mu \in \mathbb{R}^N$ defined by

$$\bar{\zeta}_i^\mu = \begin{cases} x_{\mathrm{F}}^0, & \text{if } \zeta_i^\mu = 0, \\ x_{\mathrm{F}}^1, & \text{if } \zeta_i^\mu = 1. \end{cases} \tag{3.3}$$

In vector format, $\bar{\zeta}^\mu := (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta^\mu + x_{\mathrm{F}}^0 \mathbb{1}_N$.

The firing rates x_{F}^0 and x_{F}^1 are the neural manifestation of the activity of population of neurons experiencing different input currents [71]. The input currents are the incoming activity that each neuron processes and filters by means of its activation function. Therefore, we can mathematically relate input currents and firing rates in the following way. Let $\mathrm{I}_0 \in \mathbb{R}$ represent a weak input current, and $\mathrm{I}_1 \in \mathbb{R}$ a strong input current, so that $\mathrm{I}_0 < \mathrm{I}_1$. Then the states of the neurons associated with low and high firing rates are $x_{\mathrm{F}}^0 = \phi(\mathrm{I}_0)$ and $x_{\mathrm{F}}^1 = \phi(\mathrm{I}_1)$, respectively.

Remark 3.1.3 (**On the linear independence of the memories**). Exploiting the geometric constraints placed on the prototypical memories, namely the equal sparsity and equal correlation constraints, it is possible to prove the linear independence of both the prototypical memory vectors and the retrievable memory vectors. See the Appendix of the current chapter for more details.

From prototypical memories to synaptic matrices. To a set of prototypical memories $\{\zeta^\mu\}_{\mu=1}^P$ satisfying the sparsity and correlation Assumption 3.1.2 with average activity p and equal

correlation r, we associate an $N \times N$ dimensional *covariance-based synaptic matrix*

$$W = \frac{\alpha}{p(1-r)N} \sum_{\mu=1}^{P} (\zeta^\mu - \beta \mathbb{1}_N)(\zeta^\mu - \beta \mathbb{1}_N)^\top + \frac{\gamma}{N} \mathbb{1}_N \mathbb{1}_N^\top, \tag{3.4}$$

where α, β, $\gamma \in \mathbb{R}$ and they are interpreted as follows.

- the *correlation strength* $\alpha \in \mathbb{R}$ which modulates The covariance part of W given by the outer products of the shifted prototypical memories.
- The *synaptic bias* $\beta \in \mathbb{R}$ which imposes an offset on the synaptic weights and biases the network activity towards the value p.
- The *homeostatic strength* $\gamma \in \mathbb{R}$ which controls the magnitude of the component of the covariance-based synaptic matrix that modulates the synaptic response proportionally to the global network activity. This contribution to the global activity is referred to as homeostatic from the electrochemical balancing mechanism observed in real neuronal networks. Typically, homeostatic terms are discussed in the literature [35], [39] as global inhibitory terms, with homeostatic strength $\gamma \leq 0$.

Given the definitions of covariance-based synaptic matrix and retrievable memories, the following theorem characterizes when the retrievable memories are equilibria of the *firing rate* system (FR).

Theorem 3.1.4 (Equilibria assignment through covariance-based synaptic weights). *Consider the* firing rate *model* (FR) *with activation functions satisfying the positivity and monotonicity Assumption 3.1.1. Consider prototypical memories* $\{\zeta^\mu\}_{\mu=1}^P$ *satisfying the sparsity and correlation Assumption 3.1.2 with average activity* p.

If there exist currents $\mathrm{I}_0 < \mathrm{I}_1$, *with corresponding low and high firing rates* $x_F^0 := \phi(\mathrm{I}_0)$, $x_F^1 := \phi(\mathrm{I}_1)$, *and correlation, bias, and homeostatic strengths* α, β, *and* γ *satisfying*

$$\alpha := \frac{\mathrm{I}_1 - \mathrm{I}_0}{x_F^1 - x_F^0}, \tag{3.5a}$$

$$\beta := p \frac{r x_F^1 + (1-r) x_F^0}{p x_F^1 + (1-p) x_F^0}, \tag{3.5b}$$

$$\gamma := \frac{\beta \mathrm{I}_1 + (1-\beta)\mathrm{I}_0}{p x_F^1 + (1-p) x_F^0}. \tag{3.5c}$$

then each retrievable memory $\bar{\zeta}^\mu$ *is an equilibrium point of the* firing rate *model* (FR) *with the covariance-based synaptic matrix* (3.4).

Proof. Using the definitions of $\bar{\zeta}^\nu$ and W in (3.4) and (3.3), it is a matter of lengthy but straightforward calculations to check that the following identity holds

$$\begin{aligned} W\bar{\zeta}^\nu &= \frac{\alpha}{p(1-r)N}\left(\zeta^\nu - \beta\mathbb{1}_N\right)\left(\zeta^\nu - \beta\mathbb{1}_N\right)^\top\left(x_{\mathrm{F}}^0\mathbb{1}_N + (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta^\nu\right) \\ &+ \frac{\alpha}{p(1-r)N}\sum_{\mu\neq\nu}\left(\zeta^\mu - \beta\mathbb{1}_N\right)\left(\zeta^\mu - \beta\mathbb{1}_N\right)^\top\left(x_{\mathrm{F}}^0\mathbb{1}_N + (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta^\nu\right) \\ &+ \frac{\gamma}{N}\mathbb{1}_N\mathbb{1}_N^\top(x_{\mathrm{F}}^0\mathbb{1}_N + (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta^\nu) \\ &= (\mathrm{I}_1 - \mathrm{I}_0)\zeta^\nu + \mathrm{I}_0\mathbb{1}_N. \end{aligned}$$

From the above identity, it follows that

$$\begin{aligned} \Phi(W\bar{\zeta}^\nu) &= \Phi((\mathrm{I}_1 - \mathrm{I}_0)\zeta^\nu + \mathrm{I}_0\mathbb{1}_N) \\ &= (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta^\nu + x_{\mathrm{F}}^0\mathbb{1}_N = \bar{\zeta}^\nu, \end{aligned} \tag{3.6}$$

where we used the fact that the entries of the vector $(\mathrm{I}_1 - \mathrm{I}_0)\zeta^\nu + \mathrm{I}_0\mathbb{1}_N$ takes values in $\{\mathrm{I}_0, \mathrm{I}_1\}$ and the identities $x_{\mathrm{F}}^0 = \phi(\mathrm{I}_0)$, $x_{\mathrm{F}}^1 = \phi(\mathrm{I}_1)$. To conclude, observe that equation (3.6) implies that the vectors $\{\bar{\zeta}^\mu\}_{\mu=1}^P$ are equilibria of (FR). □

We now specialize Theorem 3.1.4 to the simple yet relevant case of $p = r$, and observe that the resulting matrix has a structure similar to (1.41), i.e. the canonical choice in the computational neuroscience literature.

Corollary 3.1.5 (Classic synaptic matrix). *Let $p = r$. Under the same assumptions of Theorem 3.1.4 the syanptic matrix W reduces to*

$$W = \frac{\alpha}{p(1-p)N}\sum_{\mu=1}^{P}(\zeta^\mu - p\mathbb{1}_N)(\zeta^\mu - p\mathbb{1}_N)^\top + \frac{\gamma}{N}\mathbb{1}_N\mathbb{1}_N^\top, \tag{3.7}$$

Proof. Observe that for $p = r$, the bias paramater β is

$$\beta = p\frac{px_{\mathrm{F}}^1 + (1-p)x_{\mathrm{F}}^0}{px_{\mathrm{F}}^1 + (1-p)x_{\mathrm{F}}^0} = p \tag{3.8}$$

□

Some comments on the parameters α and γ are in order. First, in (3.5a) the parameter α is positive and is inversely proportional to the slope of the straight line intersecting the activation function $\phi(\cdot)$ at the coordinates I_0 and I_1. Notice then that when $p = r$ the homeostatic parameter

becomes

$$\gamma := \frac{p\mathrm{I}_1 + (1-p)\mathrm{I}_0}{px_{\mathrm{F}}^1 + (1-p)x_{\mathrm{F}}^0}. \tag{3.9}$$

Geometrically, the inverse of the parameter γ is the slope of the line that passes through the origin and intersects the straight line $x = \alpha^{-1}(\mathrm{I} - \mathrm{I}_0) + x_{\mathrm{F}}^0$ in the coordinate $\mathrm{I}_p = p\mathrm{I}_1 + (1-p)\mathrm{I}_0$ (see Figure 3.1).

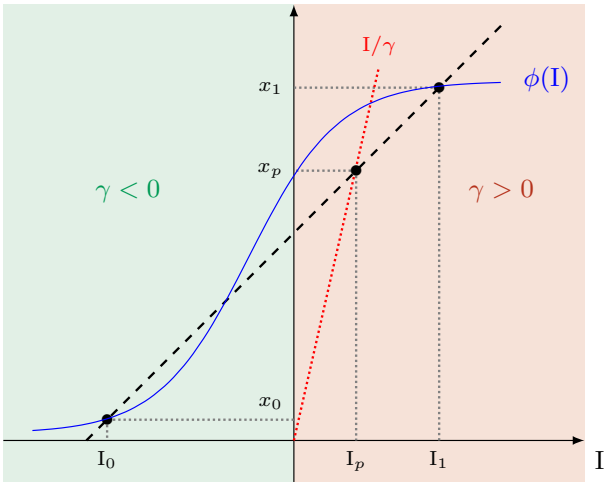


Figure 3.1: **Interpretation of the homeostatic parameter γ.** Graphical interpretation of the parameter γ in (3.5c) when $p = r$. γ^{-1} coincides with the slope of the straight line intersecting the origin and the point (I_p, x_p), where $\mathrm{I}_p := p\mathrm{I}_1 + (1-p)\mathrm{I}_0$ and $x_p := px_{\mathrm{F}}^1 + (1-p)x_{\mathrm{F}}^0$. Note that, if the point (I_p, x_p) belongs to the green area, then γ is negative, whereas if the point (I_p, x_p) belongs to the red area, then γ is positive.

The canonical case of Dayan & Abbott The matrix in (3.7) closely resembles the one in [39, Section 7.4], which is the canonical choice throughout the theoretical neuroscience literature. We next show that the latter is indeed a special case of (3.7) for $p = r$. The construction presented in [39] assumes the following:

(i) there exists z^* such that $\phi(z) = 0$ for $z \leq z^*$;

(ii) there are scalars $\lambda, \delta > 0$ such that $\mathrm{I}_0 = -\delta(1 + p\lambda) \leq z^*$, $\mathrm{I}_1 = \delta(\lambda - 1 - p\lambda)$ and $x_{\mathrm{F}}^1 := \phi(\mathrm{I}_1) = \delta$. Since it is assumed that $\mathrm{I}_0 \leq z^*$, then $x_{\mathrm{F}}^0 := \phi(\mathrm{I}_0) = 0$.

From the previous relations it can be seen that I_0, I_1 cannot be chosen arbitrarily. Indeed, the above defined I_0 and I_1 have to satisfy the equation $p\mathrm{I}_1 + (1-p)\mathrm{I}_0 + \phi(\mathrm{I}_1) = 0$. This implies that for this type of construction we can start from any $\phi(\cdot)$ satisfying condition (i) and any real number I_1 such that

$$\frac{p\mathrm{I}_1 + \phi(\mathrm{I}_1)}{p-1} \leq z^*$$

and take $\mathrm{I}_0 := \frac{p\mathrm{I}_1+\phi(\mathrm{I}_1)}{p-1}$. It can be seen that the parameters α, γ in this case become

$$\alpha = \frac{\mathrm{I}_1 - \mathrm{I}_0}{x_{\mathrm{F}}^1} = \lambda,$$
$$\gamma = -\frac{1}{p}$$

and the covariance-based synaptic matrix then becomes

$$\begin{aligned} W &= \frac{\lambda}{p(1-p)N}\sum_{\mu=1}^{P}(\zeta^\mu - p\mathbb{1}_N)(\zeta^\mu - p\mathbb{1}_N)^\top - \frac{1}{pN}\mathbb{1}_N\mathbb{1}_N^\top \\ &= \frac{\lambda}{p(1-p)\delta^2 N}\sum_{\mu=1}^{P}(\bar{\zeta}^\mu - p\delta\mathbb{1}_N)(\bar{\zeta}^\mu - p\delta\mathbb{1}_N)^\top - \frac{1}{pN}\mathbb{1}_N\mathbb{1}_N^\top \end{aligned} \tag{3.10}$$

where from the first to the second line in Equation (3.10) we have multiplied and divided by δ^2 the terms in the summation and used the fact that $\bar{\zeta}^\mu = \delta\zeta^\mu$. We have obtained in this way exactly the expression of W given in [39].

A bridge from math to biology We now want to address the biological interpretation of the map (3.7) and gain some insight on how the model can capture different aspects of neural processing. As it stands, the positivity of the *firing rate* model offers a valuable interpretative tool to bridge dynamical system theory and neural processes. Specifically, it offers the possibility to understand the role of the different components of the covariance-based synaptic matrix and how they interact with the network activity to generate neuronal rates of firing. Separating the different components

$$\begin{aligned} W \;=\; & \underbrace{\frac{\alpha}{p(1-p)N}\sum_{\mu}^{P}\zeta^\mu\zeta^{\mu\top}}_{\text{Long term potentiation network } W^{\mathrm{LTP}}} \\ & - \underbrace{\frac{\alpha}{(1-p)N}\left(\sum_{\mu}^{P}\mathbb{1}_N\zeta^{\mu\top} + \zeta^\mu\mathbb{1}_N^\top\right)}_{\text{Long term depression network } W^{\mathrm{LTD}}} + \underbrace{\left(\frac{\alpha}{(1-p)N}Pp + \frac{\gamma}{N}\right)\mathbb{1}_N\mathbb{1}_N^\top}_{\text{Global homeostatic network } W^{\mathrm{OM}}}, \end{aligned} \tag{3.11}$$

we observe that

- W^{LTP} is a relatively sparse network structure that takes into account how specific neuronal sub-clusters positively self-excite to produce fixation of the activity on a given pattern. Indeed, from the outer product of the patterns, we have that $W_{ij}^{\mathrm{LTP}} \neq 0$ only if there exists $\mu = 1, \dots, P$ such that $\zeta_i^\mu = 1$ and $\zeta_j^\mu = 1$.

- W^{LTD} is the network structure responsible for synaptic depression and is the sum of two terms, both regulating the global network activity with weights proportional to the parameters α, p. The first term, given by the outer products $\mathbb{1}_N \zeta^{\mu\top}$, provides global long term depression proportional to the correlation of the network activity and the different memory patterns. The second term, given by the outer products $\zeta^{\mu}\mathbb{1}_N^{\top}$, provides selective long term depression to the units of the memory patterns proportional to the summed network activity.

- W^{OM} is the homeostatic network and it exercises global regulation of the *firing rate* network via stimulation that depends on the normalized network rate

$$s(t) = \frac{\mathbb{1}_N^{\top} x_{\text{F}}(t)}{N} \tag{3.12}$$

Elaborating further on the biological constraints that lead to the construction of the covariance-based synaptic matrix, we observe that the W in (3.4) and (3.7) depend explicitly from the prototypical memory vectors $\{\zeta^{\mu}\}_{\mu=1}^{P}$. Notably, the form of W resembles the synaptic matrix of the classic Hopfield model [27], which can be constructed using the biologically plausible Hebbian learning rule [29]. Hebbian learning is a co-variation learning rule based on the principle articulated by D. O. Hebb "cells that fire together, wire together" [72], implying the use of local information to alter the strength of the synaptic couplings. In our case, we have that the covariance-based synaptic matrix is simply given by a one-shot Hebbian learning rule. Consistent with our interpretation, Hebbian learning as a learning framework adheres to the biological plasticity processes known as long-term potentiation (LTP) and long-term depression (LTD) [73], [74], known to be responsible for learning and forgetting in mammals. These local synaptic modifications aim to stabilize the connection strengths between neurons, enabling the retrieval of certain patterns through the collective dynamics of the network. Finally, we note that the symmetry of the synaptic matrix does not allow to distinguish excitatory from inhibitory neurons in the network. For greater biological plausibility, the synaptic matrix should adhere to Dale's law, which requires that all synaptic weights from a given neuron (i.e., within a column) share the same sign. A detailed study of such excitatory-inhibitory networks can be found in [75].

When the antimemories are equilibria Consider a set of prototypical memories (3.1) satisfying the equal sparsity and the equal correlation constraints of Assumption 3.1.2 with $p = r$ and let W be as in Corollary 3.1.5. Assume that $\phi(\cdot)$ in (FR) satisfies Assumption 3.1.1. For any prototypical memory ζ we define the corresponding antimemory as $\zeta^{\text{ant}} := \mathbb{1}_N - \zeta$ that is the vector in which we exchange the zeros with ones and vice versa. We wonder whether $\bar{\zeta}^{\text{ant}} := (x_{\text{F}}^1 - x_{\text{F}}^0)\zeta^{\text{ant}} + x_{\text{F}}^0 \mathbb{1}_N$ is an equilibrium point of (FR). Notice that this is exactly

what happens for the *voltage* models. First notice that $\bar{\zeta} + \bar{\zeta}^{\text{ant}} = (x_{\text{F}}^0 + x_{\text{F}}^1)\mathbb{1}_N$ so that $\bar{\zeta}^{\text{ant}} = (x_{\text{F}}^0 + x_{\text{F}}^1)\mathbb{1}_N - \bar{\zeta}$ and hence $W\bar{\zeta}^{\text{ant}} = (x_{\text{F}}^0 + x_{\text{F}}^1)W\mathbb{1}_N - W\bar{\zeta}$. Since we know that $W\mathbb{1}_N = \gamma\mathbb{1}_N$ and $W\bar{\zeta} = (\text{I}_1 - \text{I}_0)\zeta + \text{I}_0\mathbb{1}_N$, we can argue that $W\bar{\zeta}^{\text{ant}} = ((x_{\text{F}}^0 + x_{\text{F}}^1)\gamma - \text{I}_0)\mathbb{1}_N - (\text{I}_1 - \text{I}_0)\zeta$.

Now if i is such that $\zeta_i = 0$ and $\zeta_i^{\text{ant}} = 1$, then, after some computations we see that

$$(W\bar{\zeta}^{\text{ant}})_i = \frac{p\text{I}_1 x_{\text{F}}^1 + p\text{I}_1 x_{\text{F}}^0 + (1-2p)\text{I}_0 x_{\text{F}}^1}{px_{\text{F}}^1 + (1-p)x_{\text{F}}^0}$$

In order $\bar{\zeta}^{\text{ant}}$ to be an equilibrium point we need to impose that $\phi((W\bar{\zeta}^{\text{ant}})_i) = \bar{\zeta}_i^{\text{ant}} = x_{\text{F}}^1$ that holds if $(W\bar{\zeta}^{\text{ant}})_i = \text{I}_1$ which is equivalent to

$$\frac{p\text{I}_1 x_{\text{F}}^1 + p\text{I}_1 x_{\text{F}}^0 + (1-2p)\text{I}_0 x_{\text{F}}^1}{px_{\text{F}}^1 + (1-p)x_{\text{F}}^0} = \text{I}_1$$

After some simple computations we see that this is equivalent to the condition

$$(1-2p)(\text{I}_0 x_{\text{F}}^1 - \text{I}_1 x_{\text{F}}^0) = 0 \tag{3.13}$$

Notice that this holds if and only if $p = 1/2$ or $\text{I}_0 x_{\text{F}}^1 = \text{I}_1 x_{\text{F}}^0$. If we start from the hypothesis that i is such that $\zeta_i = 1$ and $\zeta_i^{\text{ant}} = 0$, then, after some computations we see that the equilibrium point condition holds again if and only if (3.13) holds true. We can argue that this condition is necessary and sufficient for having that all the antimemories provide equilibrium points for (FR). Notice that, while this condition holds true for the *voltage* models, it generically fails for general *firing rate* models when $p \neq 1/2$. In addition, while for the *voltage* model we have that $\text{I}_0 x_{\text{F}}^1 = \text{I}_1 x_{\text{F}}^0$ and so anti-memories are equilibrium points regardless the value of p, this is not true for *firing rate* models since the condition $\text{I}_0 x_{\text{F}}^1 = \text{I}_1 x_{\text{F}}^0$ generally fails.

Existence of homogeneous equilibria A well-known issue in associative memory networks with a Hebbian synaptic matrix is the emergence of spurious equilibria [76], [77], which are equilibria that do not correspond to the intended memories. We show now how the covariance-based synaptic matrix invariably produces at least one spurious equilibrium point with equal entries. Such equilibria will be called *homogeneous equilibria*.

Lemma 3.1.6 (Homogeneous equilibria). *With the same notation and under the same assumptions as Corollary 3.1.5,*

(i) there exists at least one solution z to the equation

$$\gamma^{-1} z = \phi(z), \tag{3.14}$$

(ii) for each solution z, the vector $\bar{x} = \gamma^{-1} z \mathbb{1}_n$, is an equilibrium point for the firing rate *model* (FR).

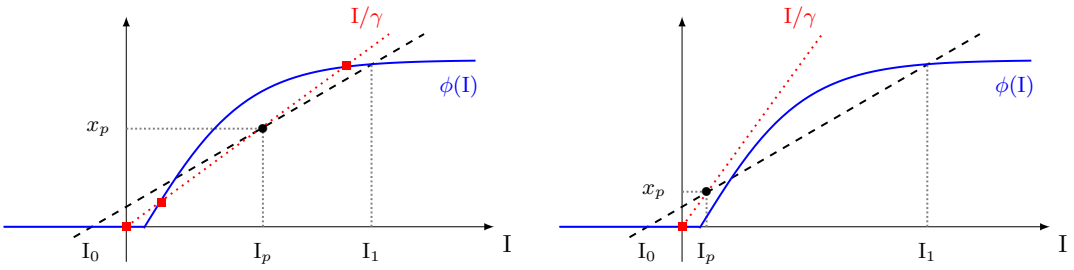
Proof of Lemma 3.1.6. (i) Let $f(z) := \phi(z) - \gamma^{-1} z$ that is continuous. Using the definition of γ in case $r = p$ it can be seen that

$$f(\mathrm{I}_0) f(\mathrm{I}_1) = -p(1-p) \frac{(\mathrm{I}_0 x_{\mathrm{F}}^1 - \mathrm{I}_1 x_{\mathrm{F}}^0)^2}{(p\mathrm{I}_1 + (1-p)\mathrm{I}_0)^2} \leq 0$$

Then $f(\mathrm{I}_0), f(\mathrm{I}_1)$ have opposite signs and by continuity of f there must be $z \in [\mathrm{I}_0, \mathrm{I}_1]$ such that $f(z) = 0$ and hence such z satisfies equation (3.14).

(ii) Assume now that z satisfies equation (3.14) and let $\bar{x} := \gamma^{-1} z \mathbb{1}_N$. Since when $r = p$ we have that $W\mathbb{1}_N = \gamma \mathbb{1}_N$, then $W\bar{x} = z\mathbb{1}_N$. Hence $\Phi(W\bar{x}) = \Phi(z\mathbb{1}_N) = \phi(z)\mathbb{1}_N = \gamma^{-1} z \mathbb{1}_N = \bar{x}$ proving in this way that $\bar{x}$ is an equilibrium point. □

Equation (3.14) has a simple graphical interpretation. Indeed, its solutions result from the intersections between the graph of $\phi(z)$ and the straight line $\gamma^{-1} z$. Observe that in case $\gamma < 0$, then $f(z) := \phi(z) - \gamma^{-1} z$ is strictly increasing and such that $f(-\infty) = -\infty$ and $f(+\infty) = +\infty$. Hence there always exists exactly one solution to equation (3.14). If instead $\gamma > 0$, then there might be multiple solutions (see Figure 3.2).



(a) γ large (b) γ small

Figure 3.2: **Condition for the existence of homogeneous equilibria.** Graphical illustration of the existence of homogeneous equilibria as γ varies. The red squares denote the solutions of equation (3.14), each associated with a homogeneous equilibrium point.

Memories not used in the construction of W Given a set of prototypical memories $\{\zeta^\mu\}_{\mu=1}^P$, consider a vector $\zeta \in \mathbb{R}^N$ different from each ζ^μ and with the property that the augmented set $\{\zeta, \zeta^\mu\}_{\mu=1}^P$ satisfies the equal sparsity and equal correlation Assumption 3.1.2. Define the covariance-based synaptic matrix W as a function of the prototypical memories $\{\zeta^\mu\}_{\mu=1}^P$ and the rescaled vector $\bar{\zeta} = (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta + x_{\mathrm{F}}^0 \mathbb{1}_N$. It is straightforward to verify that $W\bar{\zeta} =$

$[p\mathrm{I}_1 + (1-p)\mathrm{I}_0]\mathbb{1}_N$ from which it follows that $\Phi(W\bar{\zeta}) = \phi(p\mathrm{I}_1 + (1-p)\mathrm{I}_0)\mathbb{1}_N \neq \bar{\zeta}$. Hence, spurious equilibria such as ζ cannot exist.

3.2 Stability of the *firing rate* model

In this section, we examine the stability of equilibria of the *firing rate* dynamics (FR) with synaptic matrix W constructed as in the previous section.

3.2.1 On the local stability of retrievable memories

A sufficient condition that ensures local asymptotic stability of an equilibrium point can be derived via Lyapunov's indirect method [78, Theorem 4.7]. Specifically, if the Jacobian matrix of (FR) evaluated at an equilibrium point $\bar{x_{\mathrm{F}}}$, namely

$$\begin{aligned} J(\bar{x}_{\mathrm{F}}) &= -I + J_x\Phi(W\bar{x}_{\mathrm{F}}) \\ &= -I + \mathrm{diag}(\Phi'(W\bar{x}_{\mathrm{F}}))W, \end{aligned} \tag{3.15}$$

has all its eigenvalues with strictly negative real part, then $\bar{x}$ is locally asymptotically stable. The previous condition formally relates to the following lemma on the ordering of the synaptic matrix W.

Lemma 3.2.1 (Local asymptotic stability of the retrievable memories). *Consider a set of vectors $\{\zeta^\mu\}_{\mu=1}^P$ satisfying (3.2) and let W be as in Theorem 3.1.4. Assume that $\phi(\cdot)$ in (FR) satisfies Assumption 3.1.1. If*

$$\eta W \prec I_N \tag{3.16}$$

$$\tag{3.17}$$

where $\eta := \max\{\phi'(\mathrm{I}_0), \phi'(\mathrm{I}_1)\}$. Then the equilibria $\{\bar{\zeta}^\mu\}_{\mu=1}^P$ of (FR) are locally asymptotically stable.

Proof. It is well known that, according to Lyapunov's indirect method, a sufficient condition that ensures local asymptotic stability of an equilibrium point $\bar{\zeta}^\mu$ is that the Jacobian matrix (3.15) evaluated at $\bar{\zeta}^\mu$ has eigenvalues with negative real part. Observe that the Jacobian matrix can be written as

$$\begin{aligned} J(\bar{\zeta}^\mu) &= -I_N + \mathrm{diag}(\Phi'(W\bar{\zeta}^\mu))W \\ &= -I_N + \underbrace{\mathrm{diag}(\Phi'((\mathrm{I}_1 - \mathrm{I}_0)\zeta^\mu + \mathrm{I}_0\mathbb{1}_N))}_{=:D} W \end{aligned} \tag{3.18}$$

where we used the identity $W\bar{\zeta}^\mu = (\mathrm{I}_1 - \mathrm{I}_0)\zeta^\mu + \mathrm{I}_0\mathbb{1}_N$ established in the proof of Theorem 3.1.4. This means that D is a diagonal matrix with diagonal entries that are $\phi'(\mathrm{I}_0)$ or $\phi'(\mathrm{I}_1)$. From Assumption 1 we know that $\eta \geq 0$. We distinguish three cases:
a) If $\eta = 0$, then (3.16) holds true. Moreover, $D = 0$ and hence $J(\bar{\zeta}^\mu) = -I_N$ which implies that $\bar{\zeta}^\mu$ is locally asymptotically stable, proving that the thesis holds true in this case.
b) Assume now that $\eta > 0$ but that one of the two derivatives $\phi'(\mathrm{I}_0)$ or $\phi'(\mathrm{I}_1)$ is zero. This means that some of the diagonal entries of D are zero. Assume with no loss of generality that the first N_1 diagonal entries of D are positive. Hence we can write

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix}.$$

where D_1 is diagonal and with positive diagonal entries. Then

$$\begin{aligned} J(\bar{\zeta}^\mu) = -I_N + DW &= \begin{bmatrix} -I_{N_1} & 0 \\ 0 & -I_{N_2} \end{bmatrix} + \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \\ &= \begin{bmatrix} -I_{N_1} + D_1 W_{11} & D_1 W_{12} \\ 0 & -I_{N_2} \end{bmatrix}. \end{aligned}$$

where $N_1, N_2 \in \mathbb{N}$ and $N_1 + N_2 = N$. Hence the eigenvalues of $J(\bar{\zeta}^\mu)$ have negative real part if and only if $-I_{N_1} + D_1 W_{11}$ has this property and this happens if and only if $D_1^{1/2} W_{11} D_1^{1/2} \prec I_{N_1}$ or, equivalently, if and only if $W_{11} \prec D_1^{-1}$. In this way we have shown that $W_{11} \prec D_1^{-1}$ implies the local asymptotic stability of the equilibrium point $\bar{\zeta}^\mu$. Now observe that condition (3.16) implies $W_{11} \prec \eta^{-1} I_{N_1}$. Since $D_1 \preceq \eta I_{N_1}$, then $D_1^{-1} \succeq \eta^{-1} I_{N_1}$ and hence we can argue that $W_{11} \prec D_1^{-1}$ that is what we need for proving the local asymptotic stability.
c) Assume now that both the derivatives $\phi'(\mathrm{I}_0)$ and $\phi'(\mathrm{I}_1)$ are both nonzero. This means that the matrix D is invertible. The proof of the local asymptotic stability can be obtained following the same arguments used in the previous point. □

Building on this result, we next establish a sufficient condition for the local stability of the retrievable memories as equilibria of (FR).

Theorem 3.2.2 (Local stability condition for generalized prototypical memories). *Consider a set of vectors* $\{\zeta^\mu\}_{\mu=1}^P$ *satisfying* (3.2) *and let* W *be as in Theorem 3.1.4. Assume that* $\phi(\cdot)$ *in*

(FR) *satisfies Assumption 3.1.1 and let*

$$b := \frac{\mathrm{I}_0 x_F^1 - \mathrm{I}_1 x_F^0}{x_F^1 - x_F^0} \tag{3.19}$$

$$c := \alpha \frac{p-r}{1-p} \frac{1}{p x_F^1 + (1-p) x_F^0} \tag{3.20}$$

$$d := \alpha \frac{p-r}{1-p} \frac{\beta x_F^1 + (1-\beta) x_F^0}{p x_F^1 + (1-p) x_F^0} \tag{3.21}$$

$$\eta := \max\{\phi'(\mathrm{I}_0), \phi'(\mathrm{I}_1)\} \tag{3.22}$$

Let z_1, z_2 be the roots of the polynomial

$$\mathcal{P}(z) := (z-\gamma)(z-\alpha) + P(dz - \alpha d - bc)$$

(i) If z_1, z_2 are real then

$$\eta \max\{\alpha, z_1, z_2\} < 1 \tag{3.23}$$

implies that the equilibria $\{\bar{\zeta}^\mu\}_{\mu=1}^P$ of (FR) are locally asymptotically stable.

(ii) If z_1, z_2 are are not real then

$$\eta\alpha < 1 \tag{3.24}$$

implies that the equilibria $\{\bar{\zeta}^\mu\}_{\mu=1}^P$ of (FR) are locally asymptotically stable.

Proof. Let $\lambda_{\max}(\cdot)$ denote the maximum eigenvalue of a symmetric matrix. According the Lemma 3.2.1, $\lambda_{\max}(\eta W) = \eta\lambda_{\max}(W) < 1$ implies local asymptotic stability. Hence the theorem is proved if we prove that the eigenvalues of W may only take one of the values $0, \alpha, z_1, z_2$. To this aim consider the matrix $T := [\mathbb{1}_N, \bar{\zeta}^1, \dots, \bar{\zeta}^P] \in \mathbb{R}^{N\times(P+1)}$. Since it can be verified that $W\mathbb{1}_N = c\sum_{\mu=1}^P \bar{\zeta}^\mu + (\gamma - Pd)\mathbb{1}_N$, $W\bar{\zeta}^\nu = \alpha\bar{\zeta}^\nu + b\mathbb{1}_N$ for all $\nu = 1, \dots, P$, and $Ww = 0$ for all $w \in \mathrm{Im}(T)^\perp$, then we can argue that the subspaces $\mathrm{Im}(T)$ and $\mathrm{Im}(T)^\perp$ are W-invariant. Moreover, it holds

$$WT = T \underbrace{\begin{bmatrix} \gamma - Pd & b\mathbb{1}_P^T \\ c\mathbb{1}_P & \alpha I_P \end{bmatrix}}_{=:\, \bar{W}\in\mathbb{R}^{(P+1)\times(P+1)}}$$

Let now $y \neq \mathbb{0}$ such that $Wy = \lambda y$ for $\lambda \neq 0$ and observe that $y \in \mathrm{Im}(T)$, since $\mathrm{Im}(W) \subseteq$

$\mathrm{Im}(T)$. Define $x := T^\top y$ and observe that, since $y \in \mathrm{Im}(T)$, then, $x \neq \mathbb{0}$. We have that

$$\begin{aligned} \lambda x^\top &= \lambda y^\top T \\ &= y^\top W T = y^\top T \bar{W} = x^\top \bar{W} \end{aligned} \tag{3.25}$$

We proved in this way that $\mathrm{eig}(W)/\{0\} \subseteq \mathrm{eig}(\bar{W})/\{0\}$, where $\mathrm{eig}(\cdot)$ means the set of eigenvalues of a square matrix. The characteristic polynomial of $\bar{W}$ is $(z-\alpha)^{P-1}((z-\gamma+Pd)(z-\alpha)-Pbc)$ that has roots α, z_1, z_2. The thesis follows from the fact that $\alpha \geq 0$. □

Note moreover that case (ii) in the previous theorem is nongeneric in the sense that it applies in a very specific situation. [1] Moreover, if $p = r$, then $c = d = 0$, so that $z_1 = \gamma$, $z_2 = \alpha$ and the Theorem 3.2.2 reduces to the following.

Corollary 3.2.3 (**Local stability condition**). *With the same notation and under the same assumptions as Corollary 3.1.5, if*

$$\max\{\phi'(\mathrm{I}_0), \phi'(\mathrm{I}_1)\} \max\{\alpha, \gamma\} < 1, \tag{3.26}$$

then each retrievable memory $\bar{\zeta}^\mu$ is locally asymptotically stable for the firing rate *model* (FR).

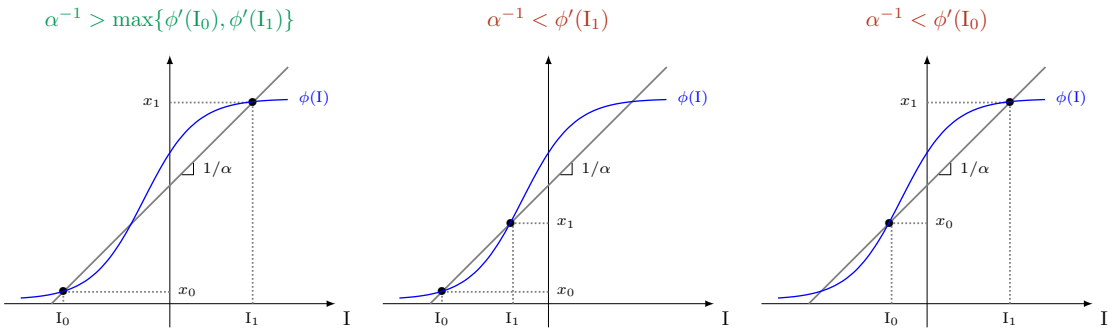
Remark 3.2.4 (**The canonical case of [39]**). For the memories construction presented in [39, Section 7.4], since $\gamma = -\frac{1}{p} < 0$, $\alpha = \lambda > 0$, and $\phi'(\mathrm{I}_0) = 0$, the stability condition (3.26) is simply

$$\phi'(\mathrm{I}_1) < \lambda^{-1}.$$

Theorem 3.2.3 indicates that the derivative of the activation function $\phi(\cdot)$ at the input coordinates I_0 and I_1 plays an important role in the stability of the retrievable memory patterns $\{\bar{\zeta}^\mu\}_{\mu=1}^P$. Specifically, the smaller are $\phi'(\mathrm{I}_0)$ and $\phi'(\mathrm{I}_1)$, the smaller is the left-hand side of (3.26), suggesting that the retrievable memories are more likely to be stable. In addition, the slope of the straight line intersecting the activation function at I_0 and I_1 also affects the condition (3.26) through parameter α. In particular, when $\gamma \leq \alpha$, stability is guaranteed if both $\phi'(\mathrm{I}_0) < 1/\alpha$ and $\phi'(\mathrm{I}_1) < 1/\alpha$, which means that the straight line intersecting the activation function at the points $(\mathrm{I}_0, x_{\mathrm{F}}^0)$ and $(\mathrm{I}_1, x_{\mathrm{F}}^1)$ has to intersect it from below (see Figure 3.3).

Remark 3.2.5 (**The case $\gamma \leq \alpha$**). We have seen that when $\gamma \leq \alpha$, the stability condition simplifies to $\max\{\phi'(\mathrm{I}_0), \phi'(\mathrm{I}_1)\} < \alpha^{-1}$. It can be proved that $\gamma \leq \alpha$ holds if and only if $\mathrm{I}_0 x_1 \leq \mathrm{I}_1 x_0$. Observe that this condition always holds if $\mathrm{I}_0 \leq 0 \leq \mathrm{I}_1$.

[1] Namely, the equality $Pp - (P-1)r = 1$ has to hold.



(a) (b) (c)

Figure 3.3: **The local stability condition.** Graphical interpretation of the local stability condition (3.26) when $\gamma \leq \alpha$. In panel (a) condition (3.26) is satisfied, while in panels (b)-(c) condition (3.26) is not satisfied and hence Theorem 3.2.3 cannot be applied.

Remark 3.2.6 (**Instability condition**). Techniques similar to those in the proof of Theorem 3.2.3 lead to the following statement: if

$$\max\{\phi'(I_0)[p\alpha + (1-p)\gamma], \phi'(I_1)[(1-p)\alpha + p\gamma]\} > 1, \tag{3.27}$$

then the equilibria $\{\bar{\zeta}^\mu\}_{\mu=1}^P$ of (FR) are unstable.

Remark 3.2.7 (**Homogeneous equilibria**). The local stability of homogeneous equilibria introduced in Section 3.1 can be analyzed with similar arguments used for proving Theorem 3.2.3. Specifically, if $\bar{x} = \gamma^{-1} z \mathbb{1}_N$ an equilibrium point for (FR), where z is a real number satisfying equation (3.14), then it is easy to see that $\bar{x}$ is locally stable if

$$\phi'(z) \max\{\alpha, \gamma\} < 1.$$

It can also be proved that such equilibrium point is unstable if instead

$$\phi'(z) \max\{\alpha, \gamma\} > 1,$$

In the special case in which $\gamma \leq \alpha$, the stability condition simplifies to $\phi'(z) < \alpha^{-1}$.

3.2.2 On the global stability of retrievable memories

Proving global convergence of the *firing rate* dynamics starts from observing that certain sets are forward-invariant for the dynamics, i.e. any trajectory of the system initiated within the set remains confined to it.

Lemma 3.2.8 (Forward-invariance of the FR dynamics). *Let Assumption 3.1.1 hold and assume that $\phi(\cdot)$ takes values in a bounded interval $\mathcal{I}$. Then there exists a large enough M such that the set $[\phi(-M), \phi(M)]^N$ is forward invariant for the dynamics* (FR).

Proof. Since $\mathcal{I}$ is bounded, then there exists $x_M > 0$ such that $\mathcal{I} \subseteq [-x_M, x_M]^N$. Let $\bar{M} := \|W\|_\infty x_M$, where $\|\cdot\|_\infty$ is the infinity norm of a vector or of a matrix. We want to prove that for any $M \geq \bar{M}$ we have that $[\phi(-M), \phi(M)]^N$ is forward invariant for the dynamics (FR). First observe that if $x_{\mathrm{F}} \in [\phi(-M), \phi(M)]^N$, then $x_{\mathrm{F}} \in \mathcal{I}^N \subseteq [-x_M, x_M]^N$ and hence $\|W x_{\mathrm{F}}\|_\infty \leq \|W\|_\infty \|x_{\mathrm{F}}\|_\infty = \bar{M}$ which implies that $\phi([W x_{\mathrm{F}}]) \in [\phi(-\bar{M}), \phi(\bar{M})]^N \subseteq [\phi(-M), \phi(M)]^N$. Let now $x_{\mathrm{F}} \in [\phi(-M), \phi(M)]^N$ such that $x_{\mathrm{F}i} = \phi(-M)$. Then

$$[\dot{x}_{\mathrm{F}}]_i = -x_{\mathrm{F}i} + \phi([W x_{\mathrm{F}}]_i) \geq -\phi(-M) + \phi(-M) = 0$$

On the other hand, if we take instead $x_{\mathrm{F}} \in [\phi(-M), \phi(M)]^N$ such that $x_{\mathrm{F}i} = \phi(M)$, then

$$[\dot{x}_{\mathrm{F}}]_i = -x_{\mathrm{F}i} + \phi([W x_{\mathrm{F}}]_i) \leq -\phi(M) + \phi(M) = 0$$

The above conditions imply that for any x_{F} belonging to the boundary of $[\phi(-M), \phi(M)]^N$ the direction of the vector derivative in the dynamics (FR) is either tangent or points inside the set $[\phi(-M), \phi(M)]^N$. The forward invariance of $[\phi(-M), \phi(M)]^N$ then follows from Nagumo's Theorem [53, Theorem 4.7]. □

We now present a result on the global behavior of the trajectories of (FR) based on an energetic characterization of the *firing rate* model.

The function $E_{\mathrm{FR}}(x_{\mathrm{F}})\colon \mathrm{Im}(\phi)^n \to \mathbb{R}$, defined as

$$\mathrm{E}_{\mathrm{FR}}(x_{\mathrm{F}}) = -\frac{1}{2} x_{\mathrm{F}}^\top W x_{\mathrm{F}} + \sum_{i=1}^{N} \int_0^{x_{\mathrm{F}i}} \phi^{-1}(z)\, \mathrm{d}z \tag{3.28}$$

will serve as energy for the *firing rate* model (FR). Here, $\mathrm{Im}(\phi)$ denotes the image of the activation function ϕ, and ϕ^{-1} any right inverse of ϕ.[2] Note that (3.28) coincides with the classic *voltage* Energy (1.6) expressed in the variable $x_{\mathrm{F}} = \Phi(x_{\mathrm{H}})$.

Theorem 3.2.9 (Global convergence to equilibria). *Consider the* firing rate *model* (FR) *with activation function $\phi(\cdot)$ satisfying the positivity and monotonicity Assumption 3.1.1. Assume moreover that the activation function takes value in a bounded interval $\mathcal{I}$.*

If $x_F(0) \in \mathcal{I}^N$, then $\dot{\mathrm{E}}_{\mathrm{FR}} \leq 0$ along the flow of (FR) *and each trajectory of* (FR) *converges to the set of equilibrium points of* (FR).

[2] A right inverse of a function $f\colon X \to Y$ is a function $g\colon Y \to X$ such that $f(g(y)) = y$ for all $y \in Y$.

Proof. By Lemma 3.2.8, for any $x_{\mathrm{F}}(0) \in \mathcal{I}^N$, there exists $M > 0$ such that the set $\mathcal{S} := [\phi(-M), \phi(M)]^N$ is forward invariant for (FR) and contains $x_{\mathrm{F}}(0)$. This implies that $E_{\mathrm{FR}}(x_{\mathrm{F}})$ in (3.28) and $\nabla \mathrm{E}_{\mathrm{FR}}(x_{\mathrm{F}}) = (-Wx_{\mathrm{F}} + \Phi^{-1}(x_{\mathrm{F}}))^\top$ are well-defined when evaluated along the trajectories of (FR).

The derivative of E_{FR} along each such trajectory is

$$\begin{aligned}\dot{\mathrm{E}}_{\mathrm{FR}}(x_{\mathrm{F}}) &= \nabla \mathrm{E}_{\mathrm{FR}}(x_{\mathrm{F}})\dot{x}_{\mathrm{F}} \\ &= (-Wx_{\mathrm{F}} + \Phi^{-1}(x_{\mathrm{F}}))^\top(-x_{\mathrm{F}} + \Phi(Wx_{\mathrm{F}})) \\ &= \sum_{i=1}^{N}(-[Wx_{\mathrm{F}}]_i + \phi^{-1}(x_{\mathrm{F}i}))(-x_i + \phi([Wx_{\mathrm{F}}]_i)),\end{aligned}$$

where we used the symmetry of W. Since $\phi(\cdot)$ is weakly increasing and $\phi^{-1}(\cdot)$ is a right inverse of $\phi(\cdot)$, if $\phi^{-1}(x_{\mathrm{F}i}) \geq [Wx_{\mathrm{F}}]_i$ then $\phi(\phi^{-1}(x_{\mathrm{F}i})) = x_{\mathrm{F}i} \geq \phi([Wx_{\mathrm{F}}]_i)$, and if $\phi^{-1}(x_{\mathrm{F}i}) \leq [Wx_{\mathrm{F}}]_i$ then $x_{\mathrm{F}i} \leq [Wx_{\mathrm{F}}]_i$. This implies that $(-[Wx_{\mathrm{F}}]_i + \phi^{-1}(x_{\mathrm{F}i}))(-x_{\mathrm{F}i} + \phi([Wx_{\mathrm{F}}]_i)) \leq 0$ for all $i = 1, \dots, N$ and for all $x_{\mathrm{F}} \in \mathcal{S}$. As a consequence, $\dot{\mathrm{E}}_{\mathrm{FR}}(x_{\mathrm{F}}) \leq 0$ for all $x_{\mathrm{F}} \in \mathcal{S}$. Moreover, $\dot{\mathrm{E}}_{\mathrm{FR}}(x_{\mathrm{F}}) = 0$ if and only if $\phi^{-1}(x_{\mathrm{F}i}) = [Wx_{\mathrm{F}}]_i$ or $\phi([Wx_{\mathrm{F}}]_i) = x_{\mathrm{F}i}$ for all $i = 1, \dots, N$. Since $\phi^{-1}(x_{\mathrm{F}i}) = [Wx_{\mathrm{F}}]_i$ implies $\phi([Wx_{\mathrm{F}}]_i) = x_{\mathrm{F}i}$, we conclude that $\dot{\mathrm{E}}_{\mathrm{FR}}(x_{\mathrm{F}}) = 0$ if and only if $\phi([Wx_{\mathrm{F}}]_i) = x_{\mathrm{F}i}$ for all $i = 1, \dots, N$, that is, if and only if x_{F} is an equilibrium point of (FR). Since $\mathrm{E}_{\mathrm{FR}}(x_{\mathrm{F}})$ is Lipschitz in $\mathcal{S}$, the thesis follows from the LaSalle invariance principle [79, Corollary 1]. □

3.3 Illustrative examples

The aim of this section is to provide hints to the reader on the design choices that result in effective associative memory *firing rate* networks. The examples will be based on the hypothesis of $p = r$, and the synaptic design will follow Corollary 3.1.5. Choosing the proper combination of parameters to set up an effective *firing rate* model can be challenging due to the presence of five free parameters; namely, the average activation p, the input currents I_0, I_1, the *gain factor* ρ and the *activation current* $\mathrm{I}^\star$. The latter two parameters were not previously discussed, but as we will see while we study two relevant examples, they are of critical importance in determining the stability properties of the system. Therefore, we will begin by fixing three out of the five free parameters, and study the stability properties of the system as the other two varies. Specifically, taking inspiration from neurobiological reality [36] we will fix the average neural activation as $p = 0.2$. We will then continue with the definition of two paradigmatic examples of activation function, and numerically study the stability properties of the system as the slope and offset parameters vary. The stability condition (3.26) is independent of network size; instead, the numerical stability of the *firing rate* model is studied for a network with $N = 1000$ and $P = 6$.

We considered the *firing rate* system to have numerically stable retrievable memories when the Jacobian of the vector field (FR) evaluated at each of the memories has all eigenvalues with negative real part. Additionally, we will present the energy function associated to specific *firing rate* models where the slope and offset parameters have been fixed to values that ensure either the stability or instability of the memories as equilibria for the system. In order to do so, we will evaluate the energy on a mesh interpolating the space between two memories as

$$x_{\mathrm{F}} = t_1 \zeta^1 + t_2 \zeta^2, \quad t_1, t_2 \in [0, 1]. \tag{3.29}$$

The energy profiles are plotted only for values of t_1, t_2 such that $x_{\mathrm{F}} \in [0, 1]^N$.

Finally, we test the retrieval performances of the *firing rate* model using the average overlap parameter

$$m^{\nu}(t) = \frac{x_{\mathrm{F}}(t)^{\top} \zeta^{\nu}}{pN} \tag{3.30}$$

where the average is taken over all the $\nu = 1, \dots, P$. In this testing phase, the prototypical memories $\{\zeta^{\mu}\}_{\mu=1}^{P}$ are drawn randomly so that the equal sparsity and equal correlation constraints are satisfied in expectation. Each retrieval is characterized by dynamics initialized arbitrarily close to the pattern $\bar{\zeta}^{\nu}$ and, if the memory is a stable equilibria of the system, then by the equal sparsity constraint there exists a $\bar{t} > 0$ such that $m^{\nu}(t) \approx 1$ for all $t > \bar{t}$. Conversely, by the equal correlation constraint we would expect that $m^{\mu}(t) \approx p$ for all $t > \bar{t}$ and $\mu \neq \nu$.

3.3.1 Rectified activation functions

Consider the rectified hyperbolic tangent activation (Fig. 3.4)

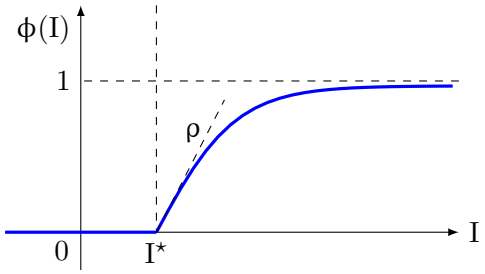


Figure 3.4: **The ReTanh(x).** Rectified hyperbolic tangent activation function (3.31).

$$\phi(\mathrm{I}) = \begin{cases} \tanh(\rho(\mathrm{I} - \mathrm{I}^{\star})), & x > \mathrm{I}^{\star}, \\ 0, & x \leq \mathrm{I}^{\star}, \end{cases} \tag{3.31}$$

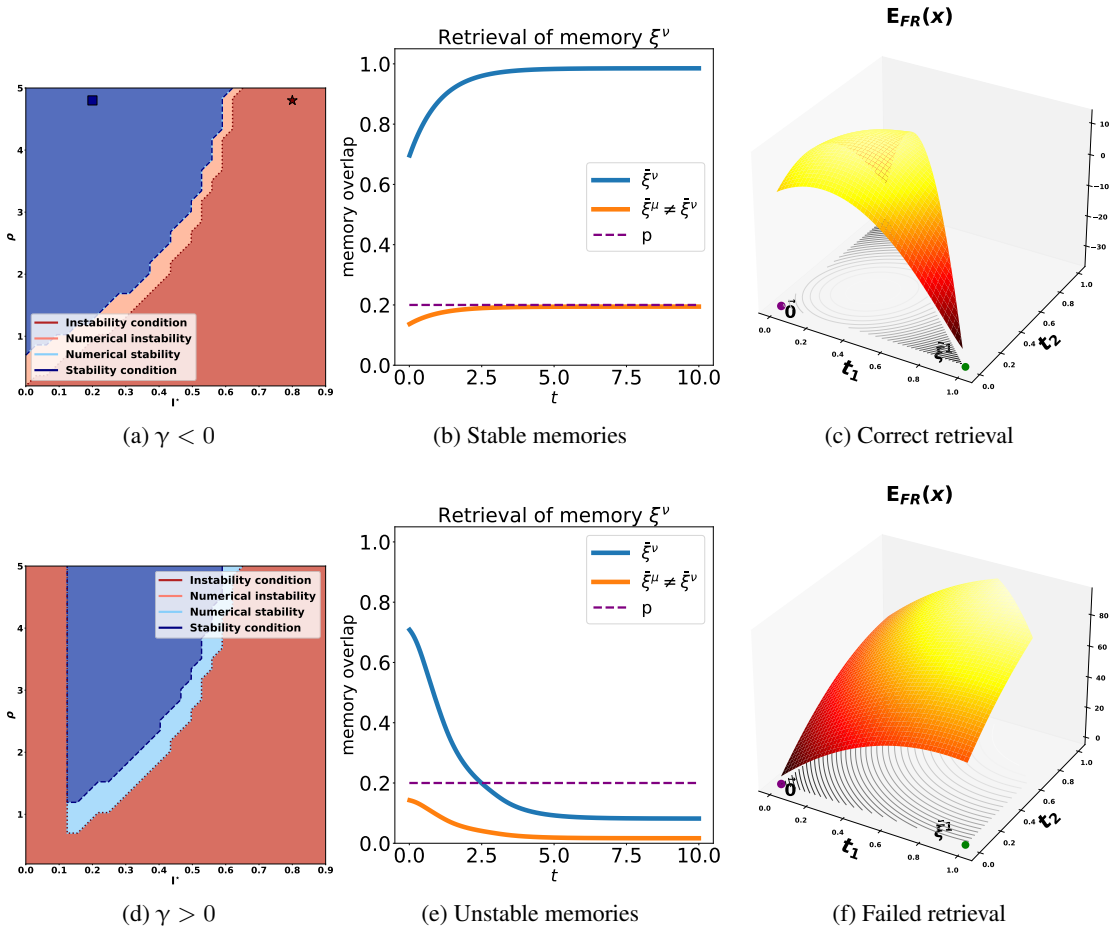
where the parameter $I^\star \in \mathbb{R}$ is the largest input current that yields a zero output and will be referred to as *activation current*, and $\rho \in \mathbb{R}$, $\rho > 0$, equals the maximal derivative of $\phi(\cdot)$ and will be termed *gain factor* (see also Fig. 3.4). This function is typically used to replicate the *firing rate* response to constant input currents observed in biological neural networks [39, Ch. 2]. In Fig. 3.5(a,d) we study the stability properties of the system as a function of the parameters ρ, $I^\star$, and the sign of the homeostatic strength γ. As observable, a negative homeostatic strength γ, which coincides with inhibitory feedback, results in a wider stability region for the *firing rate* system. Figures 3.5(b) and 3.5(e) plot the energy associated with *firing rate* models with, $\rho = 4.8$, $I^\star = 0.2$ for stable retrievable memories (blue square marker in Fig. 3.5(a)) and $I^\star = 0.75$ for unstable retrievable memories (red starred marker in Fig. 3.5(a)). As observable from Fig. 3.5(c), when the network parameters ensure the stability of the memories, the *firing rate* model is capable of performing the correct retrieval of the intended memory. Instead, when the network parameters lie in the instability region, the *firing rate* model activity collapses to a state of almost total inactivation (Fig. 3.5(f)), compatible with the energy profile in Fig. 3.5(e).

3.3.2 Sigmoidal activation functions

Consider the sigmoidal activation function (Fig. 3.6)

$$\phi(I) = \frac{1}{1 + e^{-4\rho(I - I^\star - (2\rho)^{-1})}}, \quad \rho,\, I^\star \in \mathbb{R},\, \rho > 0. \tag{3.32}$$

This function has a long history in machine learning research and is a commonly employed activation function in (artificial) neural networks [80], especially for binary classification problems where the output is interpreted as a probability. The sigmoidal activation function offers a smoothed transition compared to the sharp threshold of the rectified hyperbolic tangent, providing more gradual changes in neural firing rates. As such, it makes sense to define the *activation current* $I^\star$ as the key point where neural firing rates start to considerably rise, as observable from Fig. 3.6. Mathematically, $I^\star$ is defined by $\phi(I^\star + (2\rho)^{-1}) = 1/2$, indicating the I-axis intersection with the tangent to ϕ at its inflection point. In this case, the *gain factor* ρ is scaled by 4 to match the maximal slope of the sigmoid to that of the rectified hyperbolic tangent, allowing a direct comparison of stability effects across both functions. The aim is to study how the stability of the system varies depending on the values of the parameters ρ and $I^\star$ as they vary along some fixed intervals. As shown in Fig. 3.7(a,d), areas satisfying the stability condition (3.26) align closely with regions of observed numerical stability, suggesting the accuracy of the analytical stability criterion. Furthermore, the area for the stability condition widens when $\gamma < 0$, whereas it shrinks considerably to a narrow region for $\gamma > 0$. Figures 3.7(b) and 3.7(e) plot the energy associated with *firing rate* models with, $\rho = 4.8$, $I^\star = 0.2$ for stable retrievable



(a) $\gamma < 0$ (b) Stable memories (c) Correct retrieval

(d) $\gamma > 0$ (e) Unstable memories (f) Failed retrieval

Figure 3.5: **Retrieval performance for the ReTanh activation function.** Qualitative assessment of the design of *firing rate* models with a rectified hyperbolic tangent activation function. (a, d) Phase diagram for the stability (or instability) of the *firing rate* model for (a) $\gamma < 0$ ($\mathrm{I}_0 = -0.3$ and $\mathrm{I}_1 = 0.9$) and for (d) $\gamma > 0$ ($\mathrm{I}_0 = 0.1$ and $\mathrm{I}_1 = 0.9$). The instability region is wide, and for $\gamma > 0$, it consists of two disconnected region. Specifically, sufficiently small offsets lead to instability regardless of the slope. This is exactly the case in which $\gamma < \alpha$ and the point $(\mathrm{I}_0, x_{\mathrm{F}}^0)$ is intersected from above, thus precluding stability. (b, e) Energy surface of the *firing rate* model with $\gamma < 0$ when the memories are (b) stable and (e) unstable. (c-f) Retrieval performances of the *firing rate* model for $\gamma < 0$ when the memories are (c) stable and (f) unstable. Parameters are set to $\rho = 4.8$ and $\mathrm{I}^\star = 0.2$ (stable memories) or $\mathrm{I}^\star = 0.75$ (unstable memories).

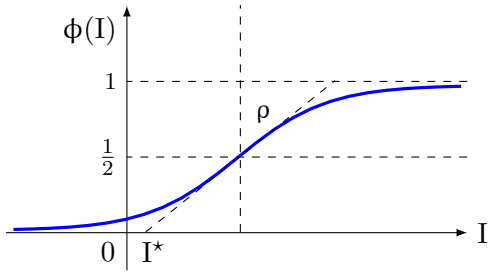


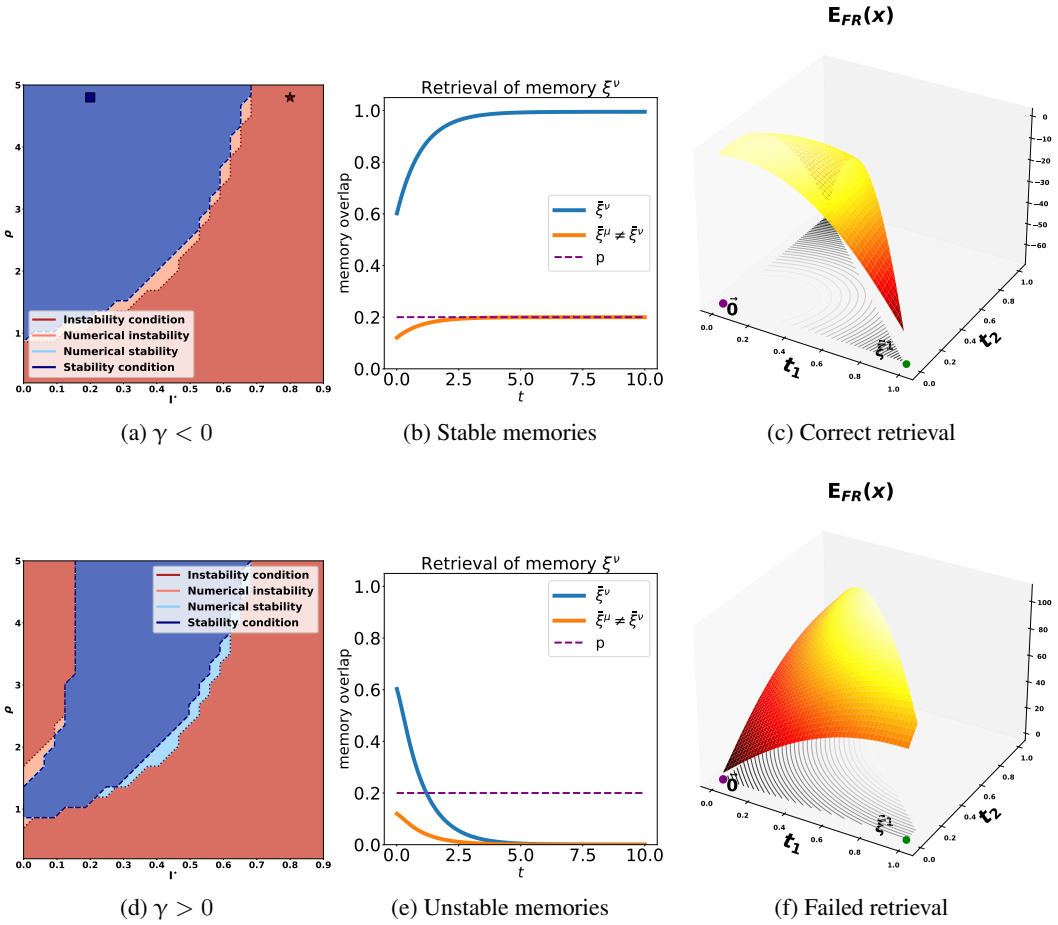
Figure 3.6: **The sigmoid.** Sigmoidal activation function (3.32).

memories (blue square marker in Fig. 3.7(a)) and $\mathrm{I}^\star = 0.8$ for unstable retrievable memories (red starred marker in Fig. 3.7(a)). Crucially, when network parameters guarantee the stability of retrievable memories (Fig. 3.7(c)), the dynamics of the *firing rate* model naturally converge toward them. Conversely, when the network parameters make the retrievable memories saddle points for the dynamics, the system's activity is repelled from these states and moves toward the origin (Fig. 3.7(f)). Finally, as shown by a comparison of Fig. 3.5(a,d) and Fig. 3.7(a,d), the smoothness of the sigmoidal activation function allows for broader stability regions across the parameter space, offering enhanced stability compared to the rectified hyperbolic tangent.

Random memories Most of the existing literature in *voltage* and *firing rate* models deals with random realizations of the memory patterns [57], [58], [81]. Specifically, the entries of the memory patterns are independent realizations from the probability distribution

$$\mathbb{P}(\zeta_i^\mu = 1) = 1 - \mathbb{P}(\zeta_i^\mu = 0) = p \qquad p \in (0,1). \tag{3.33}$$

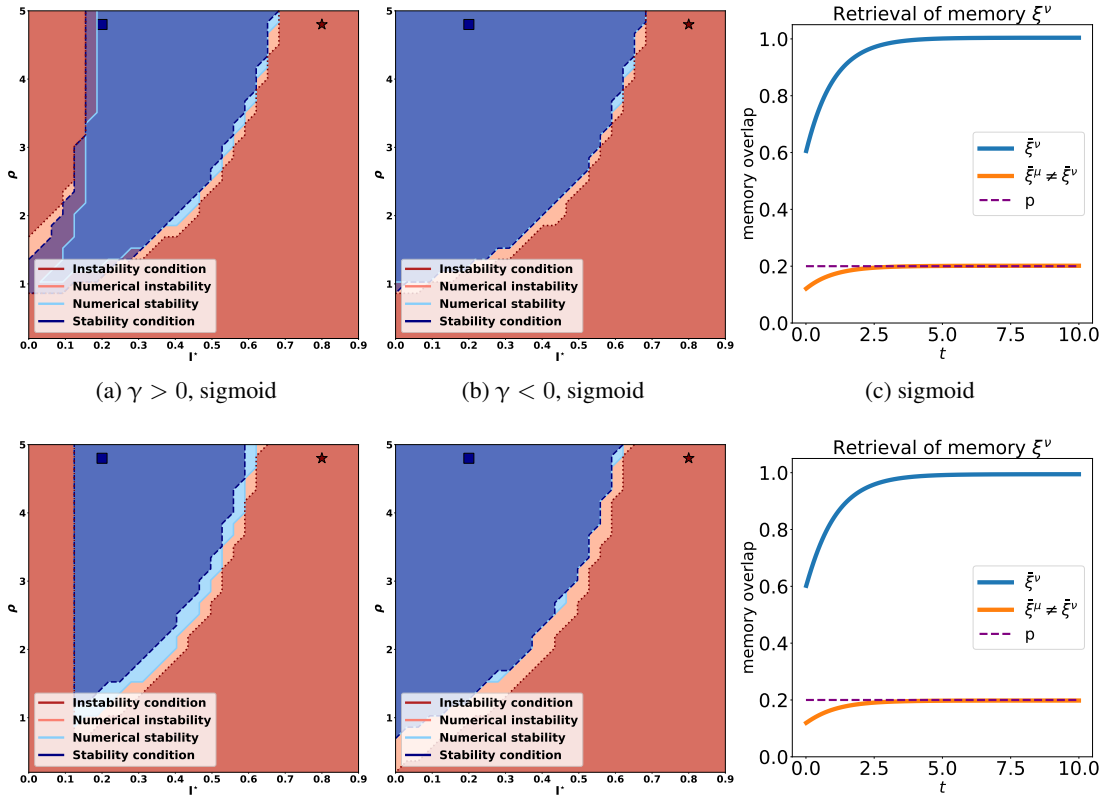
Our manuscript has instead focused on a setting with deterministic memories that satisfy the equal sparsity (3.2a) and equal correlation (3.2b) constraints. We are now interested in testing the accuracy of the analytical local stability condition and the retrieval performance of the *firing rate* model under the randomness hypothesis. In particular, we numerically compare the local stability condition (3.26) with the numerical stability of a system of $N = 1000$ neurons and $P = 72 \approx N/(2\log(N))$ [32] random memories. Fig. 3.8 clearly shows that the endogenous noise arising from crosstalk has a negligible effect on the stability estimates. Therefore, the designed *firing rate* network is robust and preserves the stability of its memories even in the random setting. The transposition of the analytical results for the deterministic setting to the canonical case of random memories further validates the complementarity of the dynamical system theory and statistical mechanical approaches for the study of complex systems.



(a) $\gamma < 0$ (b) Stable memories (c) Correct retrieval

(d) $\gamma > 0$ (e) Unstable memories (f) Failed retrieval

Figure 3.7: **Retrieval performance for the sigmoid activation function.** Qualitative assessment of the design of *firing rate* models with sigmoidal activation function. (a, d) Phase diagram for the stability (or instability) of the *firing rate* model for (a) $\gamma < 0$ ($I_0 = -0.3$ and $I_1 = 0.9$) and for (d) $\gamma > 0$ ($I_0 = 0.1$ and $I_1 = 0.9$). Specifically, for $\gamma > 0$ we have two wide, disconnected instability regions, and a narrow stability region, highlighting how excitatory homeostatic strength significantly increase the model sensitivity to parameters choice. (c, f) Energy surface associated to *firing rate* models with parameters lying in the (c) stability region of the phase portrait (f) instability region of the phase portrait. (c) For the choice of parameters $\rho = 4.8$ and $I^\star = 0.2$, the memories of the *firing rate* system are locally asymptotically stable and local minima for the Energy. (f) For the choice of parameters $\rho = 4.8$ and $I^\star = 0.8$, the memories are unstable equilibria for the *firing rate* model and saddle points of the associated Energy.



(a) $\gamma > 0$, sigmoid (b) $\gamma < 0$, sigmoid (c) sigmoid

(d) $\gamma > 0$, tanh (e) $\gamma < 0$, tanh (f) tanh

Figure 3.8: **Retreival for random memories initialization.** Effect of randomness on the stability of the memories. (a,d) For positive homeostatic terms γ, the randomness compromises the stability of memories when the network parameters are close to the boundaries of the analytical stability region. Nonetheless, the numerical stability region does not shrink considerably with respect to the deterministic memory setting. (b,e) For negative homeostatic terms γ, the randomness negligibly affects the local stability of the memories with respect to the deterministic case. (c,f) Despite the endogenous noise due to crosstalk, the network is still able to successfully retrieve all the memories for a choice of parameters in the stability region (blue squares).

3.4 Conclusion

Firing rate models provide a biologically plausible and efficient framework for representing neural activity at the population level, making them particularly well-suited for capturing key macroscopic behaviors in neuronal networks, including associative memory formation. This study introduces a positive *firing rate* model that encodes memories as neural firing rates, thus allowing for closer modeling of the averaged activity levels observed in brain networks [36, Sect. 17.2.6]. Through a synaptic matrix balancing LTP and LTD, we introduced a covariance-based approach that allows re-scaled memories to emerge as equilibrium points in network dynamics. This approach enables a reinterpretation of well-known examples from the literature [39, Sec. 7.4] and resolves the "anti-memory" phenomenon—a drawback inherent to traditional *voltage* model designs—by stabilizing only intended memories as equilibrium states. Furthermore, our analysis of local and global asymptotic stability conditions for rescaled memories provides insight on the design choices that ensure effectiveness and robustness of the *firing rate* model as an associative memory device. This study applies rigorous mathematical analysis typically reserved for *voltage* models to *firing rate* frameworks, bridging a critical gap and enabling a unified understanding of associative memory dynamics. These results are easily amenable to graphical representation, and therefore present both practitioners and theoreticians with clear intuitive guidance on the design choices that will result in an effective *firing rate* associative memory system. This study employs a deterministic approach to memory definition, diverging from traditional probabilistic frameworks. Future research could explore the model's capacity in probabilistic settings, specifically addressing the scalability and storage limits of *firing rate* associative memories. In summary, this work advances our understanding of the fundamental properties of *firing rate* networks and their application to associative memory, providing both a theoretical contribution and practical insights for researchers developing biologically plausible associative memory networks.

This chapter is based on the work [82] published in the journal *Neural Computation*. In the following chapter, we revisit the *voltage* model and address the problem of input-driven memory retrieval, a largely overlooked aspect in the literature. This perspective offers a natural connection between associative memory networks and recent developments in generative AI models.

Appendix

Proposition 3.4.1 (Linear Independence of the Memories). *Let* $p, r \in (0, 1)$ *and let the prototypical memories* $\{\zeta^\mu\}_{\mu=1}^P$ *satisfy the equal sparsity (3.2a) and the equal correlation (3.2b)*

constraints. Suppose in addition that

$$x_F^1 \neq -x_F^0 \frac{1-p}{p}. \tag{3.34}$$

Then

(i) The prototypical memories $\{\zeta^\mu\}_{\mu=1}^P$ *are linearly independent.*

(ii) The retrievable memories $\{\bar{\zeta}^\mu\}_{\mu=1}^P$ *are linearly independent.*

Proof. We start by proving the second statement.

(ii) Suppose that $x_{\mathrm{F}}^1 \neq x_{\mathrm{F}}^0$ and that the retrievable memories $\bar{\zeta}^\mu = (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\zeta^\mu + x_{\mathrm{F}}^0 \mathbb{1}_N$ for $\mu = 1, \dots, P$ are linearly dependent, so that there exists $\omega = (\omega_1, \dots, \omega_P) \in \mathbb{R}^P / \{\mathbb{0}\}$ such that

$$\mathbb{0} = \sum_{\mu=1}^P \omega_\mu \bar{\zeta}^\mu \tag{3.35}$$

Without loss of generalization suppose that $\omega_\nu \neq 0$, from which

$$\zeta^\nu = -\sum_{\mu \neq \nu} \frac{\omega_\mu}{\omega_\nu} \zeta^\mu - \frac{x_{\mathrm{F}}^0}{x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0} \mathbb{1}_N \left(1 + \sum_{\mu \neq \nu} \frac{\omega_\mu}{\omega_\nu}\right) \tag{3.36}$$

Exploiting now the equal sparsity constraint (3.2a) we obtain that

$$\begin{aligned} pN &= \mathbb{1}_N^\top \zeta^\nu \\ &= -pN \sum_{\mu \neq \nu} \frac{\omega_\mu}{\omega_\nu} - pN \frac{x_{\mathrm{F}}^0}{x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0} \frac{1}{p} \left(1 + \sum_{\mu \neq \nu} \frac{\omega_\mu}{\omega_\nu}\right) \end{aligned} \tag{3.37}$$

which implies that

$$\left(1 + \sum_{\mu \neq \nu} \frac{\omega_\mu}{\omega_\nu}\right) pN = -pN \frac{x_{\mathrm{F}}^0}{x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0} \frac{1}{p} \left(1 + \sum_{\mu \neq \nu} \frac{\omega_\mu}{\omega_\nu}\right) \tag{3.38}$$

If $\sum_\mu^P \omega_\mu \neq 0$, then simplifying we derive the following contradiction

$$1 = -\frac{1}{p} \frac{x_{\mathrm{F}}^0}{x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0} \tag{3.39}$$

since by hypothesis $x_{\mathrm{F}}^0 / (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0) \neq -p$. Then we are only left to verify that the hypothesis $\sum_\mu^P \omega_\mu = 0$ leads to another contradiction. Using again the hypothesis $\omega_\nu \neq 0$ and

expressing $\omega_\nu = -\sum_{\mu\neq\nu}\omega_\mu$, we write the condition of linear dependence as

$$\begin{aligned}\mathbb{0}_N &= \sum_{\mu\neq\nu}\omega_\mu(\bar{\zeta}^\mu - \bar{\zeta}^\nu)\\ &= \sum_{\mu\neq\nu}\omega_\mu(x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)(\zeta^\mu - \zeta^\nu)\end{aligned} \tag{3.40}$$

from which it is immediate to reach the contradiction

$$\begin{aligned}0 &= \mathbb{0}_N^\top\zeta^\nu\\ &= (x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)\sum_{\mu\neq\nu}\omega_\mu(\zeta^\mu - \zeta^\nu)^\top\zeta^\nu\\ &= pN\underbrace{(x_{\mathrm{F}}^1 - x_{\mathrm{F}}^0)}_{\neq 0}\underbrace{(1-r)}_{\neq 0}\underbrace{\left(-\sum\nolimits_{\mu\neq\nu}\omega_\mu\right)}_{\omega_\nu}\end{aligned} \tag{3.41}$$

(i) To see that also the prototypical memories are linearly independent, it suffices to take $x_{\mathrm{F}}^1 = 1$ and $x_{\mathrm{F}}^0 = 0$. Then this reduces to verify that $\sum_\mu^P \omega_\mu = 0$, which we have already proven leads to a contradiction.

□

Notice that the condition (3.34) is automatically satisfied by any *firing rate* system with non-negative activation function.

4

Input-driven memory retrieval in the *voltage* model

The introduction of this thesis presented a comprehensive view of the *voltage* Hopfield model, focusing on how memory retrieval unfolds as a dynamical process over an energy landscape. That discussion, however, largely set aside the role of external inputs. In particular, it remains unclear to what extent retrieval can be driven purely by inputs, independently of the initial condition, and whether inputs alone can reliably select the desired attractor. Understanding this relationship is essential not only for theoretical completeness but also for connecting associative memory models with their biological and computational counterparts.

Since the early 1980s, the expression "associative memory network" has been inseparable from the Hopfield model [26], [27], and a substantial body of work has examined its dynamical and statistical properties [51], [81], [83]. Through tools borrowed from statistical mechanics, voltage-based formulations provided a compelling explanation for the emergence of multiple stable memories as a function of synaptic couplings, thereby furnishing a concrete dynamical retrieval mechanism grounded in an underlying energy function. More recently, a machine-learning–driven revival has expanded this classical framework: higher-order interactions [43] and multi-layer architectures [25], [84] have endowed the model with dramatically enhanced capacity [52] and forged direct conceptual links to modern transformer architectures and their attention mechanisms [45]. These developments have also motivated new hypotheses on functional interactions between neurons and astrocytes [4], illustrating the breadth of phenomena that attractor dynamics can illuminate.

Within this broad landscape, the role of external inputs emerges as a natural and pressing question: how do inputs shape retrieval, bias attractors, and potentially override dependence on initial conditions? In classic treatments on computational neuroscience [36], [39], [85], memory retrieval in the *voltage* model is implicitly described as a two-step process. First, a noisy or incomplete input is presented as a cue and adopted as an initial condition. Then, driven by an energy landscape, the network state flows towards the closest energy minimum representing the prototypical memory. While this classic two-step process is natural within an algorithmic

paradigm, it fails to explain how neuronal circuits continuously react and adapt in real time to external inputs.

In light of these limitations, we advocate for a paradigm shift from a two-step mechanism, akin to a standard algorithmic approach, to an input-driven dynamic mechanism, aligned with the principles of online algorithms and continual learning [86]–[88]. To this extent, we propose a version of the *voltage* model that is driven by external inputs. A key feature of this model is that the input shapes the energy landscape and affects the resulting gradient descent flow (see Fig. 4.1). Furthermore, our model admits a simple representation as a modern Hopfield network [25], [45]; this representation provides a conceptual bridge with the recent literature on transformer models and machine learning. Finally, the addition of noise reveals the advantageous integration of past and present information by our model, thereby reducing misclassification errors induced by inconsistent or 'glitchy' inputs.

4.1 Primer on the *voltage* model for memory retrieval

Voltage models are fundamental tools in the study of high level, distributed memories retrieval [27]. They simplify neural dynamics into the interplay of three components: a dissipative flow, constantly polarizing the network state towards its resting value, a synaptic flow, which takes into account the weighted sum of the incoming activity from other neurons in the network, and an external input. Namely,

$$\begin{cases} \dot{x}_{\mathrm{H}}(t) = \underbrace{-x_{\mathrm{H}}(t)}_{\text{dissipation}} + \underbrace{W\Psi(x_{\mathrm{H}}(t))}_{\text{synaptic interconnections}} + \underbrace{u(t)}_{\text{ext. input}} \\ x_{\mathrm{H}}(0) = x_0 \in \mathbb{R}^N \end{cases} \tag{4.1}$$

where the prototypical memories $\{\xi^\mu\}_{\mu=1}^P$, $\xi^\mu \in \{-1,+1\}^N$ are assumed to be orthogonal (see Example 1.2.14) or with entries that are independent and identically distributed (see Example 1.2.16).

The prototypical memories are stored in the synaptic matrix through one-shot Hebbian learning map $\Sigma_{\mathrm{H}} \mapsto \mathcal{W}(\Sigma_{\mathrm{H}}) = W$ as

$$W = \frac{1}{N}\sum_{\mu=1}^{P} \xi^\mu \xi^{\mu\top} \tag{4.2}$$

Under suitable assumptions on the activation function Ψ [25] and the external input (constant), convergence to any of the stored memories is guaranteed [26] by the existence of the energy

function (1.1) (see Fig. 4.1D)

$$\mathrm{E}_H\Big(x_{\mathrm{H}}; W\Big) = -\frac{1}{2}\Psi(x)^\top W \Psi(x) + (x_{\mathrm{H}} - u(t))^\top \Psi(x) - \sum_{i=1}^{N} \int_0^{x_{\mathrm{H}i}} \Psi_i(z)\, dz \tag{4.3}$$

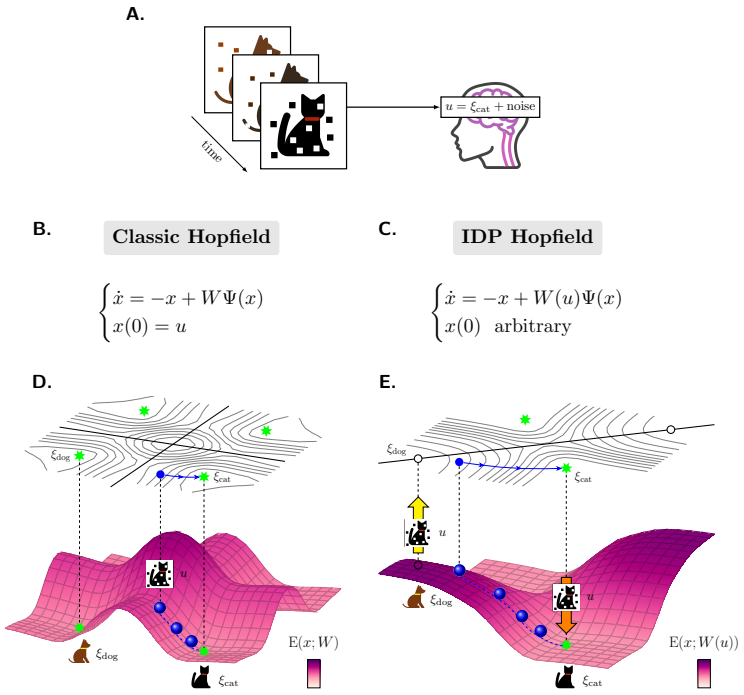


Figure 4.1: **Comparison between classic *voltage* and IDP *voltage* models.** (A) A slowly morphing sequence of noisy images is presented as a input to the observer, who updates its belief state to retrieve the memory closest to the current image u. This adaptation process occurs continuously. (B) In the classic model, the network state is set to an initial condition $x(0)$ equal to the current image u and then the *voltage* dynamics performs the memory retrieval task. (C) In the proposed input-driven plasticity model, the network initial condition is arbitrary, the image u modifies the synaptic weights $W(u)$, and the *voltage* dynamics with modified synaptic weights performs the memory retrieval task. This dynamics is well posed and naturally tracks the morphing images also when the image is time-varying $u = u(t)$. (D) In the classic model, the *voltage* dynamics is a gradient descent for the energy $\mathrm{E}(x_{\mathrm{H}}; W)$: the blue ball, representing the neural state, rolls from an initial condition towards a stable minimum point (cat memory). Therefore, the retrieval process is successful when the noisy image $x(0) = u$ (dotted cat) lies in the region of attraction of the correct memory (cat memory). (E) In the proposed model, the noisy image u directly modifies the synaptic matrix $W(u)$ and in turn the energy landscape $\mathrm{E}(x_{\mathrm{H}}; W(u))$, thereby extending the region of attraction of the correct memory. The retrieval process is successful from generic initial conditions when the correct memory is the unique minimum of the landscape. Specifically, in the panel, the noisy image (dotted cat) renders the correct memory a minimum (cat memory) and the incorrect memory (dog memory) no longer an equilibrium of $\mathrm{E}(x_{\mathrm{H}}; W(u))$.

given an appropriate initial condition x_0.

The model, first proposed as a discrete time dynamic system [27], has been particularly successful and captivating at explaining pattern reconstruction starting from a partial or corrupted cue [36], [39]. In its simplicity, the original theory effectively framed memory retrieval in the cascade of reactions that lead a network of elementary computational units to fix in a meaningful collective state, i.e. the retrievable memory. Nonetheless, most of the analysis on the model dynamics and retrieval capabilities requires the assumption $u \equiv \mathbb{0}_N$.

The mechanistic explanation of how the input affects the evolution of the *voltage* dynamics has been largely overlooked in the literature. As we will see, the introduction of an external input conflicts with the previous definition of prototypical memories as stable patterns of activity encoded in W. If the input is time-varying, then the energy function is no longer Lyapunov, namely it is not decreasing along the system trajectories, and hence we may lose convergence to any stable pattern. On the other hand, if the input is constant, it may distort the energy landscape in such a way that the minima are not related to any of the prototypical memories. Indeed, when the input is a heterogeneous mixture of memories, the energy landscape is inevitably corrupted and none of the stored memories is retrievable (see Fig. 4.2A). A possible solution is to clamp the input [89] for a brief time interval δt as an external driver for the dynamics, then un-clamp it and let the dynamics naturally evolve, namely

$$\dot{x}_{\mathrm{H}}(t) = -x_{\mathrm{H}}(t) + W\Psi(x_{\mathrm{H}}(t)) + u(t)\zeta(t) \tag{4.4}$$

$$\zeta(t) \propto \mathbb{1}_{[0,\delta t]}(t) \tag{4.5}$$

where $\mathbb{1}_S(t)$ is the characteristic function associated to any subset S of the real numbers defined by letting $\mathbb{1}_S(t) = 1$ if $t \in S$ and $\mathbb{1}_S(t) = 0$ otherwise. The external driver shoots the state trajectory in the direction of the subsequent input, acting as bias for the next retrieval. The model (4.4) is successful at discriminating the dominant component of very mixed inputs (see Fig. 4.2B), but it does so at the cost of introducing network-wide synchrony. In addition, as observable from Fig. 4.2A-B, each subsequent input instantaneously alters the network activity, canceling the information about any previous activity. Instead, in Fig. 4.2C our framework displays a remarkable capability of successfully retrieving the correct memory given the continuous external input, and it will be presented in the next section.

4.2 The Input-Driven Plasticity (IDP) *voltage* model

The aim of this work is that of studying a biologically plausible mechanism that, given a mixed input, first maintains fixation for short transients on what was previously retrieved, that is past information, and then gradually merges it with the current input, that is present information. This

gradual integration should, at a certain point, favor the present information and retrieve the correct memory. We thereby introduce an externally modulated *voltage* model, named input-driven plasticity (IDP) *voltage* model, that tries to capture this exact phenomenology. The IDP *voltage* model is defined as

$$\dot{x}_{\mathrm{H}}(t) = -x_{\mathrm{H}}(t) + W(u(t))\Psi(x_{\mathrm{H}}(t)) \tag{4.6}$$

where for a piecewise continuous, bounded input $u(t) \in \mathbb{R}^N$ the *input modulated synaptic matrix* is

$$W(u(t)) = \frac{1}{N}\sum_{\mu=1}^{P} \alpha_\mu(t)\xi^\mu\xi^{\mu\top} \tag{4.7}$$

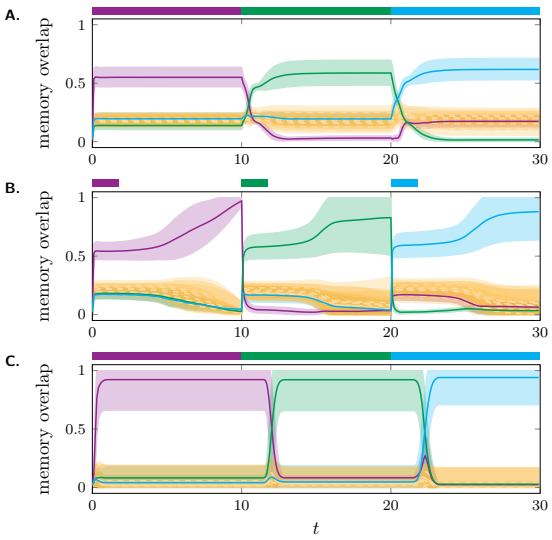


Figure 4.2: **Exploration of the response of different associative memory models to a time-varying input.** Each retrieval plot displays a sample average of the memory overlap parameter over 50 retrieval tasks, with the associated standard deviation shown through the shaded, color-matched surrounding area. Panel (A), (B), and (C) show the response of three different models to a time-varying input. The simulation time is segmented, with each sub-interval featuring a constant input. The horizontal bars on top of each retrieval plot display the duration of the external input and the color associated to its dominant saliency weight. At each external input switch, the previous dominant saliency weight shrinks below the stability threshold $\alpha_{\text{stability}}$. (A) Classic *voltage* model (V). The network dynamics converge to a mixed state, precluding exact retrieval of any individual memory. (B) Input-modulated *voltage* model (4.4). The modulator (4.5) shuts off the input after intervals of $t = 2$ simulation time. The network dynamics then freely recall the prototypical memory associated with the dominant component in the input. (C) IDP *voltage* model (4.6). The network naturally recalls the prototypical memory associated with the dominant component in the input. Remarkably, the model exhibits memory retention even after input switching, showcasing an intriguing integrative feature for past and present information.

where $\alpha_\mu(t) := [\xi^{\mu\top} u(t)]^2 \geq 0$ are called the saliency weights, in accordance with previous literature [90], [91]. These previous studies have explored the role of saliency weights and their effect on the dynamics (4.6) within the context of the morphing problem, focusing on how these weights drive transitions between memories as they evolve. However, a comprehensive analysis of how the magnitude of saliency weights affects the existence and stability of memories as equilibrium points remains missing. To address this gap, we begin by characterizing the input, followed by a detailed examination of the specific conditions on saliency weights necessary to ensure the existence and stability of memories as equilibria for eq. (4.6). We say that an input is homogeneous when it is substantially aligned with one of the prototypical memories and almost orthogonal to the others, i.e. when there exists a $\nu = 1, \dots, P$ such that $\alpha_\nu \gg 0$ and $\alpha_\mu \approx 0$ for all $\mu \neq \nu$. Conversely, we say that the input is well-mixed when it has similar saliency weights associated to all the prototypical memories, i.e. $\alpha_\mu \approx \alpha_\nu$ for all μ, ν. Furthermore, under a timescale separation hypothesis for neural dynamics and biologically trackable inputs, i.e. under the hypothesis that neural dynamics are much faster than input dynamics, the latter can be considered constant throughout each retrieval time interval. Unlike Hopfield's foundational works based upon a static synaptic matrix, we build on insights from [4], [90], [92] to introduce elementary dynamics into the synaptic structure. Differently from these treatments, we assume the synaptic changes to depend on an externally generated input rather than on a feedback activity of the associative memory system, such as that coming from intra-network astrocytes. This dynamic approach yields advantageous properties for system behavior, particularly by demonstrating how externally modulated synaptic changes can promote the retrieval of certain patterns while inhibiting others. To rigorously quantify how the input determines specific retrieval properties at the network level, we analytically study the existence and stability of memories, specifically for the case of monotonically increasing, odd activation functions $\Psi(x_H) = (\psi([x_H]_1), \dots, \psi([x_H]_N))$.

Assumption 4.2.1 (Activation function). Let the activation function $\psi \in C^2(\mathbb{R})$ satisfy the Assumption 2.1.1. We assume the following additional conditions:

$$0 < \psi'(z) \leq 1, \quad z \in \mathbb{R}, \quad \psi'(0) = 1 \tag{4.8}$$

$$\psi''(z) \begin{cases} < 0 & z > 0 \\ = 0 & z = 0 \\ > 0 & z < 0 \end{cases} \tag{4.9}$$

A function satisfying the above conditions is, for instance, $\psi(z) = \tanh(z)$. We now turn to the problem on the existence of retrievable memories.

4.2.1 Existence of equilibria for IDP *voltage* dynamics

Given an orthogonal prototypical memory ξ^ν, a retrievable memory associated to ξ^ν is a vector of the form $x_{\mathrm{H}\nu} = \gamma_\nu \xi^\nu$ for $\gamma_\nu \neq 0$ which is an equilibrium for eq. (4.6).

Theorem 4.2.2 (Existence of retrievable memories). *Let ξ^ν be a prototypical memory. Then $x_{H\nu} = \gamma_\nu \xi^\nu$ is an equilibrium for eq. (4.6) for some $\gamma_\nu \in \mathbb{R} \setminus \{0\}$ if and only if $\alpha_\nu > 1$. In this case, γ_ν satisfies*

$$\frac{\gamma_\nu}{\alpha_\nu} = \psi(\gamma_\nu) \tag{4.10}$$

Proof. The vector $x_{\mathrm{H}\nu} = \gamma_\nu \xi^\nu$ is an equilibrium for eq. (4.6) if and only if

$$\begin{aligned} x_{\mathrm{H}\nu} &= \gamma_\nu \xi^\nu \\ &= W(u)\Psi(\gamma_\nu \xi^\nu) \\ &= \frac{1}{N} \sum_{\mu=1}^{P} \alpha_\mu \xi^\mu \xi^{\mu\top} \cdot \xi^\nu \psi(\gamma_\nu) \\ &= \alpha_\nu \xi^\nu \psi(\gamma_\nu) \end{aligned} \tag{4.11}$$

and hence if and only if

$$\gamma_\nu = \alpha_\nu \psi(\gamma_\nu) \tag{4.12}$$

The theorem is proved if we prove that

$$\exists \gamma > 0 \ : \ \gamma = \alpha\psi(\gamma) \quad \Leftrightarrow \quad \alpha > 1 \tag{4.13}$$

We start from ($\Rightarrow$). Observe that since $\psi''(z) < 0$ for all $z > 0$ then $\psi'(z)$ is strictly decreasing for $z > 0$ and hence $\psi(z) < z$ for $z > 0$. This proves that if there exists $\gamma > 0$ such that $\gamma = \alpha\psi(\gamma)$, then $\alpha = \gamma/\psi(\gamma) > 1$. We prove now ($\Leftarrow$). Consider the function $f(z) := z - \alpha\psi(z)$ and observe that $f(0) = 0$, $f'(0) = 1 - \alpha < 0$ and that $\lim_{z\to+\infty} f(z) = +\infty$. These facts imply that there exists $\gamma > 0$ such that $f(\gamma) = 0$ and this concludes the proof. □

Fig. 4.3 provides a visualization of the existence condition in Theorem 4.2.2. By definition, retrievable memories need to be the corners of an hypercube [69], and we readily see which conditions on the saliency weights α's grant us their existence. If $\alpha \leq 1$, then in the semipositive interval $\mathbb{R}_+$ the dissipation line $\frac{\gamma}{\alpha}$ has only one intersection with the activation function that coincides with the center of the hypercube. Instead, for values $\alpha > 1$ we have two intersections, one coinciding with the desired corner of the hypercube.

Notice that, since each equilibrium is of the form $x_{\mathrm{H}\nu} = \gamma_\nu \xi^\nu$, it suffices to choose a

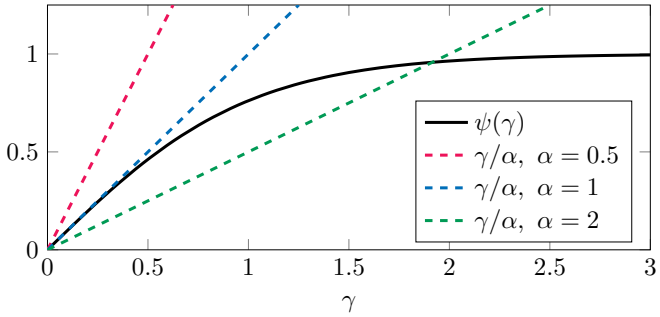


Figure 4.3: **Intersections of the activation function $\psi(\gamma) = \tanh(\gamma)$ with the dissipation line $\frac{\gamma}{\alpha}$ for different values of α.** The dissipation lines where the parameter $\alpha \leq 1$ have just one intersection with the chosen activation function. Instead, only for the value $\alpha = 2 > 1$ we obtain two intersections between the dissipation line and the activation function in the semipositive interval $\mathbb{R}_+$.

sufficiently fast saturating activation function to obtain that, in the space of membrane potentials, retrievable memories are of the form $\Psi(x_{H\nu}) = \xi^\nu \psi(\gamma_\nu) \approx \xi^\nu$. We can then conclude that

(i) given a prototypical memory ξ^ν, the associated retrievable memory is a vector of the form $x_{H\nu} = \gamma_\nu \xi^\nu$ for $\gamma_\nu \neq 0$ and is an equilibrium of eq. (4.6)

(ii) a retrievable memory exists under the current input if and only if the associated saliency weight is larger than an existence threshold:

$$\alpha_\nu > \alpha_{\text{existence}} = 1 \tag{4.14}$$

and in this case γ_ν satisfies

$$\frac{\gamma_\nu}{\alpha_\nu} = \psi(\gamma_\nu) \tag{4.15}$$

Remark 4.2.3. For simplicity, we are considering a constant input $u \in \mathbb{R}^N$. Notice however that more generally it suffices that $u : [0, T] \to \mathbb{R}^N$ is piecewise continuous and bounded. See the chapter Appendix for a detailed proof of this statement.

4.2.2 Stability of the equilibria for IDP *voltage* dynamics

We now investigate the condition for the local stability of the retrievable memories, i.e. those satisfying the existence condition. Furthermore, we observe how we can prove the existence of a stability threshold that depends on the current input.

Stability of Equilibria In order to determine when the equilibria $x_{H\nu} = \gamma_\nu \xi^\nu$ of our system are stable, we first establish a general local stability condition for *voltage*-type systems. In what

follows, given $g \in C^1(\mathbb{R}^N;\mathbb{R}^N)$, we denote the Jacobian of $g(\cdot)$ as the matrix $J_x g(x) \in \mathbb{R}^{N\times N}$ such that $[J_x g(x)]_{ij} = \frac{\partial g_i}{\partial x_j}(x)$. We now specifically address the stability of the IDP *voltage* model.

Theorem 4.2.4 (Local stability of equilibria and memories). *Consider the dynamics eq. (4.6) with a fixed input $u \in \mathbb{R}^N$. Then*

(i) an equilibrium $x_H^\star \in \mathbb{R}^N$ is locally exponentially stable if and only if

$$-[J_x\Psi(x_H^\star)]^{-1} + W(u) \prec 0 \tag{4.16}$$

(ii) if $x_{H\nu} = \gamma_\nu \xi^\nu$ with $\gamma_\nu \in \mathbb{R}\setminus\{0\}$ is an equilibrium, then $x_{H\nu}$ is locally exponentially stable if and only if

$$\psi'(\gamma_\nu) < \frac{1}{\max_{\mu=1,\dots,P}\ \alpha_\mu} \tag{4.17}$$

Proof.

(i) Consider, for fixed $u \in \mathbb{R}^N$, the field associated to the dynamics eq. (4.6)

$$F(x_{\mathrm{H}}) = -x_{\mathrm{H}} + W(u)\Psi(x_{\mathrm{H}}) \tag{4.18}$$

and compute the corresponding Jacobian at the equilibrium $x_{\mathrm{H}}^\star$

$$J_x F(x_{\mathrm{H}}^\star) = -I_N + W(u) J_x\Psi(x_{\mathrm{H}}^\star) \tag{4.19}$$

Since $J_x F(x_{\mathrm{H}}^\star) \in \mathbb{R}^{N\times N}$ is not necessarily symmetric, we apply the similarity transformation

$$J_x\Psi(x_{\mathrm{H}}^\star)^{\frac{1}{2}} J_x F(x_{\mathrm{H}}^\star) J_x\Psi(x_{\mathrm{H}}^\star)^{-\frac{1}{2}} = J_x\Psi(x_{\mathrm{H}}^\star)^{\frac{1}{2}} \underbrace{\Big(-[J_x\Psi(x_{\mathrm{H}}^\star)]^{-1} + W(u)\Big)}_{S(x_{\mathrm{H}}^\star)} J_x\Psi(x_{\mathrm{H}}^\star)^{\frac{1}{2}}$$

where we have exploited the positive definiteness of $J_x\Psi(x_{\mathrm{H}}) \in \mathbb{R}^{N\times N}$. The product

$$J_x\Psi(x_{\mathrm{H}}^\star)^{\frac{1}{2}} S(x_{\mathrm{H}}^\star) J_x\Psi(x_{\mathrm{H}}^\star)^{\frac{1}{2}} \tag{4.20}$$

is a congruence transformation that preserves the matrix inertia [93]. Now, the above product is symmetric, and in particular, the Jacobian $J_x F(x_{\mathrm{H}}^\star)$ is Hurwitz if and only if

$$S(x_{\mathrm{H}}^\star) = S(x_{\mathrm{H}}^\star)^\top \prec 0 \tag{4.21}$$

Then the thesis of the statement follows from [78, Theorem 4.15].

(ii) Note that $J_x\Psi(x_{\mathrm{H}\nu})$ is a diagonal matrix with i-th diagonal entry equal to $\psi'([x_{\mathrm{H}\nu}]_i)$. Since $\xi^\nu \in \{-1,+1\}^N$, we know that each entry of $x_{\mathrm{H}\nu} = \gamma_\nu \xi^\nu$ is equal to $\pm\gamma_\nu$. Next, we note that, for all $z \in \mathbb{R}$,

$$\psi(z) = -\psi(-z) \implies \psi'(-z) = \frac{d}{dz}\Big(-\psi(-z)\Big) = \frac{d}{dz}\psi(z) = \psi'(z) \tag{4.22}$$

Therefore, $J_x\Psi(x_{\mathrm{H}\nu}) = \psi'(\gamma_\nu) I_N$ and, in turn, the local exponential stability condition eq. (4.21) reads as

$$\frac{1}{N}\sum_{\mu=1}^{P} \alpha_\mu \xi^\mu \xi^{\mu\top} \prec \frac{1}{\psi'(\gamma_\nu)} I_N \tag{4.23}$$

Let $M := \frac{1}{\sqrt{N}}\left[\xi^1 \cdots \xi^P\right]$ and observe that condition eq. (4.23) can be written as in compact form as

$$\lambda_{\max}(M\ \mathrm{diag}(\alpha_1,\dots,\alpha_P)M^\top) < \frac{1}{\psi'(\gamma_\nu)}$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a symmetric matrix. Letting $A := M\ \mathrm{diag}(\alpha_1,\dots,\alpha_P)^{\frac{1}{2}}$ and using the identity $\lambda_{\max}(AA^\top) = \lambda_{\max}(A^\top A)$, the left-hand side of the previous inequality equals

$$\begin{aligned}\lambda_{\max}\left(A^\top A\right) &= \lambda_{\max}\left(\mathrm{diag}(\alpha_1,\dots,\alpha_P)^{\frac{1}{2}} M^\top M\, \mathrm{diag}(\alpha_1,\dots,\alpha_P)^{\frac{1}{2}}\right) \\ &= \lambda_{\max}(\mathrm{diag}(\alpha_1,\dots,\alpha_P)) = \max_{\mu=1,\dots,P} \alpha_\mu \end{aligned} \tag{4.24}$$

where we used the fact that $M^\top M = I$.

□

Stability Threshold From the previous stability results, it is natural to wonder whether there exists a critical parameter $\alpha_{\text{stability}}$ that determines the stability of the memory patterns, and how it is possible to quantify it. In what follows, to lighten the notation, we will refer to $\alpha_{\text{stability}}$ as $\alpha^\star$. For simplicity and without loss of generality, we impose the following order on the saliency weights

$$\alpha_1 > \cdots > \alpha_P \tag{4.25}$$

We now present the main theorem of the section, in which we prove the existence of a critical threshold $\alpha^\star$ for the stability of the retrievable memories of the IDP *voltage* model.

Theorem 4.2.5 (Critical saliency $\alpha^\star$). *Consider the dynamical system (4.6) and let $\alpha_1 > 1$. Then*

(i) there exists at least one locally exponentially stable equilibrium,

(ii) let $\alpha^\star := \frac{\gamma^\star}{\psi(\gamma^\star)}$ where $\gamma^\star > 0$ is such that $\psi'(\gamma^\star) = \frac{1}{\alpha_1}$. Then

$$x_{H\nu} = \gamma_\nu \xi^\nu \text{ is a locally exponentially stable equilibrium} \iff \alpha_\nu > \alpha^\star \tag{4.26}$$

Proof.

(i) We define $\alpha(\gamma) := \gamma/\psi(\gamma)$ and let $\gamma_1 > 0$ be the positive solution of

$$\alpha_1 = \alpha(\gamma_1) = \frac{\gamma_1}{\psi(\gamma_1)} \tag{4.27}$$

and observe that

$$\alpha(\gamma) < \frac{1}{\psi'(\gamma)} \quad \forall \gamma > 0 \tag{4.28}$$

As a matter of fact, it holds

$$\frac{1}{\psi'(\gamma)} - \alpha(\gamma) = \frac{\psi(\gamma) - \gamma\psi'(\gamma)}{\psi'(\gamma)\psi(\gamma)} > 0 \quad \forall \gamma > 0 \tag{4.29}$$

since $\psi'(\gamma)$, $\psi(\gamma)$, and $f(\gamma) := \psi(\gamma) - \gamma\psi'(\gamma)$ are strictly positive for all $\gamma > 0$. The fact that $f(\gamma) > 0$ for all $\gamma > 0$ follows by noting that $f(0) = 0$ and $f'(\gamma) = -\gamma\psi''(\gamma) > 0$, $\forall \gamma \neq 0$. Then it follows from eqs. (4.27), (4.28)] that

$$\psi'(\gamma_1) < \frac{1}{\alpha_1} \tag{4.30}$$

and thus, using Theorem 4.2.4, there exists at least one locally exponentially stable equilibrium point for eq. (4.6).

(ii) Since $\alpha_1 > 1$, then using eqs. (4.8), and (4.9)] we have that $\exists! \gamma^\star > 0$ such that

$$\psi'(\gamma^\star) = \frac{1}{\alpha_1} \tag{4.31}$$

Moreover, from eq. (4.9) it is immediate to observe that

$$\psi'(\gamma_\rho) < \frac{1}{\alpha_1} \iff \gamma_\rho > \gamma^\star \tag{4.32}$$

Notice now that

$$\alpha'(\gamma) = \frac{f(\gamma)}{\psi(\gamma)^2} > 0 \quad \forall \gamma > 0 \tag{4.33}$$

where $\alpha(\gamma)$ and $f(\gamma)$ have been defined in the proof of the first statement. The previous equation implies that $\alpha(\gamma)$ is monotonically increasing for $\gamma > 0$. Consequently, the stability condition of Theorem 4.2.4 is satisfied if and only if $\alpha_\nu = \alpha(\gamma_\nu) > \alpha(\gamma^\star) = \alpha^\star$.

□

Thus, the third and final core property of the IDP *voltage* model is

(iii) if multiple retrievable memories exist under the current input, then there exists a stability threshold $\alpha_{\text{stability}} > 1$ such that each memory $x_{\text{H}\mu}$ is stable if and only if $\alpha_\mu > \alpha_{\text{stability}}$.

We now want to exploit the definition of $\alpha_{\text{stability}} = \alpha^\star = \gamma^\star/\psi(\gamma^\star)$ (see Fig. 4.4 for intuition on $\gamma^\star$ and on the stability threshold $\alpha^\star$) to provide the reader with an heuristic of how a strong dominance of a saliency weight α_1 may disrupt the stable retrieval of any memory but $x_{\text{H}1} = \gamma_1 \xi^1$. Since we know that $\gamma^\star > 0$ is the solution of the equation

$$\psi'(\gamma^\star) = \frac{1}{\alpha_1} \tag{4.34}$$

then as $\alpha_1 = \alpha(\gamma_1)$ increases, we have that $\psi'(\gamma^\star)$ decreases, which by eq. (4.9) is possible only if $\gamma^\star$ is increasing as well. It is then evident from eq. (4.8) that when $\gamma^\star$ increases, so does $\alpha^\star = \alpha(\gamma^\star)$. Thus, we have proved that the growth of the stability threshold $\alpha^\star$ depends exclusively on the dominant saliency weight α_1, and moreover that the two are proportionally related to each other. Consequently, the sole growth of the dominant saliency weight α_1 is sufficient to push the stability threshold $\alpha^\star$ above all the other saliency weights, so that the only stable memory left is indeed $x_{\text{H}1} = \gamma_1 \xi^1$.

4.2.3 Levelling the Energy through input modulation

One of the most interesting features of the *voltage* model is the existence of an energy function that details how the trajectory of the system converge to the existing equilibria. The energy of the IDP model is similar to the energy of the classic *voltage* model, under the assumption that $u \in \mathbb{R}^N$ is constant. Specifically, we introduce the following definition.

Definition 4.2.6 (Energy of the IDP *voltage* model). Consider the dynamics (4.6). Then the associated energy function is

$$\text{E}_H(x_\text{H}; W(u)) = -\frac{1}{2}\Psi(x_\text{H})^\top W(u)\Psi(x_\text{H}) + x_\text{H}^\top \Psi(x_\text{H}) - \sum_{i=1}^{N} \int_0^{x_{\text{H}i}} \psi(z)\, dz \tag{4.35}$$

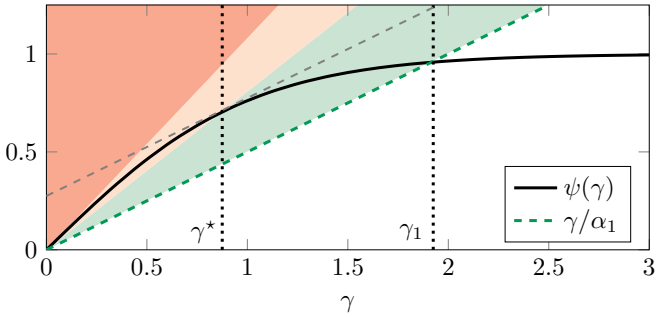


Figure 4.4: **Visual representation of the existence and stability results presented in the previous sections.** The saliency weights α such that the dissipation line γ/α belongs to: (1) the green region are associated to stable retrievable memories; (2) the yellow region are associated to unstable retrievable memories; (3) the red region to non-retrievable memories.

and we additionally define the energy per node as

$$\varepsilon(x_{\mathrm{H}}; u) = \frac{\mathrm{E}_H(x_{\mathrm{H}}; W(u))}{N} \tag{4.36}$$

The energy per node is not a necessary quantity for the following analysis, but it will substantially lighten the notation. The following result characterizes the energy landscape of the IDP *voltage* model.

Theorem 4.2.7 (Energy wells). *Given a fixed input $u \in \mathbb{R}^N$, the energy function of the IDP* voltage *model (4.6) has negative total time derivative*

$$\frac{d\mathrm{E}_H}{dt}(x_H(t); W(u(t)))|_{u(t)\equiv u} \leq 0 \tag{4.37}$$

Moreover, given two equilibria $x_{H\nu} = \gamma_\nu \xi^\nu$, $x_{H\rho} = \gamma_\rho \xi^\rho$, $\gamma_\nu, \gamma_\rho > 0$, associated to saliency weights $\alpha_\nu, \alpha_\rho > 1$ such that $\alpha_\nu > \alpha_\rho$, then

$$\varepsilon(x_{H\nu}; u) < \varepsilon(x_{H\rho}; u) \tag{4.38}$$

Proof. Consider the total time derivative of the energy function associated to the IDP *voltage*

model

$$\begin{aligned}
&\frac{d\mathrm{E}_H}{dt}(x_{\mathrm{H}}(t);W(u(t)))|_{u(t)\equiv u}\\
&=\frac{\partial \mathrm{E}_H}{\partial x_{\mathrm{H}}}(x_{\mathrm{H}}(t);W(u(t)))|_{u(t)\equiv u}\frac{dx_{\mathrm{H}}}{dt}(t)+\frac{\partial \mathrm{E}_H}{\partial u}(x_{\mathrm{H}}(t);W(u(t)))|_{u(t)\equiv u}\underbrace{\frac{du}{dt}(t)}_{\equiv 0}\\
&=\frac{\partial \mathrm{E}_H}{\partial x_{\mathrm{H}}}(x(t);W(u))\frac{dx_{\mathrm{H}}}{dt}(t)=-\frac{dx_{\mathrm{H}}}{dt}(t)^\top D\Psi(x_{\mathrm{H}}(t))\frac{dx_{\mathrm{H}}}{dt}(t)\le 0
\end{aligned}\tag{4.39}$$

where we have used the fact that $D\Psi(x_{\mathrm{H}})\succ 0$, and consequently $\dot{\mathrm{E}}=0 \iff \dot{x}_{\mathrm{H}}=0$.
We now want to explicitly compute the energy associated to a generic fixed point $x_{\mathrm{H}\nu}=\gamma_\nu\xi^\nu,\ \gamma_\nu>0$, taking into account our homogeneity assumptions on the activation function. For notational simplicity, we will use $\mathrm{E}_H(x_{\mathrm{H}};W(u))=\mathrm{E}_H(x_{\mathrm{H}})$ and $\varepsilon(x_{\mathrm{H}},u)=\varepsilon(x_{\mathrm{H}})$

$$\begin{aligned}
\mathrm{E}_H(x_{\mathrm{H}\nu})&=-\frac{1}{2}\Psi(x_{\mathrm{H}\nu})^\top W(u)\Psi(x_{\mathrm{H}\nu})+x_{\mathrm{H}\nu}^\top\Psi(x_{\mathrm{H}\nu})-\sum_{i=1}^N\int_0^{x_{\mathrm{H}\nu i}}\psi(z)\,dz\\
&=-\frac{1}{2}\psi(\gamma_\nu)\xi^{\nu\top}W(u)\xi^\nu\psi(\gamma_\nu)+\gamma_\nu\xi^{\nu\top}\xi^\nu\psi(\gamma_\nu)-N\int_0^{\gamma_\nu}\psi(z)\,dz\\
&=-\frac{N}{2}\alpha_\nu\psi(\gamma_\nu)^2+N\gamma_\nu\psi(\gamma_\nu)-N\int_0^{\gamma_\nu}\psi(z)\,dz
\end{aligned}\tag{4.40}$$

$$\tag{4.41}$$

Considering now the fixed point condition

$$\frac{\gamma_\nu}{\alpha_\nu}=\psi(\gamma_\nu)\tag{4.42}$$

and switching to the energy per node

$$\varepsilon(x_{\mathrm{H}\nu})=\frac{1}{2}\gamma_\nu\psi(\gamma_\nu)-\int_0^{\gamma_\nu}\psi(z)\,dz\tag{4.43}$$

$$=\frac{1}{2}\frac{\gamma_\nu^2}{\alpha_\nu}-\int_0^{\gamma_\nu}\psi(z)\,dz\tag{4.44}$$

Now, considering

$$F(\gamma,\alpha)=\psi(\gamma)-\frac{\gamma}{\alpha}\tag{4.45}$$

we know that $F(\gamma, \alpha_\nu) > 0, \forall \gamma \in (0, \gamma_\nu)$, from which we obtain

$$\varepsilon(x_{H\nu}) = \frac{1}{2}\frac{\gamma_\nu^2}{\alpha_\nu} - \int_0^{\gamma_\nu} \psi(z)\, dz \tag{4.46}$$

$$< \frac{1}{2}\frac{\gamma_\nu^2}{\alpha_\nu} - \int_0^{\gamma_\nu} \frac{z}{\alpha_\nu}\, dz = 0 \tag{4.47}$$

Considering now $\alpha_\nu, \alpha_\rho > 1,\ \alpha_\nu > \alpha_\rho$, we know that $F(\gamma_\rho, \alpha_\rho) = F(\gamma_\nu, \alpha_\nu) = 0$. Since F is monotonically increasing in α and decreasing in γ, then $\gamma_\nu > \gamma_\rho$. Therefore, it follows that

$$\varepsilon(x_{H\nu}) - \varepsilon(x_{H\rho}) = \frac{1}{2}\left(\frac{\gamma_\nu^2}{\alpha_\nu} - \frac{\gamma_\rho^2}{\alpha_\rho}\right) - \int_{\gamma_\rho}^{\gamma_\nu} \psi(z)\, dz \tag{4.48}$$

$$< \frac{1}{2}\left(\frac{\gamma_\nu^2}{\alpha_\nu} - \frac{\gamma_\rho^2}{\alpha_\rho}\right) - \int_{\gamma_\rho}^{\gamma_\nu} \frac{z}{\alpha_\nu}\, dz \tag{4.49}$$

$$= -\frac{1}{2}\left(\frac{\gamma_\rho^2}{\alpha_\rho} - \frac{\gamma_\rho^2}{\alpha_\nu}\right) < 0 \tag{4.50}$$

□

From Fig. 4.5 it is possible to observe how the increase of a specific saliency weight considerably deepens the energy well associated to the corresponding memory. The varying depth of the energy wells provides a clear graphical interpretation of the stability of the different memories encoded in the IDP *voltage* model. Specifically, considering panel Fig. 4.5D, it is easy to observe how an appropriate choice of an additive perturbation $\delta u \in \mathbb{R}^N$ for the dynamics (4.6) may easily push the state trajectory from the flatter to the deeper energy well. At the same time, the same perturbation δu may be too weak to reverse the previous jump, thus leaving the state trajectory inevitably confined to the basin of attraction of the memory associated to the deeper energy well. Therefore, the dominancy of one saliency weight over the others enhances the overall robustness of the model, promoting the retrieval of the correct memory even in the presence of perturbations and noise.

Remark 4.2.8. The classic *voltage* model is a specific sub-case of the IDP model, where the input impinging on the synaptic matrix is characterized by $\alpha_\mu \equiv k > 1$, for all μ.

The IDP *voltage* model naturally endows us with a direct interpretation of when the input is clearly understandable, and when instead is too vague to evoke any kind of retrieval. As previously presented, memories exist (are retrievable) only if their respective saliency weights in the input are at least unitary. Thus, for any input that is well-mixed and such that $\alpha_\mu < 1$ for all μ (Fig. 4.5A) the dynamics will converge to the origin, which we call confusion state. This state exists only in the IDP model, as the classic model always has all the retrievable memories

irrespective of the input. The possibility of falling into a confusion state on the basis of what is experienced seems quite plausible, as the clarity of what we sense in our everyday experience has

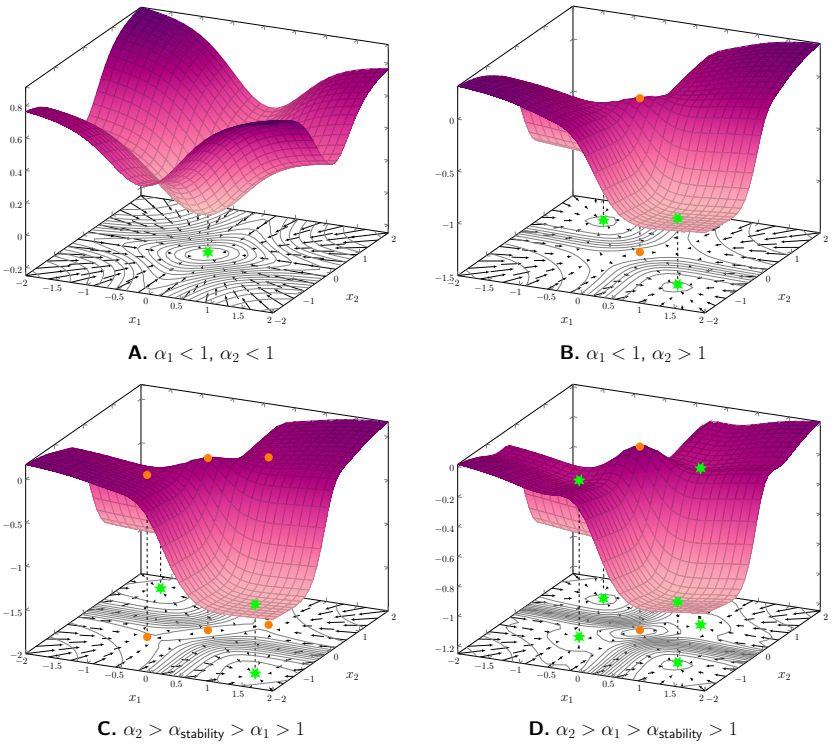


Figure 4.5: **Energy landscapes for IDP *voltage* model for varying saliency weights.** Stable and unstable equilibria are depicted as green stars and orange dots, respectively. Recall the existence threshold is $\alpha_{\text{existence}} = 1$ and, when multiple memories exist in the input, the stability threshold satisfies $\alpha_{\text{stability}} > 1$. (A) "no memories" $\alpha_1 < 1$, $\alpha_2 < 1$: when no memory is sufficiently strong in the input, the only global minimum is at the origin and it is globally attractive for the dynamics. This situation corresponds to a confusion state for the network, in which the input is not strong enough to evoke any retrieval. (B) "one memory" $\alpha_1 < 1$, $\alpha_2 > 1$: when the saliency weight for precisely one memory is above $\alpha_{\text{existence}}$, two symmetric equilibria appear corresponding to the memory and they are attractive from almost all initial conditions. In this case, the origin is a saddle point. (C) "one stable and one unstable memory" $\alpha_2 > \alpha_{\text{stability}} > \alpha_1 > 1$: two symmetric equilibria appear corresponding to the stable memory and they are attractive from almost all initial conditions. The other two symmetric equilibria are saddle points and the origin becomes an unstable maximum. (D) "two stable memories" $\alpha_2 > \alpha_1 > \alpha_{\text{stability}}$: four symmetric stable equilibria appear, corresponding to the two memories of the model. The memory associated to the dominant saliency weight carves deeper valley in the energy landscape than the other memory. When $\alpha_1 \approx \alpha_{\text{stability}}$, the shallowness of the valley associated to the first memory facilitates outward jumps due to stochastic fluctuations.

a clear bearing on what we recall. Finally, we want to address the following question: does the initial condition still matter? The energy function $\mathrm{E}(x_{\mathrm{H}}; W)$ ensures convergence to one of the stable memories from any initial conditions. On the other hand, the stability threshold $\alpha_{\text{stability}}$ determines how many stable memories we have for a given input u. Thus, the fewer the stable memories for the IDP *voltage* model, the wider their basin of attraction. In the extremal case of a single stable memory, the whole $\mathbb{R}^N$ is its basin of attraction, and any initial condition leads to the correct retrieval. Thus, appropriately initializing the system becomes irrelevant, especially in the presence of perturbations that push the dynamics out of shallow energy minima (see Fig. 4.5D).

4.2.4 Biological Plausibility of the IDP Hopfield Model

The IDP model aims to provide a high level description of the mechanisms underlying short-term plasticity within individual neural modules in response to stimuli originating from various cortical regions. Similarly to the gated synaptic theory of Tsodyks et al. [92], [94], [95], synapses are not static entities, but rather fast dynamic components that transiently mirror adjustments in the allocation of neurochemical resources. The IDP Hopfield model can be interpreted as a selective diffusion model, where the channels associated to the stimulus - possibly having directed axon bundles as their neurological proxy - either amplify or attenuate the conductance of specific sub-clusters of synapses. The fast selective tuning of synapses shapes the probability associated with the retrieval of any stored pattern. Specifically, the IDP model may present a mechanistic account of the differential influence of the mossy fibers (MF) and of the perforant pathways (PP), both providing a different manipulation of the same signal to CA3. It is well known [96] that the two signals differ in their magnitude, with the former being significantly stronger [97], albeit sparser, than the latter. As can be observed from Fig. 5.1(C), the proposed mechanism is a robust orthogonalizer of the network activity despite the presence of background noise and a mixed input. Therefore, stimulus-driven modulation of the synaptic matrix may be speculatively associated to the projective MF, which induce similar plasticity events in the CA3 network and confine potential retrievals independently of any biases originating from the PP.

In some biological models, signals represent firing rates instead of membrane potentials. In this case the Hopfield model (including the IDP Hopfield model [4.6]) is inadequate at correctly describing the system behavior because it does not preserve the positivity of signals. Motivated by this observation, we also consider the following IDP firing rate model with a non-negative activation function $\Phi : \mathbb{R}^N \to [0,1]^N$:

$$\begin{cases} \dot{x}_{\mathrm{F}}(t) = -x_{\mathrm{F}}(t) + \Phi\left(W(u(t))x_{\mathrm{F}}(t)\right) \\ x_{\mathrm{F}}(0) \in [0,1]^N \end{cases} \tag{4.51}$$

where activity is sparse and confined in the hyper-interval $[0,1]^N$. Although the firing rate model is less amenable to analytical treatment due to the lack of symmetry in the activation function, a variation of the proposed mechanism is still able to induce stimulus driven transitions among different prototypical memories. However, as observed in [98], under certain conditions the sharing of representational units among different patterns leads to their co-activation.

4.3 A modern interpretation

As discussed in the introduction, the IDP model can be naturally reformulated within the modern Hopfield framework introduced in [25]. In this formalism, a recurrent neural network is deconstructed into two interacting layers without intralayer connections. The resulting bi-layer dynamics are given by

$$\tau_x \dot{x} = -x + M_x \Psi_x(y), \tag{4.52}$$

$$\tau_y \dot{y} = -y + M_y \Psi_y(x). \tag{4.53}$$

Here, the dynamics of the x-layer (feature layer) depend only on the activity of the y-layer (memory layer), and vice versa, with no self-interactions within layers. Under the assumption of separated timescales - specifically, in the limit $\tau_y \to 0$ - the system reduces to

$$\tau_x \dot{x} = -x + M_x \Psi_x(M_y \Psi_y(x)) \tag{4.54}$$

which recovers the most general form of a recurrent neural network, including the *voltage* and *firing rate* models.

With respect to the IDP *voltage* model, the following tripartite architecture characterizes a more general model that captures the interaction of the activity in the memory layer $y \in \mathbb{R}^P$ and in the saliency layer $z \in \mathbb{R}^P$. The combined information is then exploited to drive the retrieval in the feature layer $x \in \mathbb{R}^N$. In summary, the modern Hopfield reformulation of the IDP model is

$$\tau_x \dot{x} = -x + M_x[z \odot \Psi_x(z \odot y)] \tag{4.55}$$

$$\tau_y \dot{y} = -y + M_y \Psi_y(x) \tag{4.56}$$

$$\tau_z \dot{z} = -z + M_z \Psi_z(u) \tag{4.57}$$

where the symbol $\odot$ is the Hadamard entrywise product, namely $(z \odot y)_i = y_i z_i$, and Ψ_z is an activation function implementing either a linear or non-linear processing of the input. Notice that when $M_x = M$ and $M_y = M_z = M^\top$ with $M = N^{-\frac{1}{2}}[\xi^1 \cdots \xi^P] \in N^{-1/2}\{-1,1\}^{N \times P}$, and Ψ_x, Ψ_z are identity functions, the equations (4.55),(4.56), (4.57) reduce to the IDP *voltage*

model (4.6) in the limits $\tau_y \to 0$ and $\tau_z \to 0$, with $\alpha = z \odot z$. Unraveling the activity of the IDP *voltage* network into the distinct components allows for a better qualitative understanding of the layers' contribution (see Fig. 4.6 for a block representation). The memory layer serves the function of pooling layer, projecting the activity of the feature layer into a similarity space. The pooled activity is then modulated by the input decomposition. It is clear then that the combination of pooling and modulation implements a natural trade-off between past internal activity and externally incoming information.

In the model discussed so far we have the input processing $M_z \Psi_z(u) = M^\top u$, because it allows a direct logical and visual interpretation of the relationship between the input decomposition and the network response. However, in a realistic neuronal system the relationship may not be so simple and interpretable. Indeed, the most general case would be that where $M_z \in \mathbb{R}^{P \times K}$ and $u \in \mathbb{R}^K$ is a generic input. This is the case when the input originates from a different cortical region possibly having a different output dimensionality, and therefore the relation with the prototypical memories is more subtle. A discussion of how the Ψ_z output channels are related to the memories stored in the recurrent network is beyond the scope of this article, but remains an important problem open to original solutions.

Finally, the framework of eqs. (4.55),(4.56), and (4.57) provides a dynamic description of a

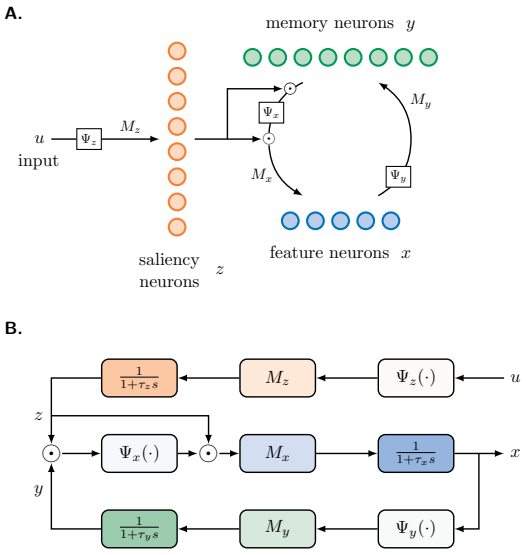


Figure 4.6: **Illustration of the modern Hopfield reformulation of eqs. (4.55),(4.56),(4.57) for the IDP *voltage* model with input filter.** The symbol $\odot$ denotes a Hadamard entrywise product. (A) Neural network representation with interconnected layers and synaptic weights M_z, M_x, and M_y. (B) Block diagram representation, where each block $\frac{1}{1+\tau s}$, with $\tau = \tau_z, \tau_x, \tau_y$, denotes a low-pass filter with cutoff frequency $1/\tau$, each block M_z, M_x, and M_y denotes a matrix-vector multiplication, and each block $\Psi_z(\cdot)$, $\Psi_x(\cdot)$, and $\Psi_y(\cdot)$ denotes an activation function.

biased self-attention mechanism in a transformer, generalizing the result (1.47) of [45]. Indeed, choosing $\Psi_y(\cdot) = \mathrm{Id}(\cdot)$, $\Psi_x(\cdot) = \mathrm{softmax}(\cdot)$, $M_x = M_y^\top = M$, and taking $\tau_x = 1$, $\tau_y \to 0$, $\tau_z \to 0$ yields

$$\dot{x} = -x + M[z \odot \mathrm{softmax}(z \odot M^\top x)] = -\nabla_x \mathrm{E}_{tr}(x; z) \tag{4.58}$$

$$\mathrm{E}_{tr}(x; z) = -\frac{1}{\beta} \log \left(\sum_{i=1}^{N} e^{\beta z_i (M^\top x)_i} \right) + \frac{x^\top x}{2} \qquad \beta > 0 \tag{4.59}$$

Constraining $z(t) \equiv \mathbb{1}_N$ we recover exactly the dynamic self-attention mechanisms studied in the provided references. Differently from the standard implementation of the attention mechanism, augmenting the dynamics with eq. (4.57) may help in controlling the retrieval of keys as a function of continuously streamed external biases pre-multiplying the queries, hence in a more online fashion.

4.4 Conclusion

Achieving robust behavior in multistable systems Multistability is a distinctive feature of the classical *voltage* model, required for its associative memory functionality. Yet, multistability is typically avoided in controlled engineering systems because it potentially leads to fragile and unpredictable behavior. While complex systems, such as the electric power grid, may in fact exhibit multiple stability regions, only one of them corresponds to proper grid functioning. It is the responsibility of the grid control infrastructure to steer the system away from undesired stability regions. The proposed IDP *voltage* model fulfills this same responsibility in the context of memory retrieval: when the input is not ambiguous, the neural state is reliably driven to the correct stability region, achieving a robust and predictable memory system.

Energy shaping in the IDP *voltage* model The IDP *voltage* model presents a simple yet effective explanation of how a direct input-driven modulation of the synapses can enrich the dynamic range of recurrent neural networks. The input driven adjustments of the synaptic couplings between neurons enforce a clear memory hierarchy, with single memories existing only if sufficiently stimulated. Furthermore, the input decomposition changes the stability properties of single memory patterns. Through the existence of the threshold $\alpha_{\text{stability}}$, a memory that was stable at a certain time instant can suddenly become a saddle point, and thus allow the network dynamics to roll towards another memory, as proposed by Karuvally et al. in [99]. During the retrieval process, the saliency attributed to individual memories via input decomposition reshapes the energy landscape of the model. This process deepens the wells associated with the most dominant input components while flattening the others. Consequently, the presence of exogenous

noise is enough to drive the dynamics towards the deepest basin of attraction and confine the network activity within it. It is worth mentioning that a model similar to the IDP *voltage* has been recently numerically studied in [100] with the aim of implementing sequential memory retrieval. In this context, the dynamics and distribution of the saliency weights reflect some previous association among prototypical memories, and confine the network activity to limit cycles.

Future work and implications The present study lays the foundation for future research aimed at fully analyzing the biologically plausible *firing rate* version of the IDP model. This study also opens pathways for the empirical validation of input-dependent associative memory models and a comprehensive characterization of learning dynamics, encompassing both long-term memory formation and short-term modulation of saliency weights. Specifically, future works will need to consider a complete theoretical investigation of the IDP *firing rate* model in a low neural activity regime compatible with experimental evidence [98]. A thorough understanding of the dynamics underlying memory retrieval in the *firing rate* model may in fact guide experimental researchers in their inquiries, and successful validation of the model may provide a solid ground for the understanding of high level cognitive processes. Most importantly, our formalization of the IDP *voltage* model paves the way for a thorough investigation into its underlying learning dynamics. Developing an effective, biologically plausible learning mechanism—such as one grounded in the Hebbian paradigm—could position the IDP *voltage* model, rooted in the modern Hopfield architecture, as an efficient biologically-inspired machine learning model.

In addition, jointly solving the problem of long-term learning and short-term modulations may provide valuable answers on the multiplicity of timescales observed throughout biological learning processes. Notably, there is already ample evidence [101] that heterogeneity in functional features and timescales may be the biological solution to the problems of catastrophic forgetting and adaptive learning. The advancement of such comprehension holds relevant implications, extending its impact beyond neuroscience and onto the domain of machine learning. Within the machine learning field, the mentioned problems prevent virtual and embedded agents [87] from deployment in contexts that require both cognitive flexibility and preservation of past knowledge. The successful combination of these two elements epitomizes the ultimate aspiration of continual learning [88], [102], an emergent machine learning paradigm that seeks to meaningfully integrate present and past data and yields the potential for great technological advancement.

This chapter is based on the work [103] published in the journal *Science Advances*. The next chapter extends the IDP *voltage* model by incorporating stochasticity, highlighting how noise facilitates smoother transitions between memories. We also provide a rigorous analysis of the associated stochastic process, demonstrating how the probability measure concentrates around the deepest minima of the energy landscape.

Appendix In this appendix, I justify how the choice of a piecewise continuous and bounded external input $u(t) \in \mathbb{R}^N$ still guarantess the existence and uniqueness of the solution to the *voltage* equation. It is well known (see [78, Theorem. 2.3, Ch. 2]) that the Cauchy problem

$$\begin{cases} \dot{x} = \mathcal{H}(t, x) \\ x(0) = x_0 \in \mathbb{R}^N \end{cases} \tag{4.60}$$

admits a global unique solution $x(t)$ if $\mathcal{H}(t, x)$ satisfies the following conditions:

(i) $\mathcal{H}(t, x)$ is piecewise continuous in t.

(ii) There exists a constant $L > 0$ such that $\|\mathcal{H}(t, x) - \mathcal{H}(t, y)\|_2 \leq L\|x - y\|_2$ for all $t \in \mathbb{R}$ and $x, y \in \mathbb{R}^N$, where $\|\cdot\|_2$ denotes the 2-norm.

(iii) There exists a maps $\zeta : \mathbb{R}^N \to \mathbb{R}$ such that $\|\mathcal{H}(t, x)\|_2 \leq \zeta(x)$.

In our case we have that $\mathcal{H}(t, x_{\mathrm{H}}) = -x_{\mathrm{H}} + W(u(t))\Psi(x_{\mathrm{H}})$. Notice that the previous condition (i) is automatically satisfied when $u(t)$ is piecewise continuous. As far as condition (ii), observe that

$$\|\mathcal{H}(t, x) - \mathcal{H}(t, y)\|_2 \leq \|x - y\|_2 + \|W(u(t))\|_2 \|\Psi(x) - \Psi(y)\|_2$$

Notice now that

$$\|W(u(t))\|_2 = \max_{\nu=1,\dots,P} \{\alpha_\nu\} \leq \max_{\mu=1,\dots,P} \{\|\xi^\mu\|_2 \|u(t)\|_2\} \leq \sqrt{N} \max_t \{\|u(t)\|_2\}$$

Noticing moreover that $\|\Psi(x) - \Psi(y)\|_2 \leq \|x - y\|_2$ we have that also condition (ii) holds. Finally, observe that in our case condition (ii) implies condition (iii) since we have that $\mathcal{H}(t, \mathbb{0}_N) = 0$.

5

Stochastic IDP *voltage* dynamics to escape shallow minima

From Chapter I to Chapter VI, this thesis has primarily examined deterministic dynamical neural networks, as their mathematical structure provides a clear foundation for analysing stability, retrieval, and the geometry of memory representations. Yet modern developments in diffusion-based machine learning models compel us to broaden this framework. Contemporary architectures increasingly rely on stochasticity as an active computational ingredient, suggesting that a complete theory of associative memory should also encompass settings in which neural dynamics are continuously perturbed by noise. Extending classical results to stochastic dynamics not only enriches the theoretical picture but may also illuminate how information is processed in models where randomness and transport play pivotal functional roles.

Within this context, it is natural to revisit the deterministic *voltage* models discussed earlier. Both the classic formulation (4.4) and the IDP *voltage* model (4.6) successfully retrieve the correct memory when presented with the same well-mixed external input (see Fig. 4.2(B,C)), which might tempt one to view them as effectively equivalent. However, this apparent similarity is a consequence of their idealized nature: in both cases, neural units evolve in isolation from the pervasive background fluctuations that characterize biological circuits and, increasingly, modern machine learning models. As a result, deterministic analyses alone cannot reveal how robust these retrieval mechanisms remain when confronted with stochastic perturbations.

To address this limitation, we introduce the stochastic dynamics associated with the IDP *voltage* model,

$$d\mathbf{X}_{\mathrm{H}}(t) = \left[-\mathbf{X}_{\mathrm{H}}(t) + W(u)\Psi(\mathbf{X}_{\mathrm{H}}(t))\right] dt + \omega, d\mathbf{B}(t), \tag{5.1}$$

where $\omega \geq 0$ denotes the noise amplitude and $\mathbf{B}(t)$ is an n-dimensional Brownian motion. This formulation enables us to probe how noise interacts with the energy landscape, how it modulates

the basin structure of stored patterns, and to what extent retrieval survives in regimes where perturbations are no longer negligible.

From Fig. 5.1A observe that when the system is insulated ($\omega = 0$) and for highly mixed inputs the retrieval process remains stuck on a previous memory as long as the associated saliency weight is above the stability threshold $\alpha_{\text{stability}}$. Equivalently, the dynamics are entrapped in a shallow minimum of the energy $\mathrm{E}_H(x_{\mathrm{H}}; W(u))$. This behavior is undesirable, as it impairs the system's ability to correctly regulate its response under conditions of high input ambiguity, and thereby undermine its reliability. We here argue that the IDP model is not only robust to noise, but actually benefits from and exploits it as a mean to escape the shallow minima and to achieve efficient retrieval in vague contexts. Indeed, as observable from Fig. 5.1B, relatively high amplitude noise ($\omega = 4$) favors the consecutive retrieval of different memories even when the inputs are highly mixed. Consistent with our intuition in Fig.4.5D, the noise drives the retrieval process out of shallow energy valleys and into the deepest energy minima. Thus, the stochastic IDP model embodies a mechanism that actively prioritizes the most relevant features in its environment while suppressing the processing of background elements, mirroring the psychological phenomenon of selective attention [104].

The observed robustness to noise for the IDP *voltage* model entails the existence of a range of noise amplitudes ω leading to correct sequential retrieval. It is then natural to wonder whether there exists optimal value of ω that minimize the time elapsing between input onset and the next retrieval. As observable from Fig. 5.1D, for an insulated system ($\omega = 0$) the time elapsing between input onset and the decay of the past memory retrieval curve is quite long, but once initiated the transient is very fast. Despite the successful switching, the prolonged dwelling on the previous memory seems somewhat undesirable for systems that have evolved to efficiently and readily adapt to their environment. When the system has relatively high noise amplitude ($\omega = 4$), the decay of the past memory retrieval curve is very swift, but the transient towards the next retrieval is fairly slow. Slow tracking of a change in the input seems equally undesirable from an adaptive perspective. Instead, for intermediate values of the amplitude parameter ($\omega = 2$) the model displays both a relatively fast decay of the past memory retrieval curve and a swift transient towards the next retrieval. In particular, it can be observed that intermediate noise amplitudes result in shorter times between input onset and retrieval of the correct memory with respect to both insulated and highly noisy systems.

In addition, brief prioritization of past information following a change in the input allows the system to correct transient errors, or glitches. Indeed, for desirable values of the noise amplitude ω the IDP model adequately weights the importance of past information, so that glitches are neglected. To test this hypothesis, we conduct a numerical experiment where the dominant saliency weight undergoes a brief perturbation (see the green glitch at $t = 8$ in Fig. 5.1C).

The IDP *voltage* model effectively suppresses these transient fluctuations, maintaining retrieval stability. This robustness arises from the multiplicative mechanism in eq. (4.55), which prevents abrupt disruptions in retrieval dynamics by integrating present external information against past feedback activity.

The presented results may additionally benefit from a more rigorous study of the properties of the stochastic *voltage* model. A valuable tool towards this aim is the Fokker-Planck equation [105]–[107], which describes the time evolution of the probability measure associated to eq. (5.1). In [108] Wong studies a particular case of the stochastic *voltage* model, where the associated probability measure converges to a stationary distribution that explicitly depends on the energy $\mathrm{E}(x_{\mathrm{H}}; W(u))$, revealing that probability mass concentrates around its minima.

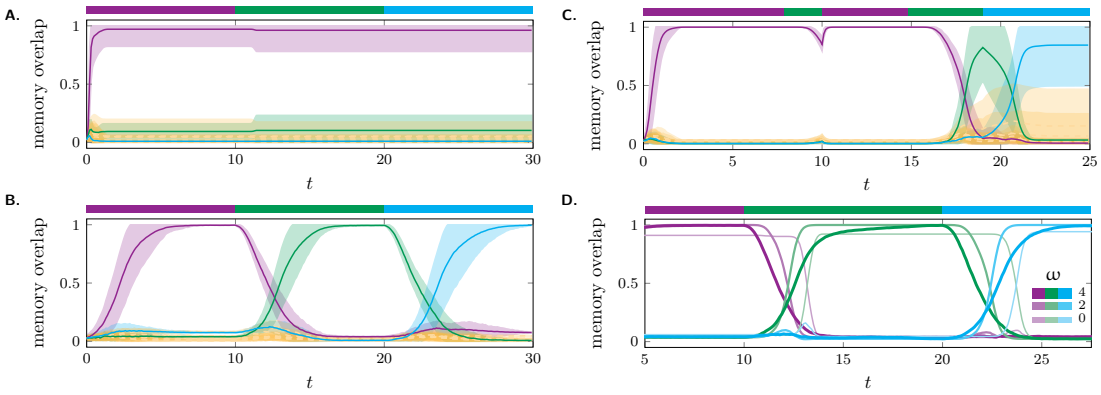


Figure 5.1: **Robustness to noise in the IDP *voltage* model.** (A-B) Differently from Fig. 4.2C, the prototypical memory associated to the red weight remains stable after each input switching. (A) IDP model insulated from noise. The model correctly retrieves the first prototypical memory, but is unable to track the dominant component of next inputs due to persistent stability of the red component. (B) IDP model with high amplitude noise. The model correctly retrieves the first prototypical memory. At input switch, the memory remains stable, but the associated energy well considerably flattens (see Fig. 4.5D as reference case). The noise pushes the activity out of the flat wells and directs it towards the deepest one, which is associated to the dominant saliency weight. The process is then repeated at each input switch. (D) Robustness to brief input perturbations (glitches). The IDP *voltage* model successfully corrects the glitch integrating past internal and present external information. (D) Noise amplitude and speed of the transient towards the next memory. Intermediate values of noise amplitude result in optimal speed for the transition between two memories. Insulated and high amplitude noise models take considerably longer to complete the switch.

5.1 From experimental validation to analytical derivation

We have shown through an extensive set of numerical experiments the beneficial effects of the noise on the performance of the IDP *voltage* model. Precisely, we have seen how the presence of noise favors convergence towards the memory associated with the dominant saliency weight, even for multistable energy landscapes. Moreover, we have observed the error correcting nature of the IDP *voltage* model, where past information is used to counter glitches in the input. It would be desirable to have a mathematical understanding of these properties. Formally the noisy version of the IDP *voltage* model is mathematically described by the Stochastic Differential Equation (SDE)

$$d\mathbf{X}_{\mathrm{H}}(t) = [-\mathbf{X}_{\mathrm{H}}(t) + W(u)\Psi(\mathbf{X}_{\mathrm{H}})]\, dt + \omega\, d\mathbf{B}(t) \tag{5.2}$$

$$= -J_x\Psi(\mathbf{X}_{\mathrm{H}}(t))^{-1}\nabla_x \mathrm{E}_H(\mathbf{X}_{\mathrm{H}}(t), W(u))\, dt + \omega\, d\mathbf{B}(t) \tag{5.3}$$

where $\omega > 0$, $\mathbf{B}(t)$ is the standard n-dimensional Brownian motion and where u is considered fixed in this context. This is an instance of the following more general SDE

$$d\mathbf{X}_{\mathrm{H}}(t) = \mathcal{H}(\mathbf{X}_{\mathrm{H}}(t))\, dt + G(\mathbf{X}_{\mathrm{H}}(t))\, d\mathbf{B}(t) \tag{5.4}$$

where $\mathcal{H} : \mathbb{R}^N \to \mathbb{R}^N$ and $G : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ are functions satisfying some regularity conditions. One way to study the solutions of SDEs is by means of the Fokker-Planck equation. Indeed, if we denote with the symbol $\mu(t, x_{\mathrm{H}})$ the probability distribution of the solution $\mathbf{X}_{\mathrm{H}}(t)$ of the SDE (5.4), then it is well known [109, Ch. 2], [110, Ch. 4] that $\mu(t, x_{\mathrm{H}})$ satisfies the partial differential equation

$$\partial_t\mu(t, x_{\mathrm{H}}) = \nabla_x \cdot \left[-\mathcal{H}(x_{\mathrm{H}})\mu(t, x_{\mathrm{H}}) + \frac{1}{2}\nabla_x \cdot \left(G(x_{\mathrm{H}})G(x_{\mathrm{H}})^\top \mu(t, x_{\mathrm{H}})\right)\right] \tag{5.5}$$

known as Fokker-Planck equation, or Kolmogorov forward equation. We say that this equation admits an asymptotic probability distribution $\mu^\star(x_{\mathrm{H}})$ if $\mu(t, x_{\mathrm{H}}) \to \mu^\star(x_{\mathrm{H}})$ as $t \to \infty$ with respect to a suitable metric. An asymptotic probability distribution $\mu^\star(x_{\mathrm{H}})$ satisfies the partial differential equation

$$\nabla_x \cdot \left[-\mathcal{H}(x_{\mathrm{H}})\mu^\star(x_{\mathrm{H}}) + \frac{1}{2}\nabla_x \cdot \left(G(x_{\mathrm{H}})G(x_{\mathrm{H}})^\top \mu^\star(x_{\mathrm{H}})\right)\right] = 0 \tag{5.6}$$

and, for this reason, is also called a stationary probability distribution of the SDE. The Fokker-Planck equation associated with the IDP *voltage* model is

$$\partial_t\mu(t, x_{\mathrm{H}}) = -\nabla_x \cdot [-x_{\mathrm{H}}\mu(t, x_{\mathrm{H}}) + W(u)\Psi(x_{\mathrm{H}}))\mu(t, x_{\mathrm{H}})] + \frac{\omega^2}{2}\triangle_x\mu(t, x_{\mathrm{H}}) \tag{5.7}$$

By studying the solutions of the this equation one could answer some of the questions posed by our numerical experiments. Specifically:

- The stationary probability distribution would help in understanding where $\mu(t, x_{\mathrm{H}})$ will tend to concentrate as the time tends to infinity. Given the structure of the drift field, it would be reasonable to suppose that the mass concentrates around the minima of $\mathrm{E}_H(x_{\mathrm{H}}; W(u))$. Is this the case?

- The previous result would be useful only if the convergence to the stationary probability distribution occurs fast enough. Indeed it is well known that for some stochastic dynamics the stationary probability distribution could be useless because it might be reached after an extremely long transient and passage through metastable states. Can we have in our context estimates of the convergence rates to the stationary probability distribution as to exclude that metastable states exist?

These questions are very hard to answer in general. Just the question concerning whether the Fokker-Planck equation admits a unique stationary probability distribution $\mu^\star(x_{\mathrm{H}})$ is not straightforward.

There is a case in which some of our questions can be answered. Indeed, it is well known that if there exists a potential function $V(x_{\mathrm{H}}) \in C^1(\mathbb{R}^N)$ such that $e^{-V(x_{\mathrm{H}})} \in L^1(\mathbb{R}^N)$ and if the SDE (5.4) satisfies

$$\mathcal{H}(x_{\mathrm{H}}) = -\nabla_x V(x_{\mathrm{H}}), \qquad G(x_{\mathrm{H}}) = \omega\, I_N,$$

where $\omega > 0$, then [111], [112] the Fokker-Planck equation (5.5) admits a unique stationary probability distribution given by

$$\mu^\star(x_{\mathrm{H}}) = \frac{1}{\mathcal{Z}} e^{-\frac{2V(x_{\mathrm{H}})}{\omega^2}}, \qquad \mathcal{Z} = \int_{\mathbb{R}^N} e^{-\frac{2V(x_{\mathrm{H}})}{\omega^2}}\, dx_{\mathrm{H}}. \tag{5.8}$$

This probability distribution is also known as the Gibbs measure. This result can be shown to hold true in the more general case in which in eq. (5.4) we have that

$$\mathcal{H}(x_{\mathrm{H}}) = -D(x_{\mathrm{H}})^{-1}\nabla_x V(x_{\mathrm{H}}) + \omega^2 \nabla_x \cdot D(x_{\mathrm{H}}(t))^{-1}, \qquad G(x_{\mathrm{H}}) = \omega D(x_{\mathrm{H}}(t))^{-\frac{1}{2}},$$

where $D(x_{\mathrm{H}})$ is a symmetric positive definite matrix for all x_{H}.

The noisy IDP *voltage* model satisfies $\mathcal{H}(x_{\mathrm{H}}) = -D\Psi(x_{\mathrm{H}})^{-1}\nabla_x E(x_{\mathrm{H}}, W(u))$ and hence, in general, it does not fit any of the previous special cases. Hence the previous results cannot be used in our framework. This is true, except for the scalar case in which $N = 1$. This case can provide some useful intuition on the problem we aim to understand. Indeed, in this case we can

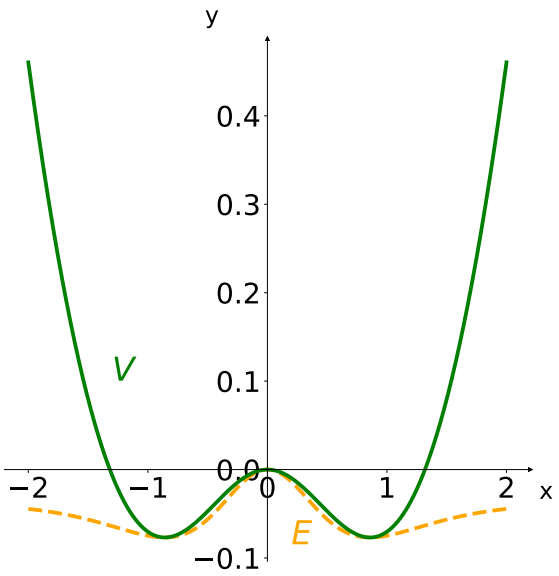


Figure 5.2: **Graphic visualization of the difference between the two energy functions** $\mathrm{E}(x_{\mathbf{H}})$ **and** $V(x_{\mathbf{H}})$**.** For the noisy IDP *voltage* model in the scalar case when $\psi(x_{\mathrm{H}}) = \tanh(x_{\mathrm{H}})$.

define the scalar synaptic parameter $w \in \mathbb{R}$ and introduce two energy functions (5.2)

$$\mathrm{E}_H(x_{\mathrm{H}}) = -\frac{1}{2} w \psi(x_{\mathrm{H}})^2 + x_{\mathrm{H}} \psi(x_{\mathrm{H}}) - \int_0^{x_{\mathrm{H}}} \psi(s)\, ds, \tag{5.9}$$

$$V(x_{\mathrm{H}}) = \frac{x_{\mathrm{H}}^2}{2} - w \int_0^{x_{\mathrm{H}}} \psi(s)\, ds \tag{5.10}$$

Observe that in this case the IDP vector field admit two representations

$$\mathcal{H}(x_{\mathrm{H}}) = \partial_x V(x_{\mathrm{H}}) = \psi'(x_{\mathrm{H}})^{-1} \partial_{x_{\mathrm{H}}} \mathrm{E}(x_{\mathrm{H}}) \tag{5.11}$$

and hence in scalar case the noisy IDP *voltage* model fits the previous special cases since it takes the form

$$d\mathbf{X}_{\mathrm{H}}(t) = -\partial_x V(\mathbf{X}_{\mathrm{H}}(t))\, dt + \omega\, dB(t) \tag{5.12}$$

Notice that here the stationary probability distribution takes the form given in eq. (5.8), but also that, since the energy $V(x_{\mathrm{H}})$ is very different from $\mathrm{E}_H(x_{\mathrm{H}})$, we can expect that also $e^{-2V(x_{\mathrm{H}})/\omega^2}$ will be very different from $e^{-2\mathrm{E}_H(x_{\mathrm{H}})/\omega^2}$. Therefore, the energy $\mathrm{E}_H(x_{\mathrm{H}})$ might not be useful in giving information about the stationary probability distribution. Indeed, if for example we choose the activation function $\psi(z) = \tanh(z)$, then the function $e^{-\mathrm{E}_H(z)}$ does not even belong to $L^1(\mathbb{R})$, since it can be verified that $\lim_{z\to+\infty} \mathrm{E}_H(z) < +\infty$ in this case. Thus, the only admissible candidate as stationary probability distribution for the associated SDE is $\mu^\star(z) = e^{-2V(z)/\omega^2}/\mathcal{Z}$,

since it is easy to check that $e^{-V(z)} \in L^1(\mathbb{R})$.

From the literature [113], we know that under some assumptions on the intensity of noise $\omega > 0$ the measure associated to the stochastic IDP *voltage* model converges to a unique stationary distribution even when $N > 1$. However, the literature does not provide tools to address the problem of where the stationary measure of the IDP Hopfield model concentrates. Specifically, we would still be far from proving the results we need to obtain a mathematical understanding of the behavior displayed in the numerical experiments 5.1, namely that the stationary probability distribution $\mu_\infty(x_{\mathrm{H}})$ concentrates most of its mass around the deepest minima of the energy function $\mathrm{E}_H(x_{\mathrm{H}}; W)$.

Notation We identify with $B_r(x) \subset \mathbb{R}^d$ the ball of radius $r > 0$ and centered at $x \in \mathbb{R}^d$. For a smooth manifold $\mathcal{M} \subseteq \mathbb{R}^d$ we identify with $T_x\mathcal{M}$ the tangent space of $\mathcal{M}$ at $x \in \mathcal{M}$, and with $\partial\mathcal{M}$ its boundary. We identify with $C^k(\mathcal{X}; \mathcal{Y})$ the class of k-differentiable functions from $\mathcal{X}$ into $\mathcal{Y}$. Let $f \in C^1(\mathbb{R}^d; \mathbb{R})$ and denote with $\nabla f(x) \in \mathbb{R}^d$ its gradient. We identify the partial derivative of f with respect to x_i as $\partial_{x_i} f(x)$. Let $g \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ and denote with $J_x g(x) \in \mathbb{R}^{d\times d}$ its Jacobian, and with $\nabla \cdot g(x) \in \mathbb{R}$ its divergence. We identify the identity matrix of dimension $d \in \mathbb{N}$ as $\mathcal{I}_d \in \mathbb{R}^{d\times d}$. We refer to a positive definite matrix $A \in \mathbb{R}^{d\times d}$ as $A \succ 0$. Let $B \in \mathbb{R}^{d\times d}$ and we denote its trace operator as $\mathrm{Tr}\,(B)$. We denote with $(\ ,\ )_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the standard Euclidean inner product. We denote the weighted inner product as $(\ ,\ )_A : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, for $A \succ 0$. We identify with $\mathcal{P}_2(\mathcal{X})$ the space of probability measures over $\mathcal{X}$ having finite second moment. Stochastic processes are identified by boldface upper roman letters, i.e. $\mathbf{X}_t$.

Theoretical Background Consider the Ito stochastic differential equation (S.D.E.)

$$\begin{cases} d\mathbf{X}_t = f(t, \mathbf{X}_t)\, dt + G(t, \mathbf{X}_t)\, d\mathbf{B}_t \\ \mathbf{X}(0) = x_0 \in \mathbb{R}^N \end{cases} \tag{5.13}$$

where $\mathbf{B}_t$ is the standard d-dimensional Brownian motion.

Assumption 5.1.1 (Lipschitzianity and sublinearity). Let $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ be the drift term, globally Lipschitz with constant $L_f > 0$ and sublinear with constant $s_f > 0$.

$$|(f(t,x) - f(t,y), x - y)_2| \le L_f \|x - y\|^2 \qquad \forall x, y \in \mathbb{R}^N \tag{5.14}$$

$$\|f(t,x)\|^2 \le s_f(1 + \|x\|^2) \tag{5.15}$$

The diffusion term $G : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^{N\times N}$ obeys analogous constraints. Namely

$$\|(G(t,x) - G(t,y)\|_2 \leq L_G\|x-y\|^2 \qquad \forall x, y \in \mathbb{R}^N \tag{5.16}$$

$$\|G(t,x)\|^2 \leq s_G(1+\|x\|^2) \tag{5.17}$$

Under these standard assumptions [114] we can guarantee the existence and uniqueness of a strong solution to eq. (5.13).

Definition 5.1.2 (Infinitesimal generator). Let $h : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ be a Lebesgue-measurable function. The infinitesimal generator of h along the process (5.13) is defined as

$$\begin{aligned}\mathcal{A}h(t,x_t) =& \lim_{\delta t\to 0} \frac{\mathbb{E}_{\mathbf{B_t}}[h(t+\delta t, \mathbf{X}_{t+\delta t})] - h(t,x_t)}{\delta t} \\ =&\partial_t h(t,x_t) + \nabla h(t,x_t)^\top f(x_t) + \frac{1}{2}\mathrm{Tr}\left(G(t,x_t)J_x(\nabla h(t,x_t))G(t,x_t)^\top\right) .\end{aligned} \tag{5.18}$$

The probability measure $\mu \in \mathcal{P}_2(\mathbb{R}_{\geq 0} \times \mathbb{R}^N)$ is related [115] to the S.D.E. (5.13) via the Kolmogorov forward equation (or Fokker-Planck equation)

$$\partial_t \mu(t,x) = \nabla \cdot \{-f(t,x)\mu(t,x) + \frac{1}{2}\nabla \cdot [D(t,x)\mu(t,x)]\} \tag{5.19}$$

$$\mu(0,x) = \mu_0(x) \qquad \forall x \in \mathbb{R}^N \tag{5.20}$$

where $D(t,x) = G(t,x)G(t,x)^\top$ and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^N)$ is the initial distribution of the states, i.e. the distribution of the initial conditions for eq. (5.13). Additionally, we introduce the notion of distance in $\mathcal{P}_2(\mathbb{R}^N)$ between two probability measures [116].

Definition 5.1.3 (Wasserstein metric). Let $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^N)$ be two probability measures having finite second moment. Define $\Gamma(\mu_1,\mu_2)$ the set of the joint probability measures γ having μ_1 and μ_2 as marginals. The p-Wasserstein metric, for $p \geq 1$, is defined as

$$W_p(\mu_1,\mu_2) = \left(\inf_{\gamma\in\Gamma(\mu_1,\mu_2)} \mathbb{E}_\gamma[\|x-y\|_p^p]\right)^{\frac{1}{p}} \tag{5.21}$$

$$= \left(\inf_{\gamma\in\Gamma(\mu_1,\mu_2)} \int_{real^N\times\mathbb{R}^N} \|x-y\|_p^p \, d\gamma(x,y)\right)^{\frac{1}{p}} \tag{5.22}$$

For the sake of brevity, from now on we will abbreviate $\Gamma(\mu_1,\mu_2)$ as simply Γ, and $\mu(t,x) = \mu_t(x)$.

5.2 Contracting drifts and convergence to stationary measures

We begin by introducing the definition of global contractivity, and henceforth consider only autonomous (time-independent) drift terms.

Definition 5.2.1 (c-strong contractivity). We say that $f : \mathbb{R}^N \to \mathbb{R}^N$ is c-strongly infinitesimally contracting w.r.t the 2-norm if it holds

$$(f(x) - f(y), x - y)_2 \leq -c\|x - y\|_2^2 \qquad \forall x, y \in \mathbb{R}^N \tag{5.23}$$

where $0 < c \leq L_f$ is the contraction rate, for L_f Lipschitz constant (5.14).

We now provide sufficient conditions for the convergence to stationarity of a probability measure (5.19) for a c-strongly contracting drift term and spatially inhomogeneous diffusion term.

Theorem 5.2.2 (Contraction of measures). *Let $\{\mathbf{X}_t\}_{t\geq 0}$ be the unique strong solution to eq. (5.13) with drift f and diffusion G satisfying Assumption 5.1.1. Let f be c-strongly contracting w.r.t. the 2-norm as for Definition 5.2.1. If $c > L_G/2$, with L_G as in (5.16), then the solution $\mu_t \in \mathcal{P}_2(\mathbb{R}^N)$ to eq. (5.19) associated to the process $\{\mathbf{X}_t\}_{t\geq 0}$ converges in the 2-Wasserstein metric to a unique stationary probability measure $\mu^\star \in \mathcal{P}_2(\mathbb{R}^N)$.*

Proof. Using a parallel coupling approach [117], we consider two stochastic processes $\{\mathbf{X}_t(x_0, \omega)\}_{t\geq 0}$ and $\{\mathbf{Z}_t(z_0, \omega)\}_{t\geq 0}$ both unique strong solutions to eq. (5.13) for different initial conditions $x_0, z_0 \in \mathbb{R}^N$, $x_0 \neq z_0$ and the same realization of the noise $\omega \in \mathbb{R}^N$. Taking the standard 2-norm $h(x) = \|x\|_2^2$ and evaluating the infinitesimal generator along the difference process $\{\mathbf{X}_t - \mathbf{Z}_t\}_{t\geq 0}$, we obtain

$$\begin{aligned}\mathcal{A}h(x_t - z_t) &= 2(f(x_t) - f(z_t), x_t - z_t)_2 \\ &\quad + \mathrm{Tr}\left((G(t, x_t) - G(t, z_t))(G(t, x_t) - G(t, z_t))^\top\right).\end{aligned} \tag{5.24}$$

Exploiting now the c-strong contractivity of the field f (5.23) and the Lipschitzianity of G (5.16)

$$\mathcal{A}h(x_t - z_t) \leq -\underbrace{(2c - L_G)}_{>0} h(x_t - z_t). \tag{5.25}$$

Using Dynkin's formula [118] on the Ito differential $dh(\mathbf{X}_t - \mathbf{Z}_t)$

$$\mathbb{E}_{\mathbf{B}_t}[h(\mathbf{X}_t - \mathbf{Z}_t)] - h(x_0 - z_0) \leq -\int_0^t (2c - L_G) h(\mathbf{X}_s - \mathbf{Z}_s)\, ds. \tag{5.26}$$

Applying a specialized version of Gronwall lemma [119] we get

$$\mathbb{E}_{\mathbf{B}_t}[h(\mathbf{X}_t - \mathbf{Z}_t)] \leq h(x_0 - z_0)e^{-(2c-L_G)t} \tag{5.27}$$

Let $\mu_t \in \mathcal{P}_2(\mathbb{R}^N)$ be the probability measure associated to the process $\{\mathbf{X}_t\}_{t\geq 0}$ and $\nu_t \in \mathcal{P}_2(\mathbb{R}^N)$ be the probability measure associated to the process $\{\mathbf{Z}_t\}_{t\geq 0}$. By the positivity of the norm

$$\begin{aligned}\mathbb{E}_{\mathbf{B}_t}[h(\mathbf{X}_t - \mathbf{Z}_t)] &\geq h(\mathbb{E}_{\mathbf{B}_t}[\mathbf{X}_t - \mathbf{Z}_t]) \\ &= h(x_t(x_0) - z_t(z_0))\end{aligned} \tag{5.28}$$

where in the last line the processes are independent from the noise realization $\omega \in \mathbb{R}^N$. Let now Γ_t denote the set of joint probability measures γ_t having μ_t and ν_t as marginals. Taking expectations of eq. (5.27) w.r.t. the joint measure, and exploiting eq. (5.28) we get

$$\begin{aligned}\mathbb{E}_{\gamma_t}[h(x_t - z_t)] &= \mathbb{E}_{\gamma_0}[h(x_t(x_0) - z_t(z_0))] \\ &\leq \mathbb{E}_{\gamma_0}[h(x_0 - z_0)]e^{-(2c-L_G)t}\end{aligned} \tag{5.29}$$

and miniminzing w.r.t. γ_0

$$\inf_{\gamma_o \in \Gamma_0} \mathbb{E}_{\gamma_t}[h(x_t - z_t)] \leq W_2^2(\mu_0, \nu_0)e^{-(2c-L_G)t}. \tag{5.30}$$

By the optimality of the infimum w.r.t. the joint measure $\gamma_t \in \Gamma_t$ we get

$$W_2^2(\mu_t, \nu_t) \leq W_2^2(\mu_0, \nu_0)e^{-(2c-L_G)t} \tag{5.31}$$

□

Observe how the presence of spatially inhomogeneous diffusion results in a "discount" of the decay rate to stationarity for the associated measure, which would otherwise equal the contraction rate in the case of homogeneous diffusion. Despite the "discount", the transient towards stationarity retains its exponential character. Define now two processes driven by the same drift term f but different diffusion terms satisfying the conditions (5.16).

$$d\mathbf{X}_t = f(t, \mathbf{X}_t)\, dt + G(t, \mathbf{X}_t)\, d\mathbf{B}_t \tag{5.32}$$

$$d\mathbf{Z}_t = f(t, \mathbf{Z}_t)\, dt + Q(t, \mathbf{Z}_t)\, d\mathbf{B}_t \tag{5.33}$$

$$\mathbf{X}(0) = x_0 \in \mathbb{R}^N, \qquad \mathbf{Z}(0) = z_0 \in \mathbb{R}^N \tag{5.34}$$

We further suppose that both G and Q have finite Frobenius norm over the entire state-space

uniformly in time.

Assumption 5.2.3 (**Noise bound**). The diffusion terms G and Q are uniformly upper bounded in the 2-norm (Frobenius norm).

$$\sup_{t,x} \|G(t,x)\|_F < +\infty \tag{5.35}$$

$$\sup_{t,x} \|Q(t,x)\|_F < +\infty \tag{5.36}$$

Then we have the following theorem that bounds the distance in the 2-Wasserstein metric between the probability measures associated to processes with equal drift term and different diffusion term (both processes satisfy the assumptions of Theorem 5.2.2).

Proposition 5.2.4 (**Contraction for different diffusions**). *Let $\{\mathbf{X}_t\}_{t\geq 0}$ be the unique strong solution to eq. (5.32) with diffusion G and associated measure μ_t. Let $\{\mathbf{Z}_t\}_{t\geq 0}$ be the unique strong solution to eq. (5.33) with diffusion term Q and associated stationary measure ν_t. Assume that L_G and L_Q defined in (5.16) satisfy Assumption 5.2.3 and take $L = \min\{L_G, L_Q\}$. If $c > L/2$ then*

$$W_2^2(\mu^\star, \nu^\star) \leq \chi^2 := \sup_{t,x} \|G(t,x) - Q(t,x)\|_F^2 \tag{5.37}$$

where $\mu^\star$ and $\nu^\star$ are the unique stationary measures to which μ_t and ν_t converge.

Proof. Assume without loss of generality that $L = L_G$ and consider the difference process $\{\mathbf{X}_t - \mathbf{Z}_t\}_{t\geq 0}$. Evaluating the infinitesimal generator of h along the difference process we obtain

$$\begin{aligned}
\mathcal{A}h(x_t - z_t) =& 2(f(x_t) - f(z_t), x_t - z_t)_2 \\
&+ \|G(t,x_t) - Q(t,z_t)\|_F^2 \\
\leq& -2c\, h(x_t - z_t) + \|G(t,x_t) - G(t,z_t)\|_F^2 \\
&+ \|G(t,z_t) - Q(t,z_t)\|_F^2 \\
\leq& -(2c - L)h(x_t - z_t) + \chi^2
\end{aligned} \tag{5.38}$$

where in the second passage we have used the strong c-contractivity of f and the triangular inequality, and in the third passage the Lipschitzianity of G and Assumption 5.2.3. Similarly to Theorem 5.2.2, we can apply Dynkin's formula and a specialized version of Gronwall lemma [119]

to reach

$$\begin{aligned}\mathbb{E}_{\mathbf{B}_t}[h(\mathbf{X}_t - \mathbf{Z}_t)] &\leq \left(h(x_0 - z_0) - \frac{\chi^2}{c}\right)^+ e^{-(2c-L)t} + \chi^2 \\ &\leq h(x_0 - z_0)e^{-(2c-L)t} + \chi^2. \end{aligned} \tag{5.39}$$

Taking expectations w.r.t. the joint measures $\gamma_t \in \Gamma_t$ and $\gamma_0 \in \Gamma_0$ and minimizing in the respective sets, we obtain that

$$W_2^2(\mu_t, \nu_t) \leq W_2^2(\mu_0, \nu_0)e^{-(2c-L)t} + \chi^2 \quad \forall t \geq 0. \tag{5.40}$$

Since W_2^2 is a metric for the space of probability measures, by the triangular inequality we have

$$\begin{aligned}W_2^2(\mu^\star, \nu^\star) \leq & W_2^2(\mu^\star, \mu_t) + W_2^2(\nu_t, \nu^\star) \\ & + W_2^2(\mu_t, \nu_t). \end{aligned} \tag{5.41}$$

Since it holds $\forall t \geq 0$, passing to the limit and exploiting the convergence of μ_t and ν_t from Theorem 5.2.2 we have

$$W_2^2(\mu^\star, \nu^\star) \leq \lim_{t \to +\infty} W_2^2(\mu_t, \nu_t). \tag{5.42}$$

Finally, using inequality (5.42) taking the limit in (5.40) we get

$$W_2^2(\mu^\star, \nu^\star) \leq \lim_{t \to +\infty} W_2^2(\mu_0, \nu_0)e^{-(2c-L)t} + \chi^2 \tag{5.43}$$

□

5.3 B_r-contracting drifts and concentration of stationary measures

We are now interested in the analysis of a more general scenario where the drift term has possibly many equilibria (see for example Fig. 5.3(b)) and is therefore not globally contracting on $\mathbb{R}^N$.

Definition 5.3.1 (**c-strong B_r-contractivity**). Let f satisfy Assumption (5.14) and define the constants $c > 0$ and a radius $r > 0$. We say that f is c-strongly B_r-contracting on $\mathbb{R}^N$ if

$$(f(x) - f(y), x - y)_2 \leq -c\|x - y\|^2 \qquad \forall x, y \in B_r(0)^c \tag{5.44}$$

Remark 5.3.2 (**Global contractivity**). Notice that the definition of c-strong global contractivity 5.2.1 and the definition of c-strong B_r-contractivity 5.3.1 coincide in the limit $r \to 0$.

The case of B_r-contractivity is particularly interesting since it encompasses more complex

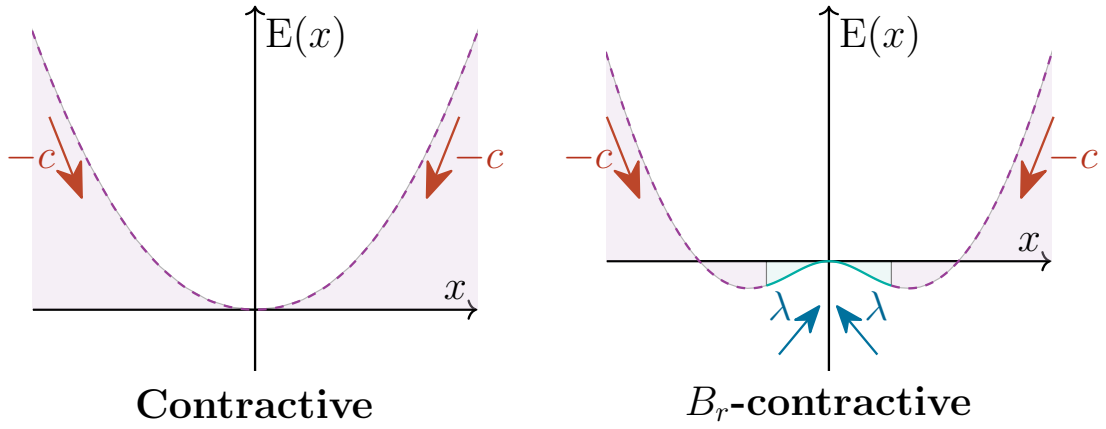


Figure 5.3: **Globally and B_r-contracting vector fields.**Visual example of potentials $\mathrm{E}(x)$ associated to globally contracting and B_r-contracting vector fields f. (a) Globally contracting vector field associated to a convex potential. (b) B_r-contractive vector field associated to a mostly convex potential, with a small concave region.

dynamics endowed with multiple equilibrium points. In particular, it allows to relate the concentration of measure around the stable equilibrium points to the properties of their basin of attraction. We will focus on stochastic dynamics characterized by constant noise $G(t,x) \equiv \omega \mathcal{I}_N$, for which it has been proved [113] that the associated measure converges to stationarity.

Assumption 5.3.3 (Stable equilibria and contraction rates on the shell). Let $x^\star$ be a stable equilibrium of $\dot{x} = f(x)$ for f drift term. There exists $c(x^\star, r) > c_{\min}(x^\star) > 0$ for $r \in (0, r_{\max}(x^\star)]$ such that

$$(f(x) - f(y), x - y)_2 \leq -c(x^\star, r)\|x - y\|_2^2 \qquad \forall x, y \in \partial B_r(x^\star) \tag{5.45}$$

Remark 5.3.4 (**From shell contraction rates to ball contraction rates**). The previous assumption requires that some of the stable equilibrium points $x^\star \in \mathcal{D}$ have associated a maximal radius $r_{\max}(x^\star)$ such that the drift term is $c(x^\star, r)$-strongly contracting for every point in the shell $\partial B_r(x^\star)$ for $r \in (0, r_{\max}(x^\star)]$. If the shell contractivity condition holds for all $r \in (0, r_{\max}(x^\star)]$, then the drift term is contracting with rate $c_{\min}(x^\star)$ for every point in the ball $B_{r_{\max}(x^\star)}(x^\star)$ (see Fig. 5.4 for a visual representation of local contractivity).

When the drift term is globally contracting and there is only a unique equilibrium point, we have that $r_{\max}(x^\star) \to +\infty$. Instead, when there are multiple equilibrium points and we only have B_r-contractivity, the radii $\{r_{\max}(x^\star) : x^\star \in \mathcal{D}\}$ are limited by the saddle regions.

Framing Assumption 5.3.3 in the context of B_r-contractivity, we have that some stable equilibrium $x^\star$ of the B_r-contracting drift term have a basin of attraction containing a ball of radius $r_{\max}(x^\star)$ where the dynamics are locally contractive. For notational simplicity, we will

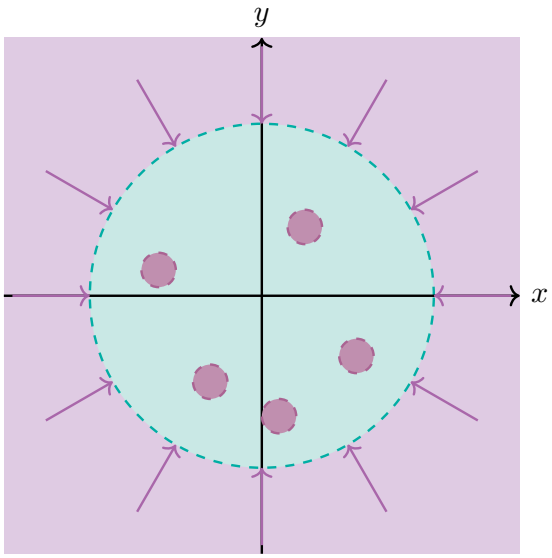


Figure 5.4: **Illustration of the contractive properties of the system in the Euclidean norm.** The pink region outside the ball and the purple subregions within it are contractive with respect to the 2-norm, as indicated by the inward-pointing arrows. In contrast, the turquoise region inside the ball does not satisfy the same contractivity property.

henceforth abbreviate integrands as $f(x)g(x) = (fg)_{|x}$ when necessary. We now prove how the radius r and the contraction rate $c(x^\star, r)$ determine if the ball $B_r(x^\star)$ attracts or dissipates mass as the measure evolves in time.

Proposition 5.3.5 (Mass sinks). *Let $\mu_t \in \mathcal{P}_2(\mathbb{R}^N)$ be the measure associated to the process (5.13) via the FPE (5.19). Let $x^\star$ satisfy Assumption 5.3.3. If there exist $r \in (0, r_{\max}(x^\star)]$ such that $c(x^\star, r) \geq \frac{N}{2}\left(\frac{\omega}{r}\right)^2$ then*

(i) For all times $t \geq 0$ it holds

$$\int_{B_r(x^\star)} \mu_t(x)\, dx \geq \int_{B_r(x^\star)} \mu_0(x)\, dx. \tag{5.46}$$

(ii) If there exists $\mu^\star \in \mathcal{P}_2(\mathbb{R}^N)$ stationary measure such that $\mu_t \xrightarrow{W_p} \mu^\star$, then

$$\int_{B_r(x^\star)} \mu^\star(x)\, dx \geq \int_{B_r(x^\star)} \mu_0(x)\, dx. \tag{5.47}$$

Proof. (i) The measure μ_t is associated to a simple S.D.E. with drift term contracting at infinity and scalar Brownian term (non-degenerate diffusion), and therefore its density is smooth

and with exponential Aronson-type bounds [120]. Therefore, it holds that $\partial_t \mu_t \in L^1(\mathbb{R}^N)$ for all $t \geq 0$ and since we are working in a compact space $B_r(x^\star)$ we can easily apply the dominated convergence theorem [121].

$$\begin{aligned}
&\partial_t \int_{B_r(x^\star)} \mu_t(x)\, dx \\
&= \int_{B_r(x^\star)} \partial_t \mu_t(x)\, dx \\
&= \int_{B_r(x^\star)} \nabla \cdot \left[-f\mu_t + \frac{\omega^2}{2} \nabla \mu_t \right]_{|x} dx \\
&= \int_{\partial B_r(x^\star)} \left[-\mu_t (f, \xi)_2 + \frac{\omega^2}{2} (\nabla \mu_t, \xi)_2 \right]_{|x} d_x\sigma \\
&= \int_{\partial B_r(x^\star)} \mu_t \left[-(f, \xi)_2 - \frac{\omega^2}{2} \mathrm{Tr}\,(J_x \xi) \right]_{|x} d_x\sigma
\end{aligned} \tag{5.48}$$

where $d_x\sigma$ is the surface measure, $\xi(x) \in [T_x \partial B_r(x^\star)]^\perp$ is the surface outer normal, and in the last passage we used Lemma 5.5.1. The outer normal to the shell $\partial B_r(x^\star)$ is $\xi(x) = (x - x^\star)/r$, so we have

$$\begin{aligned}
&\partial_t \int_{B_r(x^\star)} \mu_t(x)\, dx \\
=&\frac{1}{r} \int_{\partial B_r(x^\star)} \mu_{t|x} \left[-(f(x), x - x^\star)_2 - \frac{N}{2}\omega^2 \right] d_x\sigma \\
\geq&\frac{1}{r} \int_{\partial B_r(x^\star)} \mu_{t|x} \left[c(x^\star, r) \|x - x^\star\|_2^2 - \frac{N}{2}\omega^2 \right] d_x\sigma \\
=&\frac{1}{r} \mu_t(\partial B_r(x^\star)) \left[c(x^\star, r) r^2 - \frac{N}{2}\omega^2 \right] \geq 0
\end{aligned} \tag{5.49}$$

where from the second to the third passage we have used the fact that $f(x^\star) = 0$ and the contractivity of f in $B_r(x^\star)$. Finally, from the third to the last passage we have used the fact that $\|x - x^\star\|_2^2 \equiv r^2$ for all $x \in \partial B_r(x^\star)$ and denoted $\mu_t(\partial B_r(x^\star)) > 0$ the probability measure over the surface of the ball. Thus, we have that for all $t \geq 0$

$$\int_{B_r(x^\star)} \mu_t(x)\, dx \geq \int_{B_r(x^\star)} \mu_0(x)\, dx. \tag{5.50}$$

(ii) We know that $\mu_t \xrightarrow{W_p} \mu^\star$ and μ_t. Using the same arguments at the beginning of (i), the associated density is uniformly upper bounded in the $\|\cdot\|_{L^{1+\epsilon}}$-norm for $\epsilon > 0$ and from Vitali convergence theorem [121] we get $\mu_t \xrightarrow{L^1} \mu^\star$. This justify the exchange of integral

and limit operator, from which

$$\int_{B_r(x^\star)} \mu^\star(x)\,dx = \int_{B_r(x^\star)} \lim_{t\to+\infty} \mu_t(x)\,dx$$
$$= \lim_{t\to+\infty} \int_{B_r(x^\star)} \mu_t(x)\,dx \geq \int_{B_r(x^\star)} \mu_0(x)\,dx. \tag{5.51}$$

□

The previous result showcases an interesting interplay between the radius r and the local contraction rate $c(x^\star, r)$, where the product of the two must be sufficiently large to counterbalance the noise amplitude $\omega > 0$ weighted by the state-space dimension N. Specifically, for equilibria with small basin of attraction, the contraction rate must be significantly large for mass to concentrate within the ball. Conversely, equilibria in regions characterized by weak contraction rates can still attract mass provided that the radius r is large enough. We now focus on the simplest generalization of the pure gradient of a potential $\mathrm{E} \in C^2(\mathbb{R}^N;\mathbb{R})$ such that the B_r-contracting drift field can be expressed as

$$f(x) = -P(x)\nabla \mathrm{E}(x)$$
$$= \mathrm{diag}(p_1(x_1),\dots,p_N(x_N))\nabla \mathrm{E}(x) \qquad p_i(x_i) > 0,\ i = 1,\dots,N \tag{5.52}$$

Specifically, we would like to understand whether we can draw useful information on the stationary measure $\mu^\star \in \mathcal{P}_2(\mathbb{R}^N)$ given knowledge on $\mathrm{E}(x)$.

Theorem 5.3.6 (Concentration of mass). *Let $\{\mathbf{X}_t\}_{t\geq 0}$ be the solution of eq. (5.13) with $f(x) = -P(x)\nabla \mathrm{E}(x)$ as in (5.52) and c-strongly B_r-contractive and $G(t,x) \equiv \omega \mathcal{I}_N$. Let $P \in C^1(\mathbb{R}^N;\mathbb{R}^{N\times N})$ with $\mathrm{sign}(\partial_{x_i} p_i(x_i)) = \mathrm{sign}(x_i)$. Let $x_a, x_b \in \mathbb{R}^d$ be minima of the potential $\mathrm{E}(x)$ and choose $r > 0$ such that*

(I) $\partial B_r(x_a)$ ($\partial B_r(x_b)$) is in the same orthant of x_a (x_b).

(II) $\mathrm{E}(z + x_a) \leq \mathrm{E}(x_b) \leq \mathrm{E}(z + x_b) < 0$ for all $z \in B_r$.

If there exists $\mu^\star \in \mathcal{P}_2(\mathbb{R}^N)$ such that $\mu_t \xrightarrow{W_p} \mu^\star$ associated to the process $\{\mathbf{X}_t\}_{t\geq 0}$, then it satisfies the integral inequality

$$\int_{B_r(x_a)} \mu^\star(x)\,dx \geq \int_{B_r(x_b)} \mu^\star(x)\,dx. \tag{5.53}$$

Proof. The existence of a stationary measure $\mu^\star \in \mathcal{P}_2(\mathbb{R}^N)$ was proved in [113], and $\mu^\star$ solves

for all $y \in \mathbb{R}^N$ the integral equation

$$0 = \int_{B_r(y)} \nabla \cdot \left[P(x)\nabla \mathrm{E}(x)\mu^\star(x) + \frac{\omega^2}{2}\nabla\mu^\star(x) \right] dx \tag{5.54}$$

$$= \int_{\partial B_r(y)} \mu^\star(x) \left[(\nabla \mathrm{E}(x), \xi(x))_P - \frac{\omega^2}{2}\nabla\cdot\xi(x) \right] d_x\sigma \tag{5.55}$$

where the unit outer normal is $\xi(x) = (x-y)/r$, so that $\nabla \cdot \xi(x) = N/r$. There always exists $\epsilon > 0$ such that $\omega^2 N(2r)^{-1} \geq \epsilon$, and therefore

$$\int_{\partial B_r(y)} \mu^\star(x)(\nabla \mathrm{E}(x), \xi(x))_P \, d_x\sigma \geq \epsilon \int_{\partial B_r(y)} \mu^\star(x) \, d_x\sigma \tag{5.56}$$

hence the left surface integral is always greater or equal to zero. We now evaluate the difference of the same integral quantities centered on the two minima x_a, x_b of the potential $\mathrm{E}(x)$.

$$\begin{aligned} 0 = &\int_{\partial B_r(x_a)} \mu^\star(x) \left[(\nabla \mathrm{E}(x), \xi(x))_P - \frac{\omega^2 N}{2r} \right] d_x\sigma \\ &- \int_{\partial B_r(x_b)} \mu^\star(x) \left[(\nabla \mathrm{E}(x), \xi(x))_P - \frac{\omega^2 N}{2r} \right] d_x\sigma \end{aligned} \tag{5.57}$$

Letting now $c_\omega^{r,N} = 2r(\omega^2 N)^{-1}$, rearranging the integrals and operating the change of variables $z + x_a = x$ (and same for x_b), we obtain

$$\begin{aligned} &\int_{\partial B_r} \mu^\star(z+x_a) - \mu^\star(z+x_b) \, d_z\sigma \\ &= c_\omega^{r,N} \int_{\partial B_r} [\mu^\star(\nabla \mathrm{E}, \xi)_P]_{|z+x_a} - [\mu^\star(\nabla \mathrm{E}, \xi)_P]_{|z+x_b} \, d_z\sigma \end{aligned} \tag{5.58}$$

Focusing on the integral on the right-hand side and using integration by parts we get

$$\begin{aligned} &- \int_{\partial B_r} [\mathrm{E}\nabla \cdot (P\xi\mu^\star)]_{|z+x_a} \, d_z\sigma \\ &+ \int_{\partial B_r} [\mathrm{E}\nabla \cdot (P\xi\mu^\star)]_{|z+x_b} \, d_z\sigma \end{aligned} \tag{5.59}$$

We now focus on the divergence term inside each integrand. Specifically, we have that

$$\begin{aligned}&\nabla \cdot (P\xi\mu^\star)_{|x} = \\ &= \left[\sum_{i=1}^{d} \partial_{x_i}(P_{ii}\xi_i)\right]_{|x} \mu^\star_{|x} + [(P\xi, \nabla\mu^\star)_2]_{|x}\end{aligned} \tag{5.60}$$

Observe now that for $h_P(x) = x^\top Px$ we have, using integration by parts on the shell of a generic ball $B_r(x)$

$$\begin{aligned}&\int_{\partial B_r(x)} [(P\xi, \frac{\omega^2}{2}\nabla\mu^\star + P\nabla \mathrm{E}\mu^\star)_2]_{|x}\, d_x\sigma \\ &= r\int_{\partial B_r(x)} [(\nabla h_P(\xi), \frac{\omega^2}{2}\nabla\mu^\star + P\nabla \mathrm{E}\mu^\star)_2]_{|x}\, d_x\sigma \\ &= -r\int_{\partial B_r(x)} h_P(\xi)_{|x} \underbrace{\nabla \cdot \left[\frac{\omega^2}{2}\nabla\mu^\star + P\nabla \mathrm{E}\mu^\star\right]_{|x}}_{\text{(F.P.E.)}\equiv 0} d_x\sigma = 0\end{aligned} \tag{5.61}$$

and consequently it holds that

$$\int_{\partial B_r(x)} [(P\xi, \nabla\mu^\star)_2]_{|x}\, d_x\sigma = -\int_{\partial B_r(x)} \frac{2}{\omega^2}(P\xi, P\nabla \mathrm{E}\mu^\star)_2]_{|x}\, d_x\sigma \tag{5.62}$$

Thus, the divergence term becomes

$$\begin{aligned}&\nabla \cdot (P\xi\mu^\star)_{|x} = \\ &= \mu^\star_{|x}\left[\sum_{i=1}^{d}(\partial_{x_i}P_{ii})\xi_i + P_{ii}\partial_{x_i}\xi_i - \frac{2}{\omega^2}(P\xi, P\nabla \mathrm{E})_2\right]_{|x}\end{aligned} \tag{5.63}$$

Since we are evaluating eq. (5.60) along the surface of the ball $\partial B_r(x_a)$ ($\partial B_r(x_b)$ resp.) we have that $[(\partial_{z_i}P_{ii})\xi_i]_{z+x_a} = r^{-1}[\partial_{z_i}P_{ii}(z+x_a)z_i] > 0$ (I), $[P_{ii}\partial_{z_i}\xi_i]_{|z+x_a} = r^{-1}P_{ii}(z+x_a) > 0$. Finally, since the balls are centered at the minima of $\mathrm{E}(x)$, $\nabla \mathrm{E}_{|z+a} \parallel z$ and in particular $\mathrm{sign}(\partial_{z_i}\mathrm{E}(z+x_a)) = -\mathrm{sign}(z_i)$, from which we conclude that

$$\nabla \cdot (P\xi\mu^\star)_{|x} > 0 \qquad \forall x \in \partial B_r(x_a) \quad (\partial B_r(x_b) \text{ resp.}) \tag{5.64}$$

The integral on top of eq. (5.59) is positive, as proved by eq. (5.56). Using (II) $\mathrm{E}(z+x_a) \leq$

$\mathrm{E}(x_b) \leq \mathrm{E}(z + x_b)$ in eq. (5.59), we obtain

$$\begin{aligned}
&\int_{\partial B_r} -[\mathrm{E}\nabla \cdot (P\xi\mu^\star)]_{|z+x_a} + [\mathrm{E}\nabla \cdot (P\xi\mu^\star)]_{|z+x_b}\, d_z\sigma \\
&\geq \int_{\partial B_r} -\mathrm{E}_{|x_b} \nabla \cdot [(P\xi\mu^\star)_{|z+x_a} - (P\xi\mu^\star)_{|z+x_b}]\, d_z\sigma \\
&\geq \mathrm{E}(x_b) \int_{\partial B_r} \nabla \cdot [(P\xi\mu^\star)_{|z+x_b} - (P\xi\mu^\star)_{|z+x_a}]\, d_z\sigma = 0
\end{aligned} \tag{5.65}$$

This means that the right-hand side of eq. (5.58) is positive, and therefore

$$\int_{\partial B_r} \mu^\star(z + x_a) - \mu^\star(z + x_b)\, d_z\sigma \geq 0 \tag{5.66}$$

Finally, given a maximal radius $r_{\max}$ such that the hypothesis (I), and (II) hold, we will have that the conditions also hold for all $r \in [0, r_{\max})$, from which it follows that

$$\begin{aligned}
&\int_0^{r_{\max}} \int_{\partial B_r} \mu^\star(z + x_a) - \mu^\star(z + x_b)\, d_z\sigma\, dr \\
&= \int_{B_{r_{\max}}(x_a)} \mu^\star(x)\, dx - \int_{B_{r_{\max}}(x_b)} \mu^\star(x)\, dx \geq 0.
\end{aligned} \tag{5.67}$$

□

In plain terms, the theorem states that for systems whose drift is almost the gradient of a potential $\mathrm{E}(x)$, most of the measure mass concentrates around the deepest minima of $\mathrm{E}(x)$.

5.4 Input-driven *voltage* model and stochastic memory retrieval

Finally, we illustrate the theoretical results by applying them to the IDP *voltage* model to answer our original questions. Specifically, we adopt the framework in [103], where the interaction between an external input and the connectivity matrix governs whether the dynamics are globally or B_r-contracting. For expositional and visual clarity, we focus on a two-dimensional system.

Definition 5.4.1 (Input-driven *voltage* model). Let $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^2_{\geq 0}$ be the constant external input. The Input-driven *voltage* dynamics are

$$\begin{cases} \dot{x}_{\mathrm{H}} = -x_{\mathrm{H}} + W(u)\Psi(x_{\mathrm{H}}) \\ x_{\mathrm{H}}(0) \in \mathbb{R}^2 \end{cases} \tag{v}$$

with activation function $\Psi(x) = (\tanh(\beta x_1), \tanh(\beta x_2))$ and synaptic matrix

$$\begin{aligned} W(u) &= \frac{1}{2}u_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} + \frac{1}{2}u_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} \\ &= \frac{1}{2}u_1 \xi^1 {\xi^1}^\top + \frac{1}{2}u_2 \xi^2 {\xi^2}^\top \end{aligned} \tag{5.68}$$

The *voltage* dynamics in the plane are associated with the Energy (potential) function

$$\mathrm{E}_H(x_\mathrm{H}) = -\frac{1}{2}\Psi(x_\mathrm{H})^\top W(u)\Psi(x_\mathrm{H}) + x_\mathrm{H}^\top \Psi(x_\mathrm{H}) - \sum_{i=1}^{2} \int_0^{[x_\mathrm{H}]_i} \psi(s)\, ds \tag{5.69}$$

and its dynamics can be rewritten as $\dot{x}_\mathrm{H} = -J_x\Psi(x_\mathrm{H})^{-1}\nabla \mathrm{E}_H(x_\mathrm{H})$, where $J_x\Psi(x_\mathrm{H})^{-1}$ is the inverse of the Jacobian of the activation function. Notice that, for the given activation function, we have

$$J_x\Psi(x_\mathrm{H})^{-1} = \begin{pmatrix} [\beta(1-\tanh(\beta[x_\mathrm{H}]_1)^2)]^{-1} & 0 \\ 0 & [\beta(1-\tanh(\beta[x_\mathrm{H}]_2)^2)]^{-1} \end{pmatrix} \tag{5.70}$$

In addition

$$\partial_{x_i}(J_x\Psi(x_\mathrm{H})_{ii}^{-1}) = 2\tanh(\beta[x_\mathrm{H}]_i)[\beta(1-\tanh(\beta[x_\mathrm{H}]_i)^2)]^{-1} \qquad i = 1, 2 \tag{5.71}$$

so that $\mathrm{sign}(\partial_{x_i}(J_x\Psi(x_\mathrm{H})_{ii}^{-1})) = \mathrm{sign}([x_\mathrm{H}]_i)$. Thus, the conditions of Theorem 5.3.6 are satisfied.

We now want to leverage the result of Theorem 5.2.2 and numerically verify that when $f(x) = -x + W(u)\Psi(x)$ is the globally contracting drift term of a stochastic process, the associated time-varying measure converges to a stationary measure. Focusing on the drift term, in [103] it was proved that if $\max\{u_1, u_2\} < \beta^{-1}$, then the system has a unique globally exponentially stable equilibrium point - the origin. Choosing $\beta = 2$ and $u_1 = 0.2$, $u_2 = 0.25$, the *voltage* dynamics become globally contracting in the 2-norm (see [122, Ch. 2] for an overview on Contraction theory) with contraction rate $c = .5$. Choosing a diffusion matrix

$$G(x) = 0.4 \cdot \mathrm{diag}(\sin(x_1), \cos(x_2)) \tag{5.72}$$

that is both Lipschitz with $L_G = 0.32$ and sublinear with $s_G = 0.16$, we can observe from Fig. 5.5(b) that the stationary measure resembles a Gaussian centered at the origin. Furthermore, the logarithmic scale plot in Fig. 5.5(b) reveals the exponential convergence towards the stationary measure predicted by Theorem 5.2.2.

We are now interested in veryfying the results of Theorem 5.3.6. When $u_1 > \beta^{-1}$, then $\gamma_1 \xi^1$

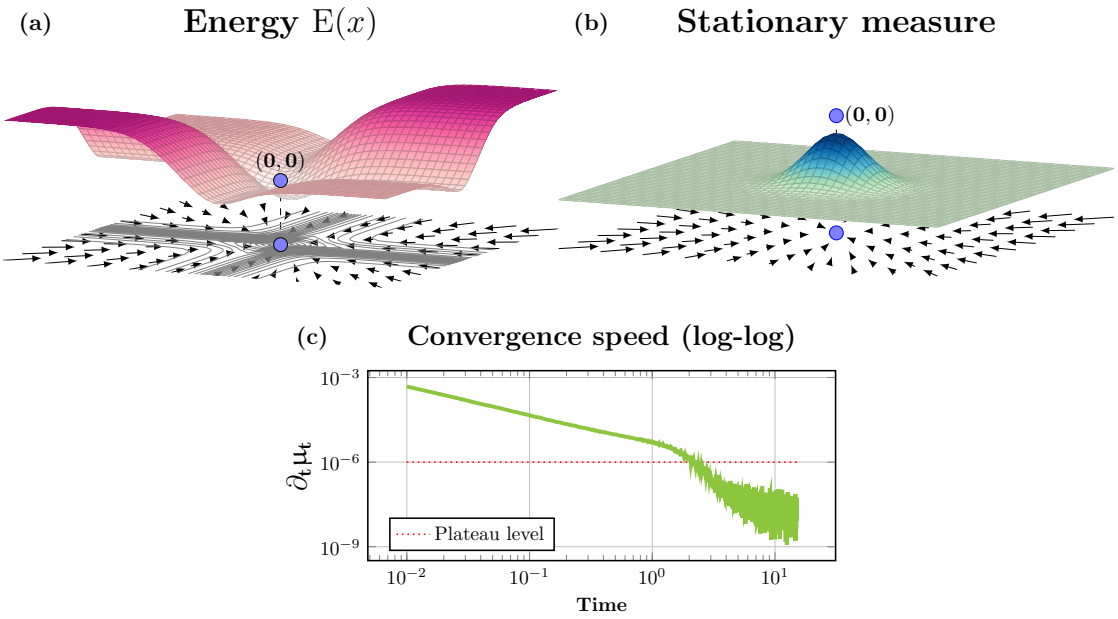


Figure 5.5: **Stochastic *voltage* model with a globally contracting drift term**. (a) Energy function associated to the *voltage* model, with a unique global minimum in the origin. (b) The stationary measure is a gaussian centered at the origin - the globally asymptotically stable equilibrium point of the drift term. The black arrows beneath panels (a-b) are the streamlines associated to the drift f. (c) Exponential trend of convergence towards stationarity, conformably with the results of Theorem 5.2.2.

and $-\gamma_1\xi^1$ are equilibrium points for the drift deterministic dynamics, with $\gamma_1 = u_1 \tanh(\beta \cdot \gamma_1)$. This condition extends to u_2 and ξ^2. The convergence of the deterministic dynamics to either of the multiple equilibrium points is guaranteed by the Energy (potential) function $\mathrm{E}_H(x_\mathrm{H})$. In addition, it was proven in [103] that if $u_2 > u_1$ then $\mathrm{E}_H(\pm\gamma_2\xi^2) < \mathrm{E}_H(\pm\gamma_1\xi^1) < 0$. By choosing $3 = u_2 > u_1 = 1 > \beta^{-1}$, constant diffusion term $G(x_\mathrm{H}) \equiv .4 \cdot \mathcal{I}_2$, and $r = 0.4$, the hypothesis (i) and (ii) of Theorem 5.3.6 are verified. Then we observe from Fig. 5.6(b) that most of the probability mass concentrates around the stable equilibria $\gamma_2\xi^2$ and $-\gamma_2\xi^2$, which are associated to the largest input u_2. The numerical simulation validates the results of Theorem 5.3.6, with Fig. 5.6(b) verifying the results in [113] on the exponential transient towards the stationary distribution.

Simulations exploit a particle method with $N = 2000$ trajectories, Gaussian kernel smoothing (variance $2 \cdot \mathcal{I}_2$), and time step $\delta t = 0.01$. The 2-norm of the difference in the measure over time was used to monitor convergence, with threshold 1e-6 indicating stationarity.

5.5 Conclusion

Despite the intuitive nature of concentration results, asymptotic behavior of measures is a fragile phenomenon that requires careful mathematical manipulation. In this chapter we have expanded the literature on the convergence of measures associated to globally contracting drift terms and spatially dependent diffusion terms. In particular, we have observed that convergence to a unique stationary measure is guaranteed as long as the contraction rate is twice as big as the Lipschitz constant of the diffusion term. Additionally, we have focused on measures associated to stochastic processes characterized by a B_r-contracting drift term and constant diffusion term. Theoretical results and numerical simulations reveal how the concentration of mass around the stable equilibria of the drift term depends on the radius of the largest ball centered at the equilibrium and within its basin of attraction as well as the local contraction rate. Moreover, if a non-convex potential $\mathrm{E}(x)$ for the drift term exists, then most of the stationary measure concentrates around its deepest minima. For the sake of completeness, we point out that Proposition 5.3.5 and Theorem 5.3.6 could be generalized to space of functions with weak derivatives (Sobolev spaces), perhaps at the cost of more burdensome notation. Future works may explore convergence to stationarity of measures associated to B_r-contracting drift terms and spatially dependent diffusion terms. Finally, numerical investigations may benefit from more extensive experiments using a portfolio of methods for the simulation of partial differential equations as well as dedicated optimal transport libraries.

This chapter is based on the works [103], [123], the first published in the journal *Science Advances* and the second currently available as pre-print on *arXiv*. The next chapter provides a

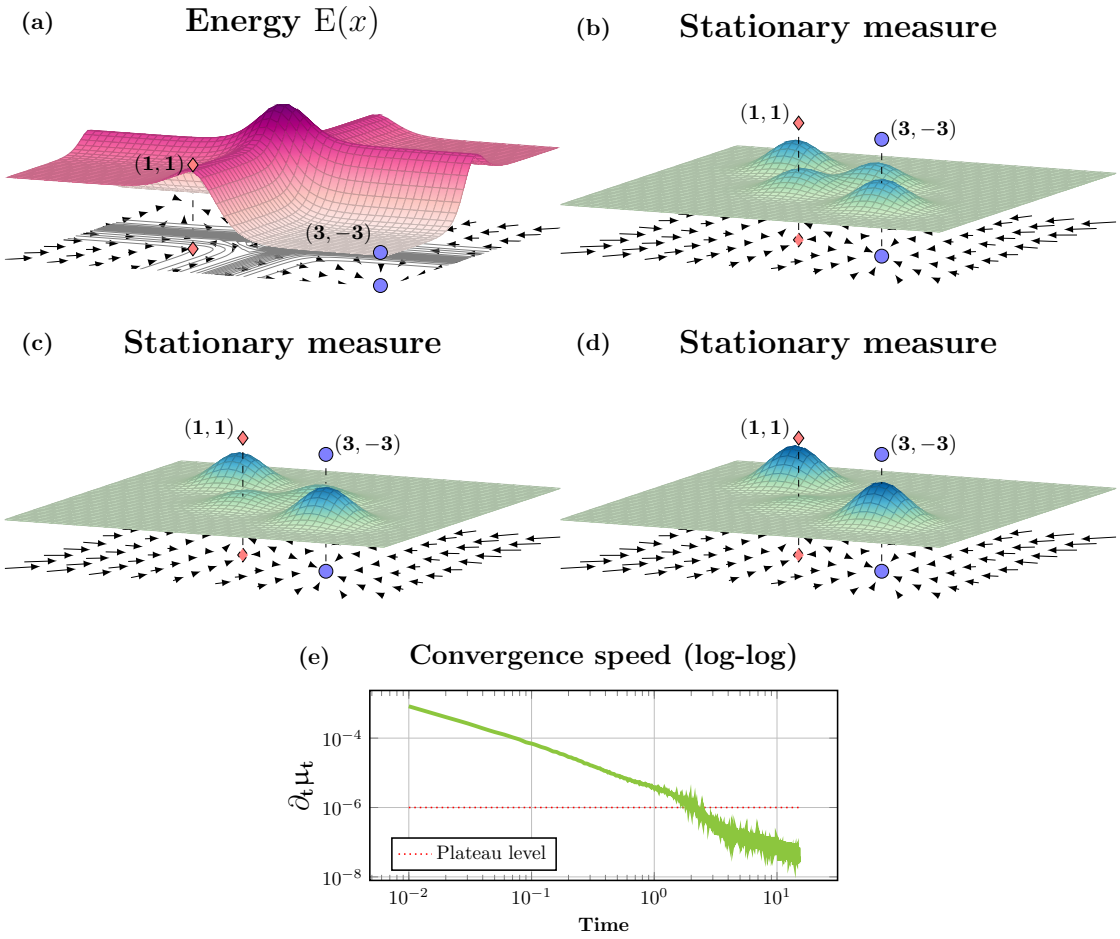


Figure 5.6: **Stochastic *voltage* model with a B_r-contracting drift term.** (a) Energy associated to the multistable *voltage* model, with antisymmetric minima w.r.t. to the origin. As observable, the minima $(3,-3)$ and $(-3,3)$ associated with the input u_2 are far deeper than the minima $(1,1)$ and $(-1,-1)$ associated with the input u_1. (b-d) The stationary measure concentrates its mass around the stable equilibrium points associated with the deepest Energy $\mathrm{E}(x)$ minima, as predicted by Theorem 5.3.6. In particular, from (b) to (d) we have stationary measures associated with increasingly higher noise amplitudes. As observable, the higher the noise amplitude, the greater the mass within the deepest basins of attraction. For noise amplitudes far beyond the one underlying simulation (d), almost all the particles concentrate and expand within the deepest basin, following the shallow geometry of $\mathrm{E}(x)$ (e) Exponential trend of convergence towards stationarity of the measure in (d), conformably with the results in [113]. The convergence speed for (b) and (c) is analogous.

synthesis of the results obtained and a forward-looking discussion of open challenges and avenues for future research.

Appendix

Lemma 5.5.1 (Integration by parts). *Let $\mathcal{M} \subseteq \mathbb{R}^N$ be a smooth manifold with boundary, and let $A : \mathcal{M} \to \mathbb{R}^{N\times N}$ and $v : \mathcal{M} \to \mathbb{R}^N$ be differentiable. Then*

$$\int_{\partial\mathcal{M}} (\nabla \cdot A(x), v(x))_2 \, d_x\sigma = -\int_{\partial\mathcal{M}} \mathrm{Tr}\left((A(x)J_x v(x)^\top)\right) \, d_x\sigma \tag{5.73}$$

Proof. The formula for the divergence of a matrix-valued field gives $(\nabla\cdot A(x))_i = \sum_{j=1}^d \partial_{x_j} A_{ij}(x)$ and consequently $(\nabla \cdot A(x), v(x))_2 = \sum_{i,j=1}^d [\partial_{x_j} A_{ij}(x)]v_i(x)$. Thus, evaluating the integral

$$\begin{aligned}
&\int_{\partial\mathcal{M}} (\nabla \cdot A, v)_{2|x} \, d_x\sigma \;= \int_{\partial\mathcal{M}} \sum_{ij}^d [\partial_{x_j} A_{ij}]_{|x} v_{i|x} \, d_x\sigma \\
=&\sum_{ij}^d \int_{\partial\mathcal{M}} \frac{d}{dx_j}[A_{ij}v_i]_{|x} - [A_{ij}\partial_{x_j} v_i]_{|x} \, d_x\sigma \\
=&\sum_{ij}^d \Big[\underbrace{\int_{\partial(\partial\mathcal{M})} [A_{ij}v_i]_{|x} \, d_x\gamma}_{=0} - \int_{\partial\mathcal{M}} [A_{ij}\partial_{x_j} v_i]_{|x} \, d_x\sigma \Big] \\
=&-\int_{\partial\mathcal{M}} \mathrm{Tr}\left((AJ_x v^\top)_{|x}\right) \, d_x\sigma
\end{aligned} \tag{5.74}$$

where in the third passage we have used generalized Stokes theorem [124] on a boundary $\partial\mathcal{M}$ of a smooth manifold $\mathcal{M}$, which has no boundary $\partial(\partial\mathcal{M})$. □

6

*Fin*al r*e*marks

The final chapter of this thesis offers a reflective perspective on the evolving dialogue between theoretical neuroscience and machine learning, with particular emphasis on the role of dynamical systems as a unifying framework. For many years, the two disciplines appeared to develop in relative isolation. Theoretical neuroscience advanced models capable of capturing biophysical detail or cognitive phenomena, yet these models often struggled to translate into practical algorithms for large-scale data analysis. Conversely, machine learning has progressed at an extraordinary pace, achieving breakthroughs in domains such as image recognition, natural language processing, and generative modeling. Despite this success, many state-of-the-art approaches remain largely detached from biological principles, offering only superficial or computer-centric interpretations of intelligence.

In recent years, however, this separation has begun to erode. A growing body of research in machine learning has drawn inspiration from neuroscience—not only in terms of architecture, but also in concepts such as energy-based formulations, recurrent dynamics, and mechanisms of memory and attention. At the same time, theoretical neuroscience increasingly borrows tools and perspectives from machine learning, both to analyze neural data and to sharpen its models of computation. This emerging convergence suggests that dynamical systems theory, which has long served as a bridge between abstract mathematics and natural processes, may provide the common ground needed to advance both fields.

Building on these considerations, the next sections will turn to a series of open problems that remain central to both theoretical neuroscience and machine learning. We will argue that viewing these challenges through the lens of dynamical systems not only provides a unifying mathematical language, but also highlights deeper connections between biological plausibility, algorithmic efficiency, and interpretability.

6.1 Sequential retrieval in associative memory models

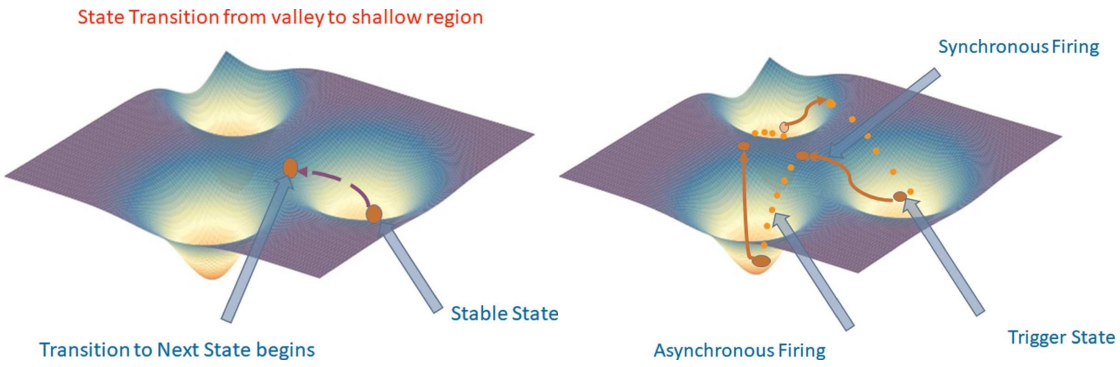
Sequential retrieval — the reliable reproduction of ordered patterns of neural activity — is a canonical problem at the intersection of neuroscience and machine learning. Unlike classical associative memory, where a network converges to a single stored attractor, sequential retrieval requires the system to traverse a prescribed trajectory of states. This idea was crystallized in David Kleinfeld's seminal paper *"Sequential State Generation by Model Neural Networks"* [125], which demonstrated how asymmetric connectivity could bias transitions between attractors, thereby extending Hopfield's paradigm from static memories to ordered sequences. Kleinfeld's insight opened the way to studying memory not only as a set of fixed points but also as structured trajectories in phase space.

Subsequent research identified several distinct mechanisms. One prominent family employs temporally asymmetric Hebbian learning rules, such as those formalized in spike-timing-dependent plasticity, to imprint directed transitions between patterns [126], [127]. A second family builds sequences from clustered attractors [128] or synfire chains [129], where activity propagates through assemblies of neurons in a feedforward fashion. A third relies on continuous attractors, in which activity bumps drift smoothly along a manifold [130] under oscillatory gating or slow adaptation. Each of these mechanisms exemplifies a dynamical-systems principle: heteroclinic chains of metastable states, low-dimensional slow manifolds, or bifurcation-driven transitions.

More recent developments bridge these models with modern machine learning. Rajan et al. [131] showed that recurrent reservoirs can generate reproducible neural sequences once shaped by learning, while modern Hopfield networks extend energy-based formulations to sequential retrieval through temporally asymmetric interaction terms [84]. Input-modulated retrieval schemes [100] further highlight how external signals can flexibly control sequence initiation, tempo, and branching, echoing the role of gating and neuromodulation in biological circuits.

From a dynamical perspective, sequential retrieval corresponds to guiding trajectories through structured regions of state space. Stability analyses clarify why some sequences are robust to noise while others collapse into single attractors. Timescale separation between fast spiking variables and slow adaptation or plasticity determines retrieval tempo [133]. Phase-space tools such as Lyapunov exponents, bifurcation analysis, and energy landscapes offer a principled way to measure robustness and capacity, while operator-theoretic methods (e.g., Koopman analysis) are emerging as diagnostics for high-dimensional sequence dynamics.

Despite progress, several open problems remain. First, scaling laws for sequence capacity are not yet fully established: although "long-sequence" Hopfield variants improve storage, the limits for overlapping or noisy sequences under biological constraints remain unclear. Second, the



(a) (b)

Figure 6.1: **Sequence retrieval in associative memory models**. Visualization of input-driven memory transition over an Energy landscape, adapted from [132]. (left) The system transitions from an attractor basin toward a saddle region, illustrating the mechanism for moving between stored memory states. (right) The dynamics follow a limit cycle over the energy landscape, demonstrating sustained oscillatory trajectories that traverse multiple regions of the landscape. Together, these panels highlight how the energy landscape shapes both discrete transitions and continuous sequence dynamics in associative memory networks.
labelfig: seq

precise plasticity rules that support reliable sequential transitions in vivo are still debated; while STDP-like rules produce asymmetry, their sufficiency for long and flexible sequences remains unresolved. Third, the control of tempo and branching is poorly understood: biological systems can replay sequences faster, slower, or in reverse, yet most models offer only limited flexibility. Finally, new methods are needed to reconstruct the geometry of sequential attractors from data, linking dynamical trajectories to experimental recordings.

In summary, Kleinfeld's original proposal of attractor transitions inaugurated a field where associative memory is reinterpreted as the generation of structured trajectories (see Fig. **??** for an exemplification of sequence retrieval). Contemporary work — from asymmetric Hebbian chains to modern Hopfield extensions and reservoir learning — demonstrates the rich interplay between biological plausibility and algorithmic utility. The dynamical-systems framework not only provides the conceptual vocabulary (attractors, manifolds, heteroclinic chains) but also offers analytical tools for addressing the central challenges of scalability, robustness, and control in sequential memory.

6.2 Hebbian learning: biological plausibility, dynamics and limits

Hebbian learning — the activity-dependent strengthening or weakening of synapses — is a foundational principle in both neuroscience and theoretical models of memory. Despite its

conceptual appeal, Hebbian learning remains poorly understood in terms of dynamics, stability, and large-scale application. Dong and Hopfield [134] analyzed networks with adapting synapses, demonstrating that coupling neural activity with slow synaptic variables reshapes phase-space, creates metastable regimes, and can be characterized using Lyapunov-like constructions. Their results showed that memory states become slow-manifold attractors rather than fixed points, and that bifurcations can emerge depending on plasticity strength and timescales, highlighting the subtle interplay between stability and learning.

Subsequent work has extended these insights. Temporally asymmetric Hebbian rules, such as STDP, introduce directionality and support sequence learning [135] but require careful tuning to avoid instability. Short-term synaptic dynamics, modeled via facilitation and depression [92], modulate effective Hebbian updates, producing transient memory and filtering effects. Neuromodulator-gated and heterosynaptic variants allow context-dependent plasticity, further expanding functional flexibility [101], [136].

From a machine-learning perspective, Hebbian rules are appealing for their locality, low computational cost, and natural suitability for continual learning. However, they face practical obstacles: instability, lack of global credit assignment, and poor scaling in deep networks. Recent approaches aim to mitigate these issues, e.g., FastHebb algorithms for efficient representation learning [137], neo-Hebbian rules for continual learning [138], and hybrid reinforcement-learning schemes exploiting local plasticity for adaptation [139].

The dynamical-systems viewpoint unifies these observations. Treating synaptic weights as slow variables coupled to fast neural activity creates a multiscale system whose behavior can be analyzed via slow-manifold reductions, Lyapunov or energy arguments, and bifurcation analysis. Timescale separation, nonlinear plasticity, and homeostatic constraints determine whether Hebbian updates sculpt useful structure or lead to runaway dynamics. Multiplicative normalization and resource constraints act as effective conserved quantities, preserving memory geometry while stabilizing the network.

Open problems remain central. Scaling laws for online, biologically constrained Hebbian networks are incomplete; stable Hebbian rules that integrate with global credit assignment in deep architectures remain an open challenge; translation to neuromorphic hardware and continual-learning benchmarks requires careful algorithmic and implementation design; and experimental validation of model predictions — e.g., slow attractor drifts or plasticity-induced bifurcations — is largely untapped.

In sum, Hebbian learning embodies both a biologically grounded mechanism and a source of theoretical puzzles. Dong and Hopfield's analysis of adapting synapses frames plasticity as an integral part of network dynamics, providing a lens for understanding stability, memory, and scalability. Progress in this area promises both a deeper understanding of biological learning and

pathways toward energy-efficient, online, and locally implementable learning rules in artificial systems.

6.3 The advantages of the dynamical systems approach in neuroscience and machine learning

The preceding sections illustrate how dynamical systems theory provides a unifying language for understanding complex neural and computational phenomena. From sequential retrieval in associative memory to Hebbian plasticity, the same conceptual toolkit—phase-space analysis, attractors, slow manifolds, bifurcations, and Lyapunov functionals—illuminates both biological mechanisms and algorithmic principles. This convergence highlights several key advantages of framing neuroscience and machine learning problems through a dynamical-systems lens.

First, *dynamical systems offer a principled framework for capturing temporal evolution*. Neural circuits are inherently time-dependent: activity patterns unfold across multiple timescales, synaptic strengths adapt slowly, and behavior emerges from the interaction of these fast and slow variables. By representing neural networks as dynamical systems, one can formalize these interactions and predict how trajectories evolve, whether in the reliable replay of sequential memories or in the emergence of metastable states under Hebbian plasticity. The sequential retrieval models discussed above exemplify this: heteroclinic chains, slow manifolds, and bifurcation-driven transitions all naturally emerge from a dynamical-systems formulation, enabling quantitative analyses of stability, robustness, and capacity.

Second, the approach provides a *common mathematical language bridging biology and computation*. Concepts such as attractors, energy landscapes, and slow-fast decompositions describe phenomena both in real neural tissue and in artificial networks. In Hebbian learning, slow synaptic adaptation coupled to fast neural dynamics generates rich multiscale behavior: the same formalism explains memory consolidation in cortex and guides algorithm design for online or continual learning. By encoding rules of plasticity as dynamical flows, researchers can explore stability, capacity, and bifurcation structure in ways that are transparent and generalizable.

Third, *dynamical systems facilitate modular and interpretable analysis of complex architectures*. High-dimensional neural networks—whether biological or artificial—can be decomposed into interacting subsystems with identifiable slow and fast modes. In sequential memory networks, this decomposition explains how network trajectories can traverse structured regions of state space robustly, how noise affects sequence fidelity, and how external signals or neuromodulators modulate tempo and branching. Similarly, in Hebbian plasticity, slow manifold analysis and Lyapunov methods provide insight into when local, activity-dependent updates stabilize network function versus when they induce drift or instability.

Fourth, the *dynamical systems perspective supports cross-disciplinary innovation*. Tools initially developed in physics and applied mathematics—bifurcation theory, operator methods, Koopman analysis, and energy-based formulations—translate naturally to both experimental neuroscience and machine learning. This enables rigorous analysis of network properties, guides the design of biologically inspired algorithms, and facilitates the benchmarking of neural plausibility versus computational efficiency. For example, modern Hopfield networks and input-modulated retrieval schemes leverage dynamical principles to combine stability, flexible sequencing, and control in ways that echo neurobiological architectures.

Finally, *dynamical systems foster the identification of open problems and experimental predictions*. By characterizing memory, plasticity, and computation in phase space, researchers can quantify how capacity scales with network size, how sequences degrade under noise, or how plasticity rules shape attractor landscapes. These predictions are not only mathematically precise but also experimentally testable, providing a bridge between theory and data.

In summary, adopting a dynamical-systems viewpoint confers multiple advantages: it captures temporal evolution across scales, unifies disparate phenomena under a common formalism, enables modular and interpretable analysis, fosters cross-disciplinary innovation, and grounds experimental predictions in rigorous mathematics. Whether applied to sequential memory, Hebbian plasticity, or emerging models of neural computation, this perspective clarifies both the principles underlying biological intelligence and the design of efficient, interpretable algorithms in machine learning. By situating both fields within a shared dynamical framework, researchers gain a powerful lens for understanding, predicting, and ultimately harnessing the complex behaviors of neural systems.

References

[1] X. Liao, A. V. Vasilakos, and Y. He, "Small-world human brain networks: Perspectives and challenges," *Neuroscience & Biobehavioral Reviews*, vol. 77, 286–300, Jun. 2017 (Cited in page 1).

[2] S. Haykin, *Neural networks and learning machines, 3/E*. Pearson Education India, 2009 (Cited in page 2).

[3] E. R. Kandel, "The molecular biology of memory storage: A dialogue between genes and synapses," *Science*, vol. 294, no. 5544, 1030–1038, Nov. 2001 (Cited in page 2).

[4] L. Kozachkov, J. J. Slotine, and D. Krotov, "Neuron–astrocyte associative memory," *Proceedings of the National Academy of Sciences*, vol. 122, no. 21, May 2025 (Cited in pages 2, 24, 25, 69, 74).

[5] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, S. Mack, *et al.*, *Principles of neural science*. McGraw-hill New York, 2000, vol. 4 (Cited in page 4).

[6] M. J. A. M. van Putten, *Dynamics of Neural Networks: A Mathematical and Clinical Approach*. Springer Berlin Heidelberg, 2020 (Cited in page 3).

[7] J. E. LeDoux, "Emotion circuits in the brain," *Annual Review of Neuroscience*, vol. 23, no. 1, 155–184, Mar. 2000 (Cited in page 3).

[8] H. Eichenbaum, "A cortical–hippocampal system for declarative memory," *Nature Reviews Neuroscience*, vol. 1, no. 1, 41–50, Oct. 2000 (Cited in page 3).

[9] E. T. Rolls, "Limbic systems for emotion and for memory, but no single limbic system," *Cortex*, vol. 62, 119–157, Jan. 2015 (Cited in page 3).

[10] L. R. Squire, "Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans.," *Psychological Review*, vol. 99, no. 2, 195–231, 1992 (Cited in page 5).

[11] H. Eichenbaum, "The hippocampus and declarative memory: Cognitive mechanisms and neural codes," *Behavioural Brain Research*, vol. 127, no. 1–2, 199–207, Dec. 2001 (Cited in page 5).

[12] E. A. Phelps, "Human emotion and memory: Interactions of the amygdala and hippocampal complex," *Current Opinion in Neurobiology*, vol. 14, no. 2, 198–202, Apr. 2004 (Cited in page 5).

[13] J. E. LeDoux, "Evolution of human emotion," in *Evolution of the Primate Brain*. Elsevier, 2012, 431–442 (Cited in page 5).

[14] W. B. Scoville and B. Milner, "Loss of recent memory after bilateral hippocampal lesions," *Journal of Neurology, Neurosurgery &; Psychiatry*, vol. 20, no. 1, 11–21, Feb. 1957 (Cited in page 6).

[15] G. Buzsáki, "Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning," *Hippocampus*, vol. 25, no. 10, 1073–1188, Sep. 2015 (Cited in page 6).

[16] D. Marr, "Simple memory: A theory for archicortex," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 262, no. 841, 23–81, Jul. 1971 (Cited in page 6).

[17] A. Treves and E. T. Rolls, "Computational analysis of the role of the hippocampus in memory," *Hippocampus*, vol. 4, no. 3, 374–391, Jun. 1994 (Cited in page 6).

[18] O. S. Vinogradova, "Hippocampus as comparator: Role of the two input and two output systems of the hippocampus in selection and registration of information," *Hippocampus*, vol. 11, no. 5, 578–598, Oct. 2001 (Cited in page 6).

[19] M. A. Yassa and C. E. L. Stark, "Pattern separation in the hippocampus," *Trends in Neurosciences*, vol. 34, no. 10, 515–525, Oct. 2011 (Cited in page 6).

[20] T. Hafting, M. Fyhn, S. Molden, M. B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, 801–806, Jun. 2005 (Cited in page 7).

[21] E. I. Moser, E. Kropff, and M. B. Moser, "Place cells, grid cells, and the brain's spatial representation system," *Annual Review of Neuroscience*, vol. 31, no. 1, 69–89, Jul. 2008 (Cited in page 7).

[22] L. R. Squire, L. Genzel, J. T. Wixted, and R. G. Morris, "Memory consolidation," *Cold Spring Harbor Perspectives in Biology*, vol. 7, no. 8, a021766, Aug. 2015 (Cited in page 7).

[23] G. W. van Hoesen, B. T. Hyman, and A. R. Damasio, "Entorhinal cortex pathology in alzheimer's disease," *Hippocampus*, vol. 1, no. 1, 1–8, Jan. 1991 (Cited in page 7).

[24] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Bulletin of Mathematical Biology*, vol. 52, no. 1–2, 25–71, 1990 (Cited in pages 8, 41).

[25] D. Krotov and J. J. Hopfield, "Large associative memory problem in neurobiology and machine learning," in *International Conference on Learning Representations*, 2020 (Cited in pages 9, 27, 69, 70, 86).

[26] J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons.," *Proceedings of the National Academy of Sciences*, vol. 81, no. 10, 3088–3092, May 1984 (Cited in pages 10, 12, 22, 27, 29, 41, 43, 69, 70).

[27] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, 2554–2558, Apr. 1982 (Cited in pages 12, 15, 27, 28, 41, 49, 69, 70, 72).

[28] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, no. 2, 147–160, Apr. 1950 (Cited in page 12).

[29] W. Gerstner and W. M. Kistler, "Mathematical formulations of hebbian learning," *Biological Cybernetics*, vol. 87, no. 5–6, 404–415, Dec. 2002 (Cited in pages 13, 22, 49).

[30] A. Hedayat and W. D. Wallis, "Hadamard matrices and their applications," *The Annals of Statistics*, vol. 6, no. 6, Nov. 1978 (Cited in page 13).

[31] S. M. Ross, *Introductory statistics*. Elsevier, 2010 (Cited in page 17).

[32] R. McEliece, E. Posner, E. Rodemich, and S. Venkatesh, "The capacity of the hopfield associative memory," *IEEE Transactions on Information Theory*, vol. 33, no. 4, 461–482, Jul. 1987 (Cited in pages 18, 62).

[33] D. Petritis, "Thermodynamic formalism of neural computing," in *Dynamics of Complex Interacting Systems*. Springer Netherlands, 1996, 81–146 (Cited in pages 18, 27, 30, 31, 38, 39).

[34] L. Meneghetti, "Towards a continuous dynamic model of the hopfield theory on neuronal interaction and memory storage," M.S. thesis, University of Padua, 2018 (Cited in page 18).

[35] D. Golomb, N. Rubin, and H. Sompolinsky, "Willshaw model: Associative memory with sparse coding and low firing rates," *Physical Review A*, vol. 41, no. 4, 1843–1854, Feb. 1990 (Cited in pages 19, 45).

[36] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014 (Cited in pages 19, 42, 58, 65, 69, 72).

[37] C. J. Colbourn and J. H. Dinitz, *CRC handbook of combinatorial designs*. CRC press, 2010 (Cited in page 20).

[38] A. Treves, "Graded-response neurons and information encodings in autoassociative memories," *Physical Review A*, vol. 42, no. 4, 2418–2430, Aug. 1990 (Cited in pages 20, 41, 42).

[39] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005 (Cited in pages 20, 42, 44, 45, 47, 48, 55, 60, 65, 69, 72).

[40] K. D. Miller and F. Fumarola, "Mathematical equivalence of two common forms of firing rate models of neural networks," *Neural Computation*, vol. 24, no. 1, 25–31, Jan. 2012 (Cited in page 21).

[41] F. Fumarola, "The synaptic weight matrix: Dynamics, symmetry breaking, and disorder," Ph.D. dissertation, Columbia University, 2018 (Cited in page 21).

[42] M. A. Cohen and S. Grossberg, "Absolute stability of global pattern formation and parallel memory storage by competitive neural networks," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 5, 815–826, Sep. 1983 (Cited in pages 22, 41, 43).

[43] D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," in *Advances in neural information processing systems*, vol. 29, 2016 (Cited in pages 22, 27, 69).

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017 (Cited in page 22).

[45] H. Ramsauer, B. Schäfl, J. Lehner, *et al.*, "Hopfield networks is all you need," in *International Conference on Learning Representations*, 2021 (Cited in pages 22, 28, 69, 70, 88).

[46] B. Hoover, Y. Liang, B. Pham, R. Panda, H. Strobelt, D. H. Chau, M. Zaki, and D. Krotov, "Energy transformer," *Advances in neural information processing systems*, vol. 36, pp. 27 532–27 559, 2023 (Cited in page 23).

[47] Q. Zhang, D. Krotov, and G. E. Karniadakis, "Operator learning for reconstructing flow fields from sparse measurements: An energy transformer approach," *Journal of Computational Physics*, vol. 538, p. 114 148, Oct. 2025 (Cited in page 23).

[48] L. Ambrogioni, "In search of dispersed memories: Generative diffusion models are associative memory networks," *Entropy*, vol. 26, no. 5, p. 381, Apr. 2024 (Cited in pages 23, 28).

[49] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021 (Cited in page 24).

[50] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Statistical mechanics of neural networks near saturation," *Annals of Physics*, vol. 173, no. 1, 30–67, Jan. 1987 (Cited in page 27).

[51] A. Crisanti, D. J. Amit, and H. Gutfreund, "Saturation level of the hopfield model for neural network," *Europhysics Letters (EPL)*, vol. 2, no. 4, 337–341, Aug. 1986 (Cited in pages 27, 69).

[52] M. Demircigil, J. Heusel, M. Lowe, S. Upgang, and F. Vermet, "On a model of associative memory with huge storage capacity," *Journal of Statistical Physics*, vol. 168, no. 2, 288–299, May 2017 (Cited in pages 27, 69).

[53] F. Blanchini and S. Miani, *Set-Theoretic Methods in Control*. Springer International Publishing, 2015 (Cited in pages 33, 57).

[54] S. Betteti, G. Baggio, and S. Zampieri, "On the capacity of continuous-time hopfield models," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, IEEE, Dec. 2024, 7853–7858 (Cited in page 38).

[55] S. I. Amari, "Learning patterns and pattern sequences by self-organizing nets of threshold elements," *IEEE Transactions on Computers*, vol. C–21, no. 11, 1197–1206, Nov. 1972 (Cited in page 41).

[56] S. I. Amari, "Neural theory of association and concept-formation," *Biological Cybernetics*, vol. 26, no. 3, 175–185, 1977 (Cited in page 41).

[57] M. V. Tsodyks and M. V. Feigel'man, "The enhanced storage capacity in neural networks with low activity level," *Europhysics Letters (EPL)*, vol. 6, no. 2, 101–105, May 1988 (Cited in pages 41, 42, 62).

[58] A. Treves, "Threshold-linear formal neurons in auto-associative nets," *Journal of Physics A: Mathematical and General*, vol. 23, no. 12, 2631–2650, Jun. 1990 (Cited in pages 41, 42, 62).

[59] D. J. Amit and M. V. Tsodyks, "Quantitative study of attractor neural networks retrieving at low spike rates: Ii. low-rate retrieval in symmetric networks," *Network: Computation in Neural Systems*, vol. 2, no. 3, 275–294, Jan. 1991 (Cited in pages 41, 42).

[60] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical Journal*, vol. 1, no. 6, 445–466, Jul. 1961 (Cited in page 42).

[61] C. Morris and H. Lecar, "Voltage oscillations in the barnacle giant muscle fiber," *Biophysical Journal*, vol. 35, no. 1, 193–213, Jul. 1981 (Cited in page 42).

[62] P. C. Bressloff and S. Coombes, "Dynamics of strongly coupled spiking neurons," *Neural Computation*, vol. 12, no. 1, 91–129, Jan. 2000 (Cited in page 42).

[63] N. Brunel, "Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons," *Journal of Computational Neuroscience*, vol. 8, no. 3, 183–208, May 2000 (Cited in page 42).

[64] A. N. Burkitt, "A review of the integrate-and-fire neuron model: I. homogeneous synaptic input," *Biological Cybernetics*, vol. 95, no. 1, 1–19, Apr. 2006 (Cited in page 42).

[65] A. N. Burkitt, "A review of the integrate-and-fire neuron model: Ii. inhomogeneous synaptic input and network properties," *Biological Cybernetics*, vol. 95, no. 2, 97–112, Jul. 2006 (Cited in page 42).

[66] Y. Roudi and P. E. Latham, "A balanced memory network," *PLoS Computational Biology*, vol. 3, no. 9, K. J. Friston, Ed., e141, Sep. 2007 (Cited in page 42).

[67] G. Mongillo, S. Rumpel, and Y. Loewenstein, "Inhibitory connectivity defines the realm of excitatory plasticity," *Nature Neuroscience*, vol. 21, no. 10, 1463–1470, Sep. 2018 (Cited in page 42).

[68] G. B. Ermentrout and D. H. Terman, *Mathematical Foundations of Neuroscience*. Springer New York, 2010 (Cited in page 42).

[69] D. Horn and J. Weyers, "Hypercubic structures in orthogonal hopfield models," *Physical Review A*, vol. 36, no. 10, 4968–4974, Nov. 1987 (Cited in pages 44, 75).

[70] X. S. Zhang, *Neural Networks in Optimization*. Springer US, 2000 (Cited in page 44).

[71] T. P. Vogels, K. Rajan, and L. F. Abbott, "Neural network dynamics," *Annual Review of Neuroscience*, vol. 28, no. 1, 357–376, Jul. 2005 (Cited in page 44).

[72] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology press, 2005 (Cited in page 49).

[73] T. V. P. Bliss and A. R. Gardner-Medwin, "Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimulation of the perforant path," *The Journal of Physiology*, vol. 232, no. 2, 357–374, Jul. 1973 (Cited in page 49).

[74] Y. Munakata and J. Pfaffly, "Hebbian learning and development," *Developmental Science*, vol. 7, no. 2, 141–148, Mar. 2004 (Cited in page 49).

[75] C. v. Vreeswijk and H. Sompolinsky, "Chaotic balanced state in a model of cortical circuits," *Neural Computation*, vol. 10, no. 6, 1321–1371, Aug. 1998 (Cited in page 49).

[76] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite numbers of patterns in a spin-glass model of neural networks," *Physical Review Letters*, vol. 55, no. 14, 1530–1533, Sep. 1985 (Cited in page 50).

[77] J. A. Hertz, *Introduction to the theory of neural computation*. Crc Press, 2018 (Cited in page 50).

[78] H. K. Khalil, *Nonlinear systems*. Upper Saddle River, NJ: Prentice-Hall, 2002, The book can be consulted by contacting: PH-AID: Wallet, Lionel (Cited in pages 52, 78, 90).

[79] J. LaSalle, "Stability theory for ordinary differential equations," *Journal of Differential Equations*, vol. 4, no. 1, 57–65, Jan. 1968 (Cited in page 58).

[80] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323 (Cited in page 60).

[81] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Information storage in neural networks with low levels of activity," *Physical Review A*, vol. 35, no. 5, 2293–2303, Mar. 1987 (Cited in pages 62, 69).

[82] S. Betteti, G. Baggio, F. Bullo, and S. Zampieri, "Firing rate models as associative memory: Synaptic design for robust retrieval," *Neural Computation*, 1–32, Aug. 2025 (Cited in page 65).

[83] A. Treves and D. J. Amit, "Metastable states in asymmetrically diluted hopfield networks," *Journal of Physics A: Mathematical and General*, vol. 21, no. 14, 3155–3169, Jul. 1988 (Cited in page 69).

[84] H. Chaudhry, J. Zavatone-Veth, D. Krotov, and C Pehlevan, "Long sequence hopfield memory," in *Advances in Neural Information Processing Systems*, vol. 36, Curran Associates, Inc., 2023, pp. 54 300–54 340 (Cited in pages 69, 116).

[85] D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, Sep. 1989 (Cited in page 69).

[86] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International conference on machine learning*, PMLR, 2017, pp. 3987–3995 (Cited in page 70).

[87] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends in Cognitive Sciences*, vol. 24, no. 12, 1028–1040, Dec. 2020 (Cited in pages 70, 89).

[88] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," *Information Fusion*, vol. 58, 52–68, Jun. 2020 (Cited in pages 70, 89).

[89] F. P. Battaglia and A. Treves, "Stable and rapid recurrent processing in realistic autoassociative memories," *Neural Computation*, vol. 10, no. 2, 431–450, Feb. 1998 (Cited in page 72).

[90] B. Blumenfeld, S. Preminger, D. Sagi, and M. Tsodyks, "Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity," *Neuron*, vol. 52, no. 2, 383–394, Oct. 2006 (Cited in page 74).

[91] H. Tang, H. Li, and R. Yan, "Memory dynamics in attractor networks with saliency weights," *Neural Computation*, vol. 22, no. 7, 1899–1926, Jul. 2010 (Cited in page 74).

[92] G. Mongillo, O. Barak, and M. Tsodyks, "Synaptic theory of working memory," *Science*, vol. 319, no. 5869, 1543–1546, Mar. 2008 (Cited in pages 74, 85, 118).

[93] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, Dec. 1985 (Cited in page 77).

[94] M. Tsodyks and H. Markram, "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability," *Proceedings of the National Academy of Sciences*, vol. 94, no. 2, 719–723, Jan. 1997 (Cited in page 85).

[95] M. Tsodyks, K. Pawelzik, and H. Markram, "Neural networks with dynamic synapses," *Neural Computation*, vol. 10, no. 4, 821–835, May 1998 (Cited in page 85).

[96] A. Treves and E. T. Rolls, "Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network," *Hippocampus*, vol. 2, no. 2, 189–199, Apr. 1992 (Cited in page 85).

[97] M. Allegra, L. Posani, R. Gómez-Ocádiz, and C. Schmidt-Hieber, "Differential relation between neuronal and behavioral discrimination during hippocampal memory encoding," *Neuron*, vol. 108, no. 6, 1103–1112.e6, Dec. 2020 (Cited in page 85).

[98] C. Gastaldi, T. Schwalger, E. De Falco, R. Q. Quiroga, and W. Gerstner, "When shared concept cells support associations: Theory of overlapping memory engrams," *PLOS Computational Biology*, vol. 17, no. 12, A. Morrison, Ed., e1009691, Dec. 2021 (Cited in pages 86, 89).

[99] A. Karuvally, T. Sejnowski, and H. T. Siegelmann, "General sequential episodic memory model," in *International Conference on Machine Learning*, PMLR, 2023, pp. 15 900–15 910 (Cited in page 88).

[100] L. Herron, P. Sartori, and B. Xue, "Robust retrieval of dynamic sequences through interaction modulation," *PRX Life*, vol. 1, no. 2, Dec. 2023 (Cited in pages 89, 116).

[101] M. K. Benna and S. Fusi, "Computational principles of synaptic memory consolidation," *Nature Neuroscience*, vol. 19, no. 12, 1697–1706, Oct. 2016 (Cited in pages 89, 118).

[102] D. Kudithipudi, M. Aguilar-Simon, J. Babb, *et al.*, "Biological underpinnings for lifelong learning machines," *Nature Machine Intelligence*, vol. 4, no. 3, 196–210, Mar. 2022 (Cited in page 89).

[103] S. Betteti, G. Baggio, F. Bullo, and S. Zampieri, "Input-driven dynamics for robust memory retrieval in hopfield networks," *Science Advances*, vol. 11, no. 17, 2025 (Cited in pages 89, 109, 110, 112).

[104] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology*, vol. 92, no. 1, 53–78, Feb. 2001 (Cited in page 92).

[105] H. Yan, L. Zhao, L. Hu, X. Wang, E. Wang, and J. Wang, "Nonequilibrium landscape theory of neural networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, Oct. 2013 (Cited in page 93).

[106] B. A. W. Brinkman, H. Yan, A. Maffei, I. M. Park, A. Fontanini, J. Wang, and G. La Camera, "Metastable dynamics of neural circuits and networks," *Applied Physics Reviews*, vol. 9, no. 1, Mar. 2022 (Cited in page 93).

[107] F. Moss and P. V. E. McClintock, *Noise in Nonlinear Dynamical Systems*. Cambridge University Press, Apr. 1989 (Cited in page 93).

[108] E. Wong, "Stochastic neural networks," *Algorithmica*, vol. 6, no. 1–6, 466–478, Jun. 1991 (Cited in page 93).

[109] P. H. Baxendale and S. V. Lototsky, *Stochastic Differential Equations: Theory and Applications: A Volume in Honor of Professor Boris L Rozovskii*. WORLD SCIENTIFIC, Apr. 2007 (Cited in page 94).

[110] G. A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Springer New York, 2014 (Cited in page 94).

[111] R. Jordan, D. Kinderlehrer, and F. Otto, "Free energy and the fokker-planck equation," *Physica D: Nonlinear Phenomena*, vol. 107, no. 2–4, 265–271, Sep. 1997 (Cited in page 95).

[112] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the fokker–planck equation," *SIAM Journal on Mathematical Analysis*, vol. 29, no. 1, 1–17, Jan. 1998 (Cited in page 95).

[113] P. Monmarché, "Wasserstein contraction and Poincaré inequalities for elliptic diffusions with high diffusivity," *Annales Henri Lebesgue*, vol. 6, 941–973, 2023 (Cited in pages 97, 103, 106, 112, 113).

[114] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*. Springer, 2014, vol. 113 (Cited in page 98).

[115] G. A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Springer, 2014 (Cited in page 98).

[116] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *Michigan Mathematical Journal*, vol. 31, no. 2, 1984 (Cited in page 98).

[117] L. Natile, M. A. Peletier, and G. Savaré, "Contraction of general transportation costs along solutions to Fokker–Planck equations with monotone drifts," *Journal de Mathématiques Pures et Appliquées*, vol. 95, no. 1, 18–35, 2011 (Cited in page 99).

[118] B. Øksendal, *Stochastic Differential Equations: an Introduction with Applications*, 5th ed. Springer, 2013 (Cited in page 99).

[119] Q.-C. Pham, "A variation of Gronwall's lemma," *arXiv preprint*, 2007 (Cited in pages 100, 101).

[120] R. F. Bass, "On aronson's upper bounds for heat kernels," *Bulletin of the London Mathematical Society*, vol. 34, no. 04, 415–419, Jul. 2002 (Cited in page 105).

[121] V. I. Bogachev, *Measure Theory*. Springer Berlin Heidelberg, 2007 (Cited in page 105).

[122] F. Bullo, *Contraction Theory for Dynamical Systems*, 1.2. Kindle Direct Publishing, 2024 (Cited in page 110).

[123] S. Betteti and F. Bullo, "Contraction and concentration of measures with applications to theoretical neuroscience," *arXiv preprint*, 2025 (Cited in page 112).

[124] J. M. Lee, *Introduction to Smooth Manifolds*. Springer, 2003 (Cited in page 114).

[125] D. Kleinfeld, "Sequential state generation by model neural networks.," *Proceedings of the National Academy of Sciences*, vol. 83, no. 24, 9469–9473, Dec. 1986 (Cited in page 116).

[126] H. Sompolinsky and I. Kanter, "Temporal association in asymmetric neural networks," *Physical Review Letters*, vol. 57, no. 22, 2861–2864, Dec. 1986 (Cited in page 116).

[127] D. Bibitchkov, J. M. Herrmann, and T. Geisel, "Pattern storage and processing in attractor networks with short-time synaptic dynamics," *Network: Computation in Neural Systems*, vol. 13, no. 1, 115–129, Jan. 2002 (Cited in page 116).

[128] M. Abeles, *Corticonics: Neural circuits of the cerebral cortex*. Cambridge University Press, 1991 (Cited in page 116).

[129] M. Diesmann, M. O. Gewaltig, and A. Aertsen, "Stable propagation of synchronous spiking in cortical neural networks," *Nature*, vol. 402, no. 6761, 529–533, Dec. 1999 (Cited in page 116).

[130] D. Spalla, I. M. Cornacchia, and A. Treves, "Continuous attractors for dynamic memories," *eLife*, vol. 10, Sep. 2021 (Cited in page 116).

[131] K. Rajan, C. D. Harvey, and D. W. Tank, "Recurrent network models of sequence generation and memory," *Neuron*, vol. 90, no. 1, 128–142, Apr. 2016 (Cited in page 116).

[132] V. M. Ladwani and V. Ramasubramanian, "M-ary hopfield neural network based associative memory formulation: Limit-cycle based sequence storage and retrieval," in *Artificial Neural Networks and Machine Learning – ICANN 2021*. Springer International Publishing, 2021, 420–432 (Cited in page 117).

[133] M. Rabinovich, R. Huerta, and G. Laurent, "Transient dynamics for neural processing," *Science*, vol. 321, no. 5885, 48–50, Jul. 2008 (Cited in page 116).

[134] D. W. Dong and J. J. Hopfield, "Dynamic properties of neural networks with adapting synapses," *Network: Computation in Neural Systems*, vol. 3, no. 3, 267–283, Jan. 1992 (Cited in page 118).

[135] N. Caporale and Y. Dan, "Spike timing–dependent plasticity: A hebbian learning rule," *Annual Review of Neuroscience*, vol. 31, no. 1, 25–46, Jul. 2008 (Cited in page 118).

[136] L. F. Abbott and S. B. Nelson, "Synaptic plasticity: Taming the beast," *Nature Neuroscience*, vol. 3, no. S11, 1178–1183, Nov. 2000 (Cited in page 118).

[137] T. Miconi, "Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks," *eLife*, vol. 6, Feb. 2017 (Cited in page 118).

[138] G. Bellec, F. Scherr, A. Subramoney, E. Hajek, D. Salaj, R. Legenstein, and W. Maass, "A solution to the learning dilemma for recurrent networks of spiking neurons," *Nature Communications*, vol. 11, no. 1, Jul. 2020 (Cited in page 118).

[139] E. Najarro and S. Risi, "Meta-learning through hebbian plasticity in random networks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 20 719–20 731 (Cited in page 118).