SAPIENZA
UNIVERSITÀ DI ROMA

Beyond Hebbian Learning: Optimizing Memory Retrieval and Network Structure in Bio-Inspired Neural Models

Sapienza University of Rome
PhD in Physics (XXXVIII cycle)

Ludovica Serricchio
ID number 1743522

Advisor
Prof. Giancarlo Ruocco
Prof. Federico Ricci-Tersenghi

Academic Year 2024/2025

Thesis defended on 19/01/2026
in front of a Board of Examiners composed by:

Prof. Silvio Franz

Prof. Matteo Paoluzzi

Prof. Fabio Saracco

Beyond Hebbian Learning: Optimizing Memory Retrieval and Network Structure in Bio-Inspired Neural Models

Ph.D. thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: — 2025

Author's email: ludovica.serricchio@uniroma1.it

Ai miei nonni
che ci sono sempre,
anche se in altre forme

Contents

Introduction: An Extended Abstract

The exploration of human cognition, particularly understanding learning and memory, has long been a central driver of neural network research. A key insight emerging from cognitive science is that human memory is fundamentally *associative*: recalling a concept readily triggers related ideas, sensations, and experiences. Modeling this associative nature of memory has become a central element of theoretical neuroscience, inspiring generations of researchers to bridge biology and computation through simplified mathematical representations of neural activity.

The modern computational approach to this problem began with the work of McCulloch and Pitts in 1943 [94], who introduced the first formal model of a neuron as a binary threshold unit capable of performing logical operations. Their model gave the foundation for treating the brain as a computational network, where collective behavior arises from simple local interactions among neurons.

A crucial conceptual advance came a few years later with Donald Hebb's postulate in 1949 [72], which proposed a physiological mechanism for learning:

> "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."

This idea, often paraphrased as "cells that fire together, wire together", represents the first plausible rule for synaptic modification, offering a mechanistic explanation for how neural assemblies might form and store information through experience-dependent plasticity. The study of collective behavior in neural systems has profoundly influenced both physics and neuroscience over the past four decades.

Among the various theoretical frameworks that have shaped our understanding of learning and memory, the *Hopfield model* [73] occupies a central position since 1980s, when John Hopfield introduced a revolutionary reinterpretation of associative memory networks. His key insight was to view a recurrent neural network of binary neurons with symmetric connections not merely as a computational system, but as a *physical* one. By defining an energy (or Lyapunov) function that monotonically decreases under the network's dynamics, Hopfield established a direct formal analogy between neural computation and the Ising model of statistical mechanics [11, 97]. This correspondence enabled the application of powerful analytical tools, such as mean-field theory and the replica method, to the study of collective computation and memory retrieval in neural systems.

A network of binary neurons with symmetric pairwise interactions could act as an associative memory, retrieving stored information from partial or corrupted input through a process of dynamical convergence toward stable attractors. These correspond to local minima of an energy landscape defined by a Hamiltonian of the form

$$H = -\frac{1}{2}\sum_{i\neq j} J_{ij}\sigma_i\sigma_j, \tag{I.1}$$

where the Ising spins $\sigma_i = \pm 1$ represent the binary activity (firing or silent) of neuron i, and the coupling coefficients J_{ij} quantify the synaptic strengths between neurons i and j. The connections are typically constructed via the *Hebbian learning rule* [99],

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu\, \xi_j^\mu. \tag{I.2}$$

The set of stored patterns $\{\boldsymbol{\xi^\mu}\}_{\mu=1}^P$, with $\xi_i^\mu = \pm 1$ drawn with equal probability, define the memory states of the network. Each pattern corresponds to a local minimum of the energy landscape and acts as an attractor of the dynamics. Starting from an arbitrary initial condition $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^N$, the system evolves deterministically or stochastically toward one of these stored configurations effectively performing error correction and memory reconstruction.
The relationship between the number of stored patterns P and the number of neurons N is quantified by the *load parameter*

$$\alpha = \lim_{N\to\infty} \frac{P}{N}. \tag{I.3}$$

This parameter distinguishes two operational regimes: (i) a *low load* regime $\alpha \ll 1$ where retrieval is robust and interference among patterns is negligible; and (ii) a *high-load* regime $\alpha = \mathcal{O}(1)$ where interference and noise dominate.
Despite conceptual clarity and analytical tractability [13, 14, 15], this model also revealed an intrinsic limitation common to many recurrent systems: when the storage load exceeds the critical threshold $\alpha_c \simeq 0.138$ [13], the system undergoes a transition to a spin-glass phase dominated by an exponentially large number of metastable, spurious states. In this regime, genuine attractors states lose stability and the network ceases to retrieve stored memories.

Structure and Objectives of thesis

Understanding and overcoming this fundamental constraint has driven decades of research aimed at extending from one side the storage capacity of Hopfield networks and on the other side the biological realism of associative memory models. Indeed, neural networks have recently achieved remarkable success in various machine learning tasks. However, the underlying reasons for their effectiveness, especially in comparison to biological neural networks, remain unclear.
In this thesis we introduce and explain a dual-strategy toward this goal, exploring two complementary research directions to enhance the performance and biological plausibility of associative memory models. The work is consequently divided into two parts:

- In **Part I** we focus on improving the classical, symmetric Hopfield model through a new algorithmic approach: the Daydreaming algorithm to optimize the synaptic coupling matrix and push performance beyond the Hebbian limit.
- In **Part II** we abandon the symmetric, fully-connected Hebbian paradigm altogether. We investigate how structural properties, specifically, asymmetry and dilution in the coupling matrix, fundamentally reshape the dynamical landscape of neural networks, moving towards more biologically realistic architectures.

Part I - Symmetric and Fully Connected Coupling Matrix

The first part of this thesis is dedicated to enhancing the storage capacity of the Hopfield model through a new algorithmic perspective, while maintaining the classical assumption of symmetric, fully connected synaptic matrices. This restriction guarantees the existence of a global energy function (Hamiltonian), allowing the network dynamics to be analyzed using equilibrium statistical mechanics.

Chapter 1: The Hopfield Model

In this first Chapter, we revisit the Hopfield model in detail, from its historical biological and physical motivations to its complete analytical formulation and phase diagram. We begin by recalling the model's biological inspiration, McCulloch-Pitts neurons with binary states, synaptic plasticity governed by Hebbian learning and the interpretation of memory retrieval as convergence to attractors in a high-dimensional energy landscape. The dynamics are formulated either deterministically (zero temperature $T = 0$) or stochastically (finite temperature $T > 0$) using Glauber single-spin updates, which respect detailed balance and converge to the Gibbs-Boltzmann distribution when couplings are symmetric.
The core of the chapter is represented by thermodynamic formulation of the model. The Hebbian coupling matrix encodes P random binary patterns as local minima of the energy landscape. We presents the derivation of the mean-field equations, the analysis of the energy landscape, and the construction of the full phase diagram in the $(T - \alpha)$ plane.
Both low load and high-load regimes are discussed, including the emergence of retrieval, paramagnetic, and spin-glass phases. In the vanishing load regime $\alpha \to 0$, we derive the mean-field free energy and self-consistent equations for the overlap order parameters, identifying pure retrieval states as stable attractors below a critical temperature $T_c = 1$. In the high-load regime $\alpha > 0$, we employ the replica method under the replica-symmetric (RS) ansatz to compute the free energy, underlining the emergence of spin-glass order parameters and the transition to a phase dominated by spurious states. We conclude the chapter with the construction of the full (T, α) phase diagram, which delineates four distinct regimes:

- **Paramagnetic phase** $T > T_g(\alpha)$: thermal fluctuations dominate, no retrieval.
- **Spin-glass phase** $T_M(\alpha) < T < T_g(\alpha)$: frozen disorder with no pattern correlation.

- **Metastable retrieval phase** $T_c(\alpha) < T < T_M(\alpha)$: retrieval states coexist with spin-glass states but remain only locally stable.
- **Ferromagnetic retrieval phase** $T < T_c(\alpha)$: pure states are global minima with well-defined basins of attraction.

A signal-to-noise analysis at $T = 0$ establishes the critical capacity $\alpha_c \approx 0.138$, beyond which retrieval becomes impossible due to interference noise from stored patterns. We also discuss the onset of replica symmetry breaking (RSB) at very low temperatures, which slightly increases the effective capacity to $\alpha_c^{\text{RSB}} \approx 0.144$ [35] but does not fundamentally resolve the storage bottleneck.

Take-home message: The Hopfield model, despite its analytical tractability, suffers from a fundamental capacity limit caused by the interference among stored patterns and the proliferation of spurious attractors in the high-load regime. This limitation motivates the search for learning rules that can stabilize memories while suppressing spurious states.

Chapter 2: Extending the Hopfield Model - The Role of Dreaming

During the last two decades of the previous century, significant efforts in statistical physics were directed towards the development of neural network models capable of memory storage and retrieval [11, 124]. These investigations have greatly improved our understanding of how memory retrieval processes can be modeled within network structures, bridging disciplinary boundaries and influencing diverse fields, including theoretical neuroscience and biology.
A central question in this research concerns the *storage capacity* of recurrent neural networks (RNNs), particularly in the context of Hopfield networks. Since its original formulation, several extensions have been proposed to improve its capacity, stability, and biological realism.
In this Chapter we review historical and contemporary approaches implemented to improve the storage capacity of Hopfield-like networks, with particular emphasis on *unlearning* (or *dreaming*) mechanisms [34, 111, 74, 140]. These processes modify the synaptic matrix by selectively weakening or removing spurious correlations accumulated during learning, effectively reorganizing the energy landscape.
The term *dreaming* was later adopted to emphasize the analogy with biological sleep, during which the brain is believed to reorganize and consolidate learned information [34]. Indeed, inspired by the hypothesis that REM sleep consolidates memories by selectively weakening spurious synaptic correlations, unlearning procedures modify the coupling matrix iteratively to destabilize spurious attractors while reinforcing genuine memory states smoothing the energy landscape around the true attractors, thereby stabilizing the retrieval of stored memories and increasing the maximal load α achievable by the system.
However, most existing dreaming procedures face significant challenges. Excessive repetition of unlearning steps leads to the deterioration of desired memories, resulting in *catastrophic forgetting*, where even genuine memories are degraded and therefore no pattern can be retrieved unless the number of dreaming iterations is carefully finely tuned [141]. A local dreaming procedure that can be iterated

indefinitely was later proposed in [92, 111], allowing for continuous learning of new examples at the cost of progressively forgetting earlier ones. Beyond the solution proposed by Parisi [111] the black-out catastrophe can also be addressed within the solvable framework introduced in [100] which provides a complementary perspective on the problem. Alternative strategies have sought to modify the nature of the synaptic interactions, for example by introducing multi-spin terms [58, 84, 2], which can raise the capacity beyond the linear regime and, in some cases, even exponentially [90]. More recent advances include dreaming algorithms inspired by the perceptron rule [53, 25, 24] achieve near-optimal results but still require careful fine-tuning of the number of unlearning iterations.
Furthermore, standard dreaming procedures struggle to efficiently handle correlated examples [142], a limitation that restricts their applicability to realistic datasets.
By describing the main characteristics and limitations of existing learning rule aimed to improve the storage capacity of the Hopfield Model, we set theoretical foundation for the introduction of the *Daydreaming algorithm* [127], a local learning rule which combines the robustness of Hebbian learning with the capacity-enhancing effects of continuous unlearning. Unlike previous methods, Daydreaming applies learning and unlearning simultaneously in each update step, avoiding the early-stopping dilemma and enabling indefinite iteration without catastrophic forgetting.

Take-home message: Unlearning represents an efficient mechanism to reorganize synaptic weights and improve retrieval robustness. However, most existing implementations face critical limitations: fine-tuning requirements, catastrophic forgetting, or inability to handle correlated data. The Daydreaming algorithm is designed to overcome these limitations by integrating learning and unlearning into a single, locally implementable update rule.

Chapter 3: Daydreaming

Within this framework, we develop and analyze the performances of the Daydreaming algorithm that overcomes the catastrophic capacity limit of the standard Hebbian learning, extending the retrieval phase of the original model at $T = 0$. Unlike classical Hebbian learning, which encodes memories once and then remains static, Daydreaming continuously refines the synaptic coupling matrix by simultaneously reinforcing stored patterns and suppressing spurious attractors. This dual-process mechanism operates without supervision and requires minimal parameter tuning. The key advantage of Daydreaming is its ability to extend the retrieval phase beyond the classical Hopfield capacity limit $\alpha_c \simeq 0.138$ up to theoretical maximum $\alpha = 1$ for uncorrelated patterns, while also exploiting correlations in structured data to arrive at a even greater effective capacity. We demonstrate these capabilities through systematic numerical experiments on synthetic datasets and real-world data (MNIST), showing that the algorithm not only stores memories efficiently but also spontaneously extracts latent structure and generalizes to unseen examples.
In this Chapter we introduce, analyze, and test the Daydreaming, the central contribution of this first part of thesis. The rule iteratively modifies the synaptic couplings to improve the stability of learned patterns while destabilizing spurious minima and is defined by a simple iterative update rule applied to the synaptic

matrix $\mathbf{J}$:

$$J_{ij}^{(u+1)} = J_{ij}^{(u)} + \frac{1}{\tau N}(\xi_i^{\mu(u)}\xi_j^{\mu(u)} - \tilde{\sigma}_i^{(u)}\tilde{\sigma}_j^{(u)}). \tag{I.4}$$

At each step u, a random stored pattern $\boldsymbol{\xi}^{\boldsymbol{\mu(u)}}$ is selected and reinforced (Hebbian term), while the network is initialized from a random configuration and evolved to a fixed point $\tilde{\boldsymbol{\sigma}}^{\boldsymbol{(u)}}$, which is then suppressed (anti-Hebbian term).
The parameter τ controls the timescale of learning, but remarkably, the algorithm's performance is largely independent of τ for sufficiently large values, making it effectively parameter-free.
We begin by establishing convergence properties. Through systematic analysis of the norm of updates and the distance from the initial Hebbian matrix, we demonstrate that the learning dynamics reach a stationary statistical ensemble after a time proportional to τ. Importantly, convergence is assessed not by fixed-point behavior of $\mathbf{J}$ itself (which continues to fluctuate due to stochasticity), but by the stationarity of retrieval maps, curves plotting final overlap m_{F} versus initial overlap m_{I}, which quantify the effective size and stability of attractor basins.
As a primary validation of the algorithm's capacity-enhancing capabilities, we analyze its performances on uncorrelated data. We demonstrate that:

- For loads $\alpha < \alpha_c \approx 0.138$ (below Hopfield capacity), Daydreaming enlarges basins of attraction compared to standard Hebbian learning.
- For loads $\alpha_c < \alpha < 1$, Daydreaming stabilizes stored patterns that would otherwise be unstable under Hebbian learning, extending the retrieval plateau in the $(m_{\mathrm{I}}, m_{\mathrm{F}})$ plane.
- At theoretical limit $\alpha \to 1$, patterns become marginally stable so we achieve the maximum possible capacity for uncorrelated data in the symmetric coupling matrix scenario.

This distinction between low $\alpha < \alpha_c$ and high $\alpha_c < \alpha < 1$ load regimes is presented here primarily to establish a clear comparison with the original Hopfield model and its well-known capacity limit. The division into low- and high-load regimes in this case serves just to quantify the algorithm's advancement. We emphasize that this separation does not reflect a qualitative change in the Daydreaming algorithm's operation, but rather quantifies its advancement across different load conditions. Within the classical retrieval phase $\alpha < \alpha_c$, Daydreaming enlarges existing basins of attraction. Beyond it $\alpha > \alpha_c$, the same mechanism creates stable attractors where the Hebbian rule fails, breaking the original model's theoretical capacity barrier.
We also analyze the energy landscape evolution during training, showing that the mean energy of stored patterns decreases relative to spurious states as training progresses, with the energy gap vanishing precisely at $\alpha = 1$.
Comparison with the Storkey [131, 132] and Pseudo-Inverse rules [115, 80] reveals that Daydreaming produces larger basins of attraction across the entire range $\alpha \in [0, 1]$, and unlike the pseudo-inverse rule, remains numerically stable at high loads.
Performance on correlated data demonstrates the algorithm's remarkable ability to exploit structure. Using the random-features model, where P patterns are generated as random linear combinations of $D < N$ hidden features, we show that:

- Correlated patterns develop larger basins of attraction than uncorrelated ones at the same load α.
- Retrieval remains robust even for $\alpha > 1$, surpassing theoretical limit for uncorrelated patterns.
- The algorithm spontaneously stabilizes the hidden features themselves as attractors, without any explicit supervision or knowledge of the feature structure.

Finally, we see the application to MNIST to validate the algorithm on real-world data. In the low load regime $\alpha < 0.8$, the network functions as a precise associative memory, storing individual training images as stable attractors. Supervised Daydreaming, training on class prototypes rather than individual images, obtains a test accuracy of approximately 67.5% on digit classification, surpassing previous Hopfield-based approaches. In the high-load regime $\alpha \gtrsim 50$, a qualitative transition occurs: training examples lose individual stability, but the network develops emergent attractors that correspond to class prototypes. Dynamics initialized from either training or test images converge to these generalized representations, demonstrating a form of unsupervised abstraction and generalization.

Take-home message: The Daydreaming algorithm achieves near-optimal memory storage and robust retrieval using only local update rules. The presented learning rule autonomously stabilizes stored memories, enlarges their basins of attraction, and eliminates spurious minima without requiring external supervision or fine-tuning. It extends the retrieval phase well beyond the standard Hopfield limit to theoretical maximum capacity for uncorrelated patterns $\alpha = 1$. Moreover, it autonomously extracts latent features and prototypes from real data, demonstrating its intrinsic ability to generalize and exploit correlations in structured data without supervision.

Part II – Asymmetric and Diluted Coupling Matrix

In the second part of this thesis we shift focus from optimizing learning rules within symmetric architectures to a more general exploration of the role of network structure itself in shaping dynamical behavior.
Biological neural systems are neither fully connected nor symmetric in their connectivity. Synaptic connections are sparse (*diluted*), often unidirectional ($J_{ij} \neq 0$ does not imply $J_{ji} \neq 0$), and exhibit varying degrees of correlation between forward and backward pathways. These structural features break the symmetry assumptions underlying the Hopfield model, eliminating the global energy function and introducing non-equilibrium dynamics with potentially richer attractor landscapes. They deeply influence the dynamical configuration space that is now characterized by fixed points and limit cycles.
Most prior works have studied dilution and asymmetry separately and not considering the topological asymmetric of the synaptic matrix. To address this gap, we introduce a generalized framework of the network structure parameterized by three independent control parameters:

- **Connection density** ρ: the fraction of existing synaptic connections, ranging from highly diluted $\rho \approx 0$ to fully connected $\rho = 1$;

- **Synaptic correlation** η: the degree of correlation between reciprocal couplings (J_{ij}, J_{ji}) when both exist, interpolating between perfect correlation ($\eta = 1$, symmetric), independence ($\eta = 0$, asymmetric), and anti-correlation ($\eta = -1$, antisymmetric);
- **Topological asymmetry** a_1: the probability that a connection exists in only one direction, quantifying the prevalence of directed (unidirectional) links independent of coupling symmetry.

This parameterization defines a three-dimensional phase diagram, visualized as a triangular prism in (ρ, a_1, η) space, that contains all possible network architectures, from the classical symmetric Hopfield model to fully asymmetric sparse networks resembling biological connectivity. Using this parametrization we systematically explore the full space of network architectures, distinguishing between the effects of diluting connections, breaking coupling symmetry, and introducing directed (unidirectional) links.
Through exhaustive numerical simulations of finite systems, we map how these parameters influence the long-term network behavior: in particular, we analyze the number, length, stability, and basin sizes of dynamical attractors. After this exploration, we identify optimal connectivity regimes that balance robust configuration space partitioning with biologically realistic constraints.

Chapter 4: Extending Neural Networks Toward Biology

In this Chapter we motivate the transition from symmetric, fully connected networks to biologically realistic architectures. We discuss the biological implausibility of the Hopfield model's assumptions: full connectivity contradicts the sparse wiring observed in cortical and hippocampal circuits (where connection probabilities are typically $\sim$ 4–10%), while perfect symmetry is inconsistent with the directed nature of most synaptic structures.
Relaxing these constraints has profound theoretical consequences: the dynamical models that are no longer Hamiltonian and detailed balance is violated. Attractors may include not only fixed points but also limit cycles of arbitrary length, and the distribution of basin sizes becomes highly non-trivial.
Experimental studies show that synaptic connectivity in both neocortex and hippocampus is highly structured rather than random. In neocortical circuits, reciprocal connections and small connectivity motifs occur more frequently than expected [129], while in hippocampal CA3 networks recurrent connections are sparse but organized into motifs such as reciprocal ones [69]. From a theoretical perspective, Brunel [29] showed that optimizing the size of basins of attraction under synaptic sign constraints naturally leads to sparse connectivity and an intermediate degree of symmetry. These findings motivate the inclusion of structured and heterogeneous couplings in theoretical models of neural networks.
We review prior work on asymmetric dense networks (which revealed transitions from ordered to chaotic dynamics as a function of coupling correlation) [106, 20, 21, 38, 68, 76] and on symmetric diluted networks which demonstrated an increasing number of attractors under specific sparsity regimes [54, 87]. However, the joint interplay between dilution, coupling asymmetry, and topological asymmetry remained

unexplored. In this chapter we set the stage for a systematic investigation of this three-dimensional parameter space, with the goal of identifying optimal regimes that maximize attractor diversity and biological plausibility.

Take-home message: Biological realism requires relaxing symmetry and full connectivity and introduces richer attractor structures. A comprehensive framework is needed to decouple the effects of density, coupling correlation, and topological structure, a challenge we address through exhaustive numerical simulation across the full parameter space.

Chapter 5: Dynamics in Asymmetric Diluted Spin Systems

In this Chapter we expose theoretical and computational framework for our investigation. We define a network of N binary spins $\sigma_i = \pm 1$ evolving under deterministic, synchronous (parallel) update dynamics:

$$\sigma_i(t+1) = \mathrm{sgn}\left(\sum_j J_{ij}\sigma_j(t)\right). \tag{I.5}$$

The coupling matrix $\mathbf{J}$ is constructed systematically to exhibit controlled levels of density ρ, synaptic correlation η, and topological asymmetry a_1. Unlike previous approaches that generate $\mathbf{J}$ as a linear combination of symmetric and antisymmetric matrices (thereby implicitly enforcing $a_1 = 0$), our construction independently controls all three parameters, allowing exploration of the full architectural space. We introduce the key observables used to characterize the attractor landscape:

- Number of attractors C;
- Cycle lengths L_i;
- Basin sizes Ω_i;
- Mean distances MD_i;
- Effective quantities $N_{\text{eff}}, L_{\text{eff}}$;
- Chaoticity parameter p_{chaos}.

Our computational strategy is based on exhaustive enumeration: for each realization of $\mathbf{J}$ and each system size N, we simulate the deterministic dynamics from all 2^N initial configurations, allowing a complete characterization of the attractor landscape. This brute-force approach, while computationally expensive (limiting us to $N \lesssim 26$), gives a rigorous statistical characterization through simple sampling-based methods.

Take-home message: Asymmetric diluted networks exhibit fundamentally different and richer dynamical properties than their symmetric counterparts, with attractor landscapes shaped by the independent and often non-trivial interplay of density, coupling correlation, and topological structure.

Chapter 6: Numerical Results in Asymmetric Diluted Networks

In this Chapter we present the core of the numerical findings of Part II and it is organized as a detailed statistical analysis of the exploration of the (ρ, η, a_1) parameter space. Section 6.1.1 validates our methodology by reproducing known results in previously studied regimes:

- Fully connected, fully asymmetric networks ($\rho = 1$, $\eta = 0$, $a_1 = 0$): recovery of the classical results of [21, 106, 68] linear growth of attractor count $C \sim 0.35N$, exponential growth of cycle lengths $L \sim e^{0.21N}$, and heterogeneous basin structure.
- Fully connected networks with varying correlation ($\rho = 1$, $\eta \in [0, 1]$): a dynamical transition at $\eta_d \approx 0.33$ separates chaotic and ordered regimes.
- Fully asymmetric networks with varying dilution ($\eta = 0$, $\rho \in (0, 1]$): increasing dilution suppresses chaotic dynamics; for $\rho \lesssim 0.2$, cycle lengths stabilize at $L \sim 4$, independent of system size.

By systematically varying ρ, a_1, and η, the study reveals how structural parameters control the crossover between ordered and disordered dynamics. A central finding is the emergence of an *optimal regime* of partial asymmetry and moderate connectivity, where retrieval remains efficient even in the absence of a Hamiltonian description. Section 6.2 presents the most significant results, focusing on networks without directed links $a_1 = 0$. We identify distinct dynamical regions: chaotic, ordered, intermediate, and highly diluted regimes.
The high-dilution regime $\rho \lesssim 0.2$ is particularly striking: it exhibits exponential attractor proliferation, short stable cycles $L \sim 4$, uniform basin partitioning $\Omega_i \approx 2^N/C$, and rapid convergence $MD = \mathcal{O}(1)$. Rényi entropy analysis, Sec. 6.2.1 confirms near-perfect equipartition among attractors.
Finally, Section 6.3 extends the analysis to the full three-dimensional space, varying topological asymmetry a_1. Even with moderate a_1, the high-dilution regime remains robust: attractor diversity and uniformity persist, demonstrating that dilution, rather than symmetry, is the dominant structural determinant of ordered dynamics and exponential growth of the number of attractors.

Take-home message: The attractor landscape of asymmetric diluted networks is governed by a complex interplay between connection density, synaptic correlation, and topological asymmetry, but exhibits universal features in extreme regimes. There exists an optimal region in parameter space even in non-Hamiltonian systems, suggesting that biological networks may operate near this balance. High dilution $\rho \lesssim 0.2$ emerges as an optimal connectivity regime combining:

- Maximal attractor diversity (exponential proliferation);
- Computational efficiency (short cycles, rapid convergence);
- Robustness to synaptic variability (weak η-dependence);
- Biological plausibility (sparsity consistent with cortical and hippocampal circuits).

Concluding Remarks

This thesis tries to create a bridge between statistical mechanics models that can be solved analytically since are modeled through the existence of an energy function and more realistic networks that are build considering asymmetric and diluted synaptic couplings. The main objective of this thesis is to overcome the Hebbian prescription:

- In Part I with the implementation of the *Daydreaming algorithm*, we demonstrate that adaptive local dynamics can push the storage capacity of symmetric networks, and even beyond, their theoretical limits.
- In Part II the exploration of asymmetric and diluted couplings reveals how structural heterogeneity and directed interactions shape the global organization of attractors in bio-inspired neural systems that are non-Hamiltonian.

Together, these results give a unified perspective: efficient memory and learning emerge from the interplay between local plasticity rules and global structural constraints with implications for both theoretical understanding and practical design of recurrent neural architectures.

In this Introduction, we have collected a brief overview of the content and objectives of this work; however, a more extended synthesis, highlighting the key take-home messages, can be found at the end of each Chapter where we collect and discuss the main results obtained and their significance. The purpose of this executive summary is therefore to serve as a roadmap, to which the reader may return during the reading of thesis when needed.

Part I

Symmetric and Fully Connected Coupling Matrix

Chapter 1

The Hopfield Model

«Speech has allowed the communication of ideas

– *"Talkin' Hawkin"*, S. Hawking. *The Endless River*, Pink Floyd.

This Chapter is dedicated to the description of the Hopfield Model from its historical biological derivation to the complete analytical resolution. The Hopfield model represents a foundational framework for understanding associative memory in neural networks. We begin tracing its lineage from foundational ideas in neuroscience and cognitive science to its formalization in statistical mechanics. We then define the model's formal structure, derive its fundamental equations of state, and analyze its equilibrium properties. We conclude by providing a complete description of the model's phase diagram in the temperature-load ($T-\alpha$) plane, distinguishing between the low load (memory) and high-load (spin-glass) regimes where the analysis is done using the mean field theory of spin glasses involving the Replica Method considering replica symmetric (RS) solutions.

1.1 Biological Motivation and Model Definition

The Hopfield model [73, 11, 102, 33] represents a fundamental contribution to theoretical neuroscience, providing a tractable statistical mechanical description of associative, or content-addressable, memory. It is an associative memory model used to describe the behavior of neurons as a neural network. We can view it as a thermodynamic extension of the earlier McCulloch-Pitts formal neuron [94], which translates its binary logic into a stochastic dynamics that operates over an ensemble of states.

The model is characterized by a collection of N units mimicking binary neurons which interact pairwise through weighted links and was made for the realization of the basic computational functions of a network of neurons. We schematize a neural network as a system of N interconnected binary units, $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{N}$, where each $\sigma_i \in \{-1, +1\}$ represents the state of a single neuron. A value of $+1$ corresponds to a *firing* neuron, while -1 signifies a *quiescent* state. These units interact via symmetric synaptic couplings J_{ij}, which encode the strength and sign (excitatory or inhibitory) of the connection from neuron j to neuron i, with the condition of no self-interactions: $J_{ii} = 0$ for all i. In its original formulation [73], neurons were

described by the states $S_i = 0$ (*not firing*) and $S_i = 1$ (*firing at maximum rate*), following McCulloch and Pitts. Each neuron receives inputs from the neighboring ones through synaptic junctions of strength J_{ij}, which determine the contribution of a signal fired by the jth neuron to the post synaptic potential which acts on the ith neuron. The synaptic efficacy J_{ij} can be both positive and negative distinguish between an excitatory and inhibitory synapse; non connected neurons have obviously $J_{ij} = 0$. The dynamic evolution of these states, in the phase space of 2^N states, is determined by the nature of these interactions among the neurons. The sum of the input signal to the ith neuron from the other $N-1$ is defined as synaptic potential V_i that can be written as:

$$V_i = \sum_{j \neq i}^{N} J_{ij} S_j. \tag{1.1}$$

If the weighted sum of the input signals exceeds a threshold, the neuron starts to produce an electrical pulse, defined *spike* that is transmitted through an axon to other neurons. For each neuron i there is a fixed threshold potential U_i that the neuron has to overcome in order to become excited: in the absence of noise, or external perturbation, each neuron fires a signal if its potential exceeds the threshold. This means that the state of a neuron updates based on whether the weighted sum of its inputs, the post-synaptic potential V_i, exceeds this threshold potential U_i and the neuron readjusts its state randomly in time setting [73]:

$$S_i = 1 \ \text{ if } \ \sum_{j \neq i}^{N} J_{ij} S_j > U_i, \qquad S_i = 0 \ \text{ if } \ \sum_{j \neq i}^{N} J_{ij} S_j < U_i. \tag{1.2}$$

The stable states of the network will be those configurations in which each of the spin variables S_i is aligned with its molecular field $h_i = V_i - U_i : S_i h_i = S_i(V_i - U_i) > 0$. This relation represents the dynamic rule that describes the time update of the neural configuration which is given by the McCulloch-Pitts neuron [94]:

$$S_i(t + \Delta t) = \theta(V_i(t + \Delta t) - U_i) \tag{1.3}$$

with $\theta(x)$ the Heaviside function and

$$V_i(t + \Delta t) = \sum_{j \neq i}^{N} J_{ij} S_j(t). \tag{1.4}$$

Conventionally the time interval Δt is chosen as $\Delta t = 1$. It is possible to notice a powerful correlation between this natural system with the Ising model in sight of this, a Hopfield network can be model by N Ising spins $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{N}$ that can assume only two states: $\sigma_i = 1$ if the neuron is excited and can transmit a signal and $\sigma_i = -1$ if it is at rest [113]. These two conditions are related to the case in which the neuron fires or not an action potential at time t.
In this new prescription the configurations obtained by setting: $\sigma_i = 2S_i - 1$. It is possible to rewrite the previous Eqs. (1.3)-(1.4) respectively as:

$$\sigma_i(t + 1) = \text{sgn}(V_i(t + 1) - U_i) \tag{1.5}$$

and

$$V_i(t+1) = \frac{1}{2}\sum_{j\neq i}^{N} J_{ij}(\sigma_j(t)+1). \tag{1.6}$$

The Eq. (1.5) is achieved considering that $\mathrm{sgn}(x) = 2\theta(x) - 1$. The threshold potential U_i can be chosen equal to $\sum_j J_{ij}$ so that the Eq. (1.5) becomes[1]:

$$\sigma_i(t+1) = \mathrm{sgn}\left(\sum_{j\neq i}^{N} J_{ij}\sigma_j(t)\right). \tag{1.7}$$

This equations represents the local update rule which governs the dynamical evolution of the network. At a given time t, we define the total input, or post-synaptic potential, received by neuron i as the sum of signals from all other neurons, weighted by their respective synaptic efficacies:

$$h_i(t) = \sum_{j\neq i}^{N} J_{ij}\,\sigma_j(t). \tag{1.8}$$

The state of neuron i at the next time step is then determined by comparing this potential to a threshold U_i and a common and simplifying choice is to set the threshold to zero ($U_i = 0$) for all neurons, leading us to the canonical update rule in Eq. (1.7).

1.1.1 Stochastic dynamics

A natural generalization of this model to a system with noise is to adopt Glauber single-spin dynamics at a finite temperature T [114, 63]. This introduces stochasticity into the time evolution of neurons, reflecting the noisy nature of real neural systems since the evolution of real neurons is not deterministic. In this framework, there is a finite probability for the network to occupy configurations beyond those strictly satisfying $\sigma_i h_i > 0$. The effect of noise is controlled by an effective temperature T, which quantifies the level of thermal fluctuations. The case of stochastic binary neurons is formally approached in [33] by redefining the threshold potential U_i adding a random component:

$$U_i(t) = U_i - \frac{1}{2}Tz_i \tag{1.9}$$

with $T = \frac{1}{\beta} \in \mathbb{R}^+$ and z_i a with a symmetric probability distribution function $P(z)$ (i.e., $P(z) = P(-z)$), zero mean $\mathbb{E}[z_i] = 0$, and unit variance $\mathrm{Var}(z_i) = 1$. This variable models the random fluctuations inherent in the biological neuron activation process. Substituting this into the deterministic update rule (and with the standard choice $U_i = \frac{1}{2}\sum_j J_{ij}$), the stochastic evolution law becomes:

$$\sigma_i(t+1) = \mathrm{sgn}\left(h_i(t) + Tz_i(t)\right). \tag{1.10}$$

The previous equation is obtained considering that $\mathrm{sgn}(\frac{x}{2}) = \mathrm{sgn}(x)$. The parameter T measures the amount of noise in the system: for $T = 0$, we recover the deterministic rule (Eq. 1.7), while for $T \to \infty$ the system evolves in a completely random manner.

[1] In this calculation the factor $\frac{1}{2}$ is reabsorbed in definition of the couplings: $J_{ij} = \frac{1}{2}J_{ij}$ [11].

The probability of finding the neuronal state $\sigma_i(t+1)$ is expressed in terms of the distribution $P(z)$ of the noise variables z_i:

$$\mathbb{P}\left[\sigma_i(t+1) = -1\right] = \mathbb{P}[h_i(t) + Tz_i < 0] = \mathbb{P}\left[z_i < -\frac{h_i(t)}{T}\right] = \int_{-\infty}^{-h_i(t)/T} P(z)dz \tag{1.11}$$

$$\mathbb{P}\left[\sigma_i(t+1) = 1\right] = 1 - \mathbb{P}\left[\sigma_i(t+1) = -1\right] = \int_{-h_i(t)/T}^{\infty} P(z)\,dz. \tag{1.12}$$

For symmetric noise distributions $P(z) = P(-z)$:

$$\mathbb{P}[\sigma_i(t+1) = \pm 1] = \mathbb{P}[h_i(t) + Tz_i(t) \gtrless 0] = \int_{-\infty}^{\sigma_i(t+1)h_i(t)/T} P(z)dz. \tag{1.13}$$

This expression can be written compactly by introducing the cumulative distribution function $F(x) = \int_{-\infty}^{x} P(z)dz$:

$$\mathbb{P}[\sigma_i(t+1)] = F\left(\beta\sigma_i(t+1)h_i(t)\right) \tag{1.14}$$

where we have used $\beta = 1/T$ and that represents the probability that $\sigma_i(t+1) = \pm 1$. The standard choice for neural networks, which guarantees convergence to the Gibbs-Boltzmann distribution, is to let the noise be governed by:

$$F(x) = \frac{1}{2}\left[1 + \tanh(x)\right]. \tag{1.15}$$

Considering that $\tanh(\sigma_i x) = \sigma_i \tanh(x)$ if $\sigma_i \in \{-1, 1\}$ we end up with the following equation:

$$\mathbb{P}[\sigma_i(t+1)] = \frac{e^{\beta h_i(t)\sigma_i(t+1)}}{e^{\beta h_i(t)} + e^{-\beta h_i(t)}}. \tag{1.16}$$

Thus, the neuron updates to $\sigma_i(t+1) = +1$ with probability $p = 1/(1+e^{-2\beta h_i(t)})$ and to $\sigma_i(t+1) = -1$ with probability $1-p$. In the zero-temperature limit $\beta \to \infty$, this probability becomes a step function, recovering the deterministic dynamics, while at infinite temperature $\beta \to 0$, $p \to 1/2$ and updates become completely random.
There are two ways of updating the variables:

- **Synchronous (Parallel) Dynamics:** All neurons are updated simultaneously using the fields $h_i(t)$ calculated from the entire configuration at time t. Neurons are independent therefore the probability of finding the state $\boldsymbol{\sigma}$ in an updated configuration is the product of the single probabilities. So the probability for a full network update is:

$$\mathbb{P}\left(\boldsymbol{\sigma}(t+1)|\boldsymbol{\sigma}(t)\right) = \prod_{i=1}^{N} \frac{e^{\beta\sigma_i(t+1)h_i(t)}}{2\cosh(\beta h_i(t))}. \tag{1.17}$$

- **Asynchronous (Sequential) Dynamics:** Neurons are updated one at a time, typically in a random sequence. A single neuron i is selected and its state is updated immediately changing the configuration used to calculate subsequent fields.

Asynchronous dynamics generally ensure detailed balance and convergence to the equilibrium Gibbs-Boltzmann distribution [88]:

$$P_{\text{eq}}(\boldsymbol{\sigma}) = \frac{1}{Z} e^{-\beta \mathcal{H}(\boldsymbol{\sigma})}, \tag{1.18}$$

where Z is the partition function and $\mathcal{H}(\boldsymbol{\sigma})$ is the Hamiltonian of model that will be presented in the following Section 1.2.

1.2 Thermodynamic Formulation

A profound connection to statistical mechanics emerges when the stable states of the network are identified with the local minima of an appropriate energy function, or Hamiltonian. For a network with symmetric couplings ($J_{ij} = J_{ji}$), this Hamiltonian is defined as:

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j = -\frac{1}{2}\sum_{i\neq j} J_{ij}\sigma_i\sigma_j. \tag{1.19}$$

The condition for a dynamical fixed point, where $\sigma_i = \text{sgn}(h_i)$ and thus $\sigma_i h_i > 0$ for all i, is precisely equivalent to the condition that $\boldsymbol{\sigma}$ is a local minimum of $\mathcal{H}(\boldsymbol{\sigma})$ with respect to single-spin flips since it's means that each neuron is aligned with its local field and no further updates occur. This mapping allows us to analyze the network's properties using the full formalism of equilibrium statistical mechanics. The attractors of the dynamics correspond to metastable states in the energy landscape defined by $\mathcal{H}(\boldsymbol{\sigma})$. Consequently, we can study the system at a finite temperature T by employing the Gibbs-Boltzmann distribution.
This formulation recasts the neural network as an interacting magnetic system, directly analogous to the Ising model. Within this analogy, the synaptic efficacies J_{ij} play the role of exchange interactions, and neurons correspond to spin configurations. The crucial distinction lies in the nature of the couplings J_{ij}; rather than being short-ranged and uniform (e.g., ferromagnetic), they are long-ranged and purposefully designed to store specific memories. The Hamiltonian $\mathcal{H}(\boldsymbol{\sigma})$ acts as a cost function, whose minima $\boldsymbol{\sigma}^*$ represent the desired output. The goal of the network dynamics is to evolve from an initial configuration $\boldsymbol{\sigma}$ towards one of these low-energy attractor states.

1.2.1 The Hebbian Learning Rule and Stored Patterns

The term associative memory, or content-addressable memory, refers to the capability of a neural system to retrieve a complete, stored memory upon receiving a partial or corrupted version presented as input [11, 33]. Starting from an initial configuration that represents a corrupted input, the network dynamics evolve towards a stable fixed point, an attractor state, that corresponds to a stored memory.
The dynamics of the network prescribes a trajectory in the space of the 2^N possible states of the system. It is, therefore, natural to identify retrieval as the convergence to, and persistence in, one of these attractor states. The ability of the system to perform as an associative memory is related to the concept of ergodicity and, therefore, to the dependence of trajectories on the initial states. If the network's

trajectories were independent of the initial state (i.e., ergodic), no pattern-specific retrieval could occur. Instead, if trajectories in the space of network states depend on the initial states then the model can recall. The existence of distinct basins of attraction—regions in state space that flow into specific attractors—allows for the classification of inputs and successful recall. These attractors coincide with the local minima of the system's Hamiltonian $\mathcal{H}(\boldsymbol{\sigma})$, they are defined *pattern* and labelled with the symbol $\boldsymbol{\xi}$. This means that the neural network can be used to process information by the creation of these attractors in the space $\{-1, 1\}^N$ of network states, through modification of the interactions J_{ij}.
To function as an associative memory, the synaptic matrix $\boldsymbol{J}$ must be constructed to store a set of P predefined memory patterns, denoted $\boldsymbol{\xi} = \{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$ with $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \dots, \xi_N^\mu)$ for $\mu = 1, \dots, P$. Each ξ_i^μ is a Rademacher variable iid (identical independent distributed) which assumes the values $\xi_i^\mu = \pm 1$ with equal probability:

$$\mathbb{P}(\xi_i^\mu) = \frac{1}{2}\left[\delta(\xi_i^\mu - 1) + \delta(\xi_i^\mu + 1)\right] \quad \forall i, \ \forall \mu. \tag{1.20}$$

A memory is stored when the dynamic rule takes every spin σ_i to assume the state ξ_i^μ:

$$\sigma_i = \operatorname{sgn}\left(\sum_{j=1}^N J_{ij}\xi_j^\mu\right) = \xi_i^\mu \ \forall i. \tag{1.21}$$

The Eq. (1.21) represents a condition of stability for the pattern $\boldsymbol{\xi}^\mu = \{\xi_i^\mu\}_{i=1}^N$ that is a stable fixed point of the dynamic rule (1.7). For each μ the pattern $\boldsymbol{\xi^\mu}$ represents a configuration of the network that has been learned. The system provides associative memory if the total of P patterns are dynamically stable configurations that respect the condition in Eq. (1.21). Storing P patterns generalizes the case of recalling a single pattern $\boldsymbol{\xi}^\alpha = \{\xi_i^\alpha\}_{i=1}^N$, so the expression in the Eq. (1.21) can be generalized in the case of P patterns.
The standard prescription for constructing $\boldsymbol{J}$ is the Hebbian learning rule [72, 88], often summarized as "cells that fire together, wire together" that suggests Ansatz for the synaptic matrix that represent the synaptic efficacy between pairs of neurons:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \quad \text{for} \quad i \neq j, \quad \text{and} \quad J_{ii} = 0. \tag{1.22}$$

The factor $1/N$ ensures that the local fields h_i in Eq. (1.8) remain of $\mathcal{O}(1)$ as the system size N grows. For analytical purposes, it is often assumed that the patterns are approximately orthogonal ($\xi^\mu \perp \xi^\nu \ \forall \mu \neq \nu$):

$$\frac{1}{N}\sum_{j=1}^N \xi_j^\mu \xi_j^\nu = \delta_{\mu\nu} + O\left(\frac{1}{\sqrt{N}}\right). \tag{1.23}$$

With these assumptions, if the state of the system coincides with the pattern μ at time t, the time evolution at the next time steps is the following:

$$\operatorname{sgn}\left(\sum_{j=1}^N J_{ij}\xi_j^\mu\right) = \operatorname{sgn}\left(\frac{1}{N}\sum_{j=1}^N\sum_{\nu=1}^P \xi_i^\nu \xi_j^\nu \xi_j^\mu\right) = \operatorname{sgn}\left(\sum_{v=1}^P \xi_i^\nu \delta_{\mu\nu}\right) = \operatorname{sgn}(\xi_i^\mu) \tag{1.24}$$

confirming the stability condition: $\sigma_i(t+\Delta t) = \xi_i^\mu \ \forall i$ and for sufficiently large N. In the case in which $\mu = 1$ the *Mattis Model* is obtained [78]; the whole set of N neurons is employed to store only 1 pattern of information (with its spin-flip symmetric one). For this reason the Hopfield model [73] is the correct neural network for associative memory and stands as a foundational framework for understanding how collective computation and emergent dynamics can give rise to memory retrieval.

1.2.2 Energy Landscape and Overlaps

Substituting the Hebbian prescription (1.22) into the Hamiltonian (1.19) bring to the Hamiltonian of the Hopfield model formed by N neurons σ_i, $i \in (1, \ldots, N)$ and P patterns $\boldsymbol{\xi}^\mu$, $\mu \in (1, \ldots, P)$ [12, 33, 11] :

$$H(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{N}\sum_{\mu=1}^{P}\sum_{1\le i<j\le N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j = -\frac{1}{2N}\sum_{\mu=1}^{P}\sum_{i,j}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j + \frac{1}{2N}\sum_{\mu=1}^{P}\sum_{j=i}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j \tag{1.25}$$

that can be written as

$$H(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{N}{2}\sum_{\mu=1}^{P}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i\right)^2 + \frac{P}{2} \tag{1.26}$$

We recognize the quantity in parentheses as the overlap between the current network configuration $\boldsymbol{\sigma}$ and the μ-th memory pattern. This quantity, also known as *Mattis magnetization*, is a spin-dependent variable that measures the resemblance between a given spin configuration and the stored patterns. It can be seen as the order parameter for the Hopfield model and is defined as:

$$m_\mu(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i \ \in [-1, 1]. \tag{1.27}$$

This overlap measures the proximity of the network state to the memory $\boldsymbol{\xi}^\mu$. The Hamiltonian in Eq. (1.26) can thus be rewritten as:

$$\mathcal{H}(\boldsymbol{\sigma}) = -\frac{N}{2}\sum_{\mu=1}^{P}(m^\mu)^2 + \text{constant}. \tag{1.28}$$

This form reveals that the energy is minimized when the network configuration $\boldsymbol{\sigma}$ has a large squared overlap with one of the stored patterns, confirming that these patterns are indeed the minima of the energy landscape, i.e., the attractors of the dynamics. The constant term $P/2$ in Eq. (1.26) removes diagonal contributions $J_{ii} = 0$ and ensures extensivity. In the low load regime $P \ll N$, presented in Sec. 1.3, this term is negligible in thermodynamic limit. At higher loads (in Sec. 1.4), interference among patterns (cross-talk noise) becomes significant.

1.2.3 Scaling, Load, and Lyapunov Function

The deterministic, zero-temperature dynamics of the Hopfield network is formally equivalent to the $T = 0$ limit of Glauber dynamics for an Ising spin system with

$$H = -\frac{1}{2}\sum_{i,j}J_{ij}\sigma_i\sigma_j, \tag{1.29}$$

and couplings given by the Hebbian prescription (1.22). Under these conditions, the update rule aligns each spin σ_i with its local field $h_i = \sum_j J_{ij}\sigma_j$ inducing a monotonic decrease in the energy $\mathcal{H}(\boldsymbol{\sigma})$ with each asynchronous update. Consequently, the noiseless system $T = 0$ in the case of symmetric interactions $J_{ij} = J_{ji}\ \forall\,(i,j)$ and $J_{ii} = 0\ \forall\, i$, has a Lyapunov function $L(\boldsymbol{\sigma}) = H(\boldsymbol{\sigma})$ that ensures the existence of stable equilibrium points.

To quantify the network's performance in pattern retrieval, we analyze the Mattis magnetizations defined for each pattern μ in Eq. (1.27). If a random configuration $\boldsymbol{\sigma}$ is extracted, it is uncorrelated to the patterns $\boldsymbol{\xi}\ \forall\ \mu = 1, ..., P$. This means that each term in the sum acts as an independent Bernoulli random variable with $\mathbb{P}(\xi_i^\mu \sigma_i = \pm 1) = \frac{1}{2}$. The Mattis magnetization $m_\mu(\boldsymbol{\sigma})$ is then equivalent to the net displacement of a one-dimensional random walk of N steps. Its expected value is zero, $\mathbb{E}[m_\mu] = 0$, and we can estimate its typical value with the square root of its variance:

$$\tilde{m}_\mu = \sqrt{\mathbb{E}[m_\mu(\boldsymbol{\sigma})^2]} = \sqrt{\frac{1}{N^2}\sum_{ij}\mathbb{E}[\xi_i^\mu \xi_j^\nu \sigma_i \sigma_j]} \sim \frac{1}{\sqrt{N}} \tag{1.30}$$

since $\mathbb{E}[\xi_i^\mu \xi_j^\nu \sigma_i \sigma_j] = \delta_{ij}$ for uncorrelated configurations. In contrast, if the configuration is perfectly aligned with a specific pattern, say $\boldsymbol{\sigma} = \boldsymbol{\xi}^1$, then $m_1 = 1$ and $\tilde{m_\mu} \sim \mathcal{O}(1/\sqrt{N})$ for $\mu \neq 1$. Substituting this into the Hamiltonian (1.26) yields:

$$H(\boldsymbol{\xi^1}|\boldsymbol{\xi}) \sim -\frac{N}{2} + O(1). \tag{1.31}$$

This means that if the network is prepared sufficiently near to a given pattern, so initializing the network from a corrupted version of a pattern (a point within its basin of attraction), then the network dynamics will end in a fixed point coincident with that pattern as exemplified in Figure 1.1.

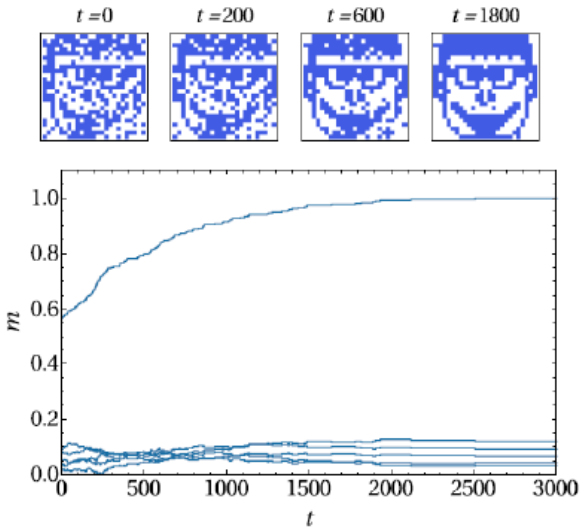


Figure 1.1 Retrieval dynamics in a Hopfield network with $P = 6$ patterns. The system is initialized from a noisy distortion of pattern $\boldsymbol{\xi}^1$. The evolution of the Mattis magnetizations $m_\mu(t)$ shows successful convergence to the target attractor $m_1 \to 1$, while the magnetizations for other patterns remain negligible. Figure taken from [1].

As can be observed in Fig. 1.1, just one of the six Mattis magnetizations dedicated to quantify the retrieval of the six stored patterns grows up to one and its corresponding pattern is retrieved.
A central parameter governing the network's memory capacity is the load parameter α, defined by the asymptotic ratio:

$$\alpha = \lim_{N\to\infty} \frac{P}{N}. \tag{1.32}$$

This parameter quantifies the relation between the number of stored patterns P and the dimensionality of the neural configurations of spins N. It distinguishes between two fundamental operational regimes:

- Vanishing load regime $\alpha \to 0$: The number of patterns P grows sublinearly with N. In this regime, the crosstalk noise between patterns is manageable, and retrieval is robust. The additive constant in the energy $\frac{P}{2}$, becomes negligible in thermodynamic limit and it is possible to omit it in the computations of the partition function and of the free energy which is the quantity that allows to analyze and describe the system completely. The system evolves to a fixed state $\boldsymbol{\sigma} = \boldsymbol{\xi}^{\boldsymbol{\mu}}$, which μ depends on the initial conditions. This regime will be discussed in Sec. 1.3.

- High-load regime $\alpha > 0$: The number of patterns scales linearly with N. This regime is characterized by increased crosstalk and interference between patterns, leading to a breakdown of retrieval capabilities at a critical capacity α_c. The high load regime will be presented in Sec. 1.4.

1.3 Low load regime

To solve the model analytically, we need to compute the free energy per neuron $f(\beta, \alpha)$ in thermodynamic limit $N \to \infty$ and derive self-consistent equations for the order parameters. We start the analysis in the low load regime and in Sec. 1.4 we enlarge the description of the Hopfield Model into the high storage case. Neglecting the constant term of the Hopfield Hamiltonian (1.26) which is not relevant in thermodynamic limit, we obtain:

$$H(\boldsymbol{\sigma}|\boldsymbol{\xi}) \sim -\frac{N}{2} \sum_{\mu=1}^{P} m_\mu^2(\boldsymbol{\sigma}). \tag{1.33}$$

The partition function Z is:

$$Z = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma}|\boldsymbol{\xi})} = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta N}{2} \sum_{\mu=1}^{P} m_\mu^2(\boldsymbol{\sigma})}. \tag{1.34}$$

The Hamiltonian depends on $\boldsymbol{\sigma}$ only through the pattern overlaps $m_\mu(\boldsymbol{\sigma})$, so we choose these as our order parameters and define the corresponding constrained partition function. Introducing the shorthand notation $\boldsymbol{m} = (m_1, ..., m_P)$, and

discarding the $\mathcal{O}(1)$ term $P/2$ from the Hamiltonian we write the constrained partition function $Z(\boldsymbol{m})$. Even in this case the mean field approximation is done:

$$Z(\boldsymbol{m}) = \sum_{\boldsymbol{\sigma}} \delta(\boldsymbol{m} - \boldsymbol{m}(\boldsymbol{\sigma})) e^{\frac{N\beta \boldsymbol{m}^2(\boldsymbol{\sigma})}{2}}. \tag{1.35}$$

where the δ-function is a P-dimensional Dirac delta

$$\delta(\boldsymbol{m} - \boldsymbol{m}(\boldsymbol{\sigma})) = \prod_{\mu=1}^{P} \delta(m_\mu - m_\mu(\boldsymbol{\sigma})). \tag{1.36}$$

The full partition function is then:

$$Z = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma}|\boldsymbol{\xi})} \xrightarrow{N\gg 1} \int d\boldsymbol{m}\, Z(\boldsymbol{m}) = \int d\boldsymbol{m}\, e^{-N\beta f_N(\boldsymbol{m})} \tag{1.37}$$

where $f_N(\boldsymbol{m})$ is the intensive, constrained free energy. The free energy $F_N(\boldsymbol{m})$ is:

$$F_N(\boldsymbol{m}) = N f_N(\boldsymbol{m}) = -\frac{1}{\beta} \ln Z(\boldsymbol{m}). \tag{1.38}$$

In thermodynamic limit ($N \to \infty$), this integral can be evaluated using the saddle-point approximation. The calculation (detailed in Appendix A.1) involves representing the δ-functions via integral representations for the auxiliary variables $\boldsymbol{x} = (x_1, \ldots, x_P)$. The final result for the intensive free energy at the saddle point is:

$$f(\boldsymbol{m}) = \frac{1}{2}\boldsymbol{m}^2 - \frac{1}{\beta}\, \mathbb{E}\left[\ln\left(2\cosh\left(\beta\, \boldsymbol{\xi} \cdot \boldsymbol{m}\right)\right)\right], \tag{1.39}$$

where the expectation $\mathbb{E}[\cdot]$ denotes an average over the quenched disorder, i.e., over the distribution of the patterns $\boldsymbol{\xi}$. Extremizing this free energy with respect to the order parameters $\boldsymbol{m}$ we get the self-consistent mean-field equations:

$$\boldsymbol{m} = \mathbb{E}\left[\boldsymbol{\xi}\, \tanh\left(\beta\, \boldsymbol{\xi} \cdot \boldsymbol{m}\right)\right]. \tag{1.40}$$

1.3.1 Stable solutions

The self-consistent mean-field equations, Eq. (1.40), admit multiple solutions. Their stability is determined by the nature of the corresponding critical points of the free energy landscape in Eq. (1.39). In this section we analyze these solutions and their stability across different temperature regimes.
For $\beta < 1$ ($T > 1$), the only stable solution is the paramagnetic state $\boldsymbol{m} = \boldsymbol{0}$. This can be shown by considering the magnitude of the overlap vector. Starting from the mean-field equation and using the properties of the pattern distribution:

$$\boldsymbol{m}^2 = \sum_{\mu} m_\mu m_\mu = \sum_{\mu} m_\mu\, \mathbb{E}\left[\xi^\mu \tanh(\beta \boldsymbol{\xi} \cdot \boldsymbol{m})\right] = \mathbb{E}\left[(\boldsymbol{\xi} \cdot \boldsymbol{m}) \tanh(\beta \boldsymbol{\xi} \cdot \boldsymbol{m})\right]. \tag{1.41}$$

Using the inequality $|x \tanh(y)| \leq |xy|$ (which holds since $|\tanh(y)| \leq |y|$), we find:

$$\boldsymbol{m}^2 \leq \mathbb{E}\left[|\boldsymbol{\xi} \cdot \boldsymbol{m}|\, |\tanh(\beta \boldsymbol{\xi} \cdot \boldsymbol{m})|\right] \leq \beta\, \mathbb{E}\left[(\boldsymbol{\xi} \cdot \boldsymbol{m})^2\right]. \tag{1.42}$$

Evaluating the expectation value:

$$\mathbb{E}\left[(\boldsymbol{\xi}\cdot\boldsymbol{m})^2\right] = \sum_{\mu,\nu} m_\mu m_\nu \mathbb{E}[\xi^\mu \xi^\nu] = \sum_\mu m_\mu^2 = \boldsymbol{m}^2, \tag{1.43}$$

where we used $\mathbb{E}[\xi^\mu\xi^\nu] = \delta_{\mu\nu}$. Combining these results we obtain the following relation:

$$\boldsymbol{m}^2 \leq \beta \boldsymbol{m}^2. \tag{1.44}$$

For $\beta < 1$ ($T > 1$), this inequality can only be satisfied if $\boldsymbol{m}^2 = 0$. Thus, the paramagnetic state is the unique solution for $T > 1$.
At the critical temperature $T_c = 1$ ($\beta_c = 1$), the paramagnetic solution becomes unstable, and new solutions emerge. To analyze the behavior near T_c, we expand the mean-field equation, Eq. (1.40), for small $\boldsymbol{m}$. Let $\tau = \beta - 1 = (1/T) - 1$, which is small near T_c. The expansion for the μ-th component is:

$$m_\mu = \mathbb{E}\left[\xi^\mu \tanh(\beta\boldsymbol{\xi}\cdot\boldsymbol{m})\right] = \sum_\nu \beta \mathbb{E}[\xi^\mu\xi^\nu] m_\nu - \frac{1}{3}\beta^3 \sum_{\nu\rho\lambda} \mathbb{E}[\xi^\mu\xi^\nu\xi^\rho\xi^\lambda] m_\nu m_\rho m_\lambda + \mathcal{O}(|\boldsymbol{m}|^5). \tag{1.45}$$

Now we use $\mathbb{E}[\xi^\mu\xi^\nu] = \delta_{\mu\nu}$ and $\mathbb{E}[\xi^\mu\xi^\nu\xi^\rho\xi^\lambda] = \delta_{\mu\nu}\delta_{\rho\lambda} + \delta_{\mu\rho}\delta_{\nu\lambda} + \delta_{\mu\lambda}\delta_{\nu\rho} - 2\delta_{\mu\nu}\delta_{\nu\rho}\delta_{\rho\lambda}$. For the forth moment, the expectation is non-zero only when the indices are pairwise equal and the last term subtracts the overcounted case when all four indices are the same. We compute $S = \sum_{\nu\rho\lambda} \mathbb{E}[\xi^\mu\xi^\nu\xi^\rho\xi^\lambda] m_\nu m_\rho m_\lambda$ term by term:

- Term 1 ($\delta_{\mu\nu}\delta_{\rho\lambda}$): $\sum_{\nu\rho\lambda} \delta_{\mu\nu}\delta_{\rho\lambda} m_\nu m_\rho m_\lambda = m_\mu \sum_\rho m_\rho^2 = m_\mu \boldsymbol{m}^2$
- Term 2 ($\delta_{\mu\rho}\delta_{\nu\lambda}$): $\sum_{\nu\rho\lambda} \delta_{\mu\rho}\delta_{\nu\lambda} m_\nu m_\rho m_\lambda = m_\mu \sum_\nu m_\nu^2 = m_\mu \boldsymbol{m}^2$
- Term 3 ($\delta_{\mu\lambda}\delta_{\nu\rho}$): $\sum_{\nu\rho\lambda} \delta_{\mu\lambda}\delta_{\nu\rho} m_\nu m_\rho m_\lambda = m_\mu \sum_\rho m_\rho^2 = m_\mu \boldsymbol{m}^2$
- Term 4 ($-2\delta_{\mu\nu}\delta_{\nu\rho}\delta_{\rho\lambda}$): $-2\sum_{\nu\rho\lambda} \delta_{\mu\nu}\delta_{\nu\rho}\delta_{\rho\lambda} m_\nu m_\rho m_\lambda = -2m_\mu^3$

Putting all together:

$$m_\mu = \beta m_\mu - \beta^3 m_\mu \left(\boldsymbol{m}^2 - \frac{2}{3} m_\mu^2\right) + \mathcal{O}(|\boldsymbol{m}|^5). \tag{1.46}$$

Substituting $\beta = 1 + \tau$ and keeping terms up to order $\mathcal{O}(\tau|\boldsymbol{m}|, |\boldsymbol{m}|^3)$ we have:

$$m_\mu \simeq (1+\tau) m_\mu - m_\mu \left(\boldsymbol{m}^2 - \frac{2}{3} m_\mu^2\right). \tag{1.47}$$

So the equation becomes:

$$m_\mu \left(\tau - \boldsymbol{m}^2 + \frac{2}{3} m_\mu^2\right) = 0. \tag{1.48}$$

For each component μ, Eq. (1.48) implies either:

$$m_\mu = 0 \quad \text{or} \quad m_\mu^2 = \frac{3}{2}(\boldsymbol{m}^2 - \tau). \tag{1.49}$$

Let n be the number of non-zero components in $\boldsymbol{m}$. By symmetry, these non-zero components must have equal magnitude, $|m_\mu| = m$. Thus, $\boldsymbol{m}^2 = nm^2$. Substituting into the non-trivial branch:

$$m^2 = \frac{3}{2}(nm^2 - \tau) \quad \Rightarrow \quad m^2\left(1 - \frac{3n}{2}\right) = -\frac{3}{2}\tau. \tag{1.50}$$

Solving for m^2 gives the amplitude for the non-zero components:

$$m = \pm\sqrt{\frac{3}{3n-2}\tau}. \tag{1.51}$$

These solutions, known as *symmetric mixture states*, represent configurations where the network activity has a uniform overlap with n different patterns because they are retrieved with the same intensity since m_μ does not depend on μ. The number of such solutions is $2^n\binom{P}{n}$, accounting for the choice of which n patterns are active and the sign of each overlap. Mixture states are connected with the simultaneous recall of more patterns; by storing more and more patterns, eventually interferences emerge and yield to a number of local minima. These additional minima have no counterpart in terms of stored patterns, so they are traditionally called spurious fixed points. An example of spurious attractor is given by the configuration:

$$\xi = \text{sign}(\xi^1 + \xi^2 + \xi^3). \tag{1.52}$$

The number of such configurations grows very fast with the number of stored patterns P; from the dynamical point of view, this means that, storing more patterns, the probability for the network dynamics to be trapped in the attraction basins of spurious states gets higher and the attracting power of pure fixed points is dramatically downsized.
An example of dynamics ending in spurious configurations is reported in the following Figure 1.2. The network is prepared in the spurious configuration (1.52). Each spin is again flipped with a probability different from zero and the network is left free to evolve for a sufficient long time. In this case, three Mattis magnetization raise sensibly over the noise. The network is not able to properly retrieve a single pattern, obtaining a useless mixture of the stored patterns.

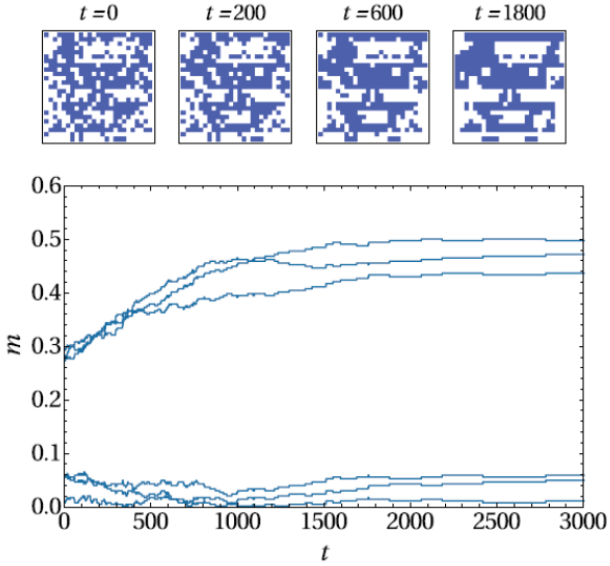


Figure 1.2 Dynamics ending in a spurious state in a Hopfield network with $P = 6$ patterns. Figure taken from [1]

.

Without loss of generality, a symmetric mixture state can be written as:

$$\boldsymbol{m} = m_n(\underbrace{1,\ldots,1}_{n \text{ times}}, \underbrace{0,\ldots,0}_{P-n \text{ times}}), \tag{1.53}$$

with:

$$m_n = \sqrt{\frac{3}{3n-2}}\sqrt{1-T}. \tag{1.54}$$

In this expression, n denotes the dimensionality of m, the number of non zero and equal components, in a particular solution below T_c [13]. The others $P-n$ are set to 0 for a total of $2^n\binom{P}{n}$ solutions. The first factor stands for the number of possible sign changes and the second one for the number of ways that the n non-zero components can be selected out of the P stored patters. These symmetric solutions are important because they are the only solutions that exist throughout the whole temperature range $T < 1$.
By substituting solutions of this type into the saddle point equation (1.40) we obtain:

$$0 = \mathbb{E}[\boldsymbol{\xi}\tanh(\beta m_n M)] \qquad \text{if } \mu > n \tag{1.55}$$

$$m_n = \mathbb{E}[\boldsymbol{\xi}\tanh(\beta m_n M)] \qquad \text{if } \mu < n \tag{1.56}$$

with:

$$M = \sum_{\nu=1}^{n} \xi^{\nu}. \tag{1.57}$$

When the values of m_n are substituted in the expression of the energy, an emergent ordering is found for the symmetric saddle points at $T = 0$ [13]:

$$f_1 < f_3 < f_5 < \cdots < \cdots < f_6 < f_4 < f_2, \tag{1.58}$$

where f_n denotes the free energy of an n-mixture. The energies of the odd mixture states increase with the number of admitted patterns while the energies of the even mixtures decrease with n. All these mixture states are solutions of the mean field equation (1.40) but they aren't necessarily attractors. In a mixture many patterns are recalled simultaneously in a mixed way and we are unable to recognize any pattern individually since these states have non-zero overlap with more than one memory.
To assess stability, we examine the Hessian matrix of second derivatives of the free energy $f(\boldsymbol{m})$: a solution is stable if the Hessian is negative definite. Those that are positive definite are minima of the energy (true attractors). Those extreme at which the second derivative of $f(\boldsymbol{m})$ has decreasing directions are unstable. For a symmetric n-mixture state, the Hessian takes a block-diagonal form (see Appendix A.2 for the full calculation). Its eigenvalues are:

- $\lambda_1 = 1 - \beta(1-q)$ with degeneracy $P-n$
- $\lambda_2 = 1 - \beta(1-q) - \beta Q$ with degeneracy $n-1$
- $\lambda_3 = 1 - \beta(1-q) + (n-1)\beta Q$ with degeneracy 1.

Expanding the eigenvalues near $T \to T_c = 1$, we can find that [13]: $\lambda_1 < 0$ and $\lambda_3 < 0$ but $\lambda_2 > 0$ for all $n > 1$. Thus, only pure states $n = 1$ are stable immediately below T_c because in this way $\lambda_2 = 0$ and the other two eigenvalues are negative. All other solutions are saddle points since the Hessian has positive and negative eigenvalues. So for $n = 1$ there is only a local stationary point that is stable and correspond to ferromagnetic states. All stationary points with $n > 1$ are unstable. As temperature decreases further, odd mixture states $n = 3, 5, 7, \ldots$ become metastable (locally stable but not global minima) at specific temperatures $0 < T_n < 1$, while even mixtures remain unstable at all temperatures. Each odd mixture becomes locally stable at its own temperature.
An analysis of the signs of these eigenvalues reveals the following phase structure at low α:

- **Paramagnetic Phase** $T > 1$: Only the paramagnetic solution $\boldsymbol{m} = 0$ is stable. The system is ergodic with random spin orientation and no pattern recall.
- **Critical Point** $T_c = 1$: Second-order phase transition. Symmetric mixture solutions emerge.
- **Retrieval Phase** $T^{(3)} < T < 1$: Paramagnetic state unstable. The $2P$ pure states (pattern and its inverse) are stable global minima. All mixture states are unstable. The pure states $n = 1$ are the only stable attractors. The network functions effectively as an associative memory, with each stored pattern having a large basin of attraction.
- **Emergence of Spurious States** $T < T^{(n)}$: Spurious states begin to appear. The pure states remain the global energy minima and even mixtures are always unstable.
  - $T^{(3)} \approx 0.461$: 3-pattern mixtures become metastable.
  - $T^{(5)} \approx 0.385$: 5-pattern mixtures become metastable.
  - $T^{(7)} \approx 0.345$: 7-pattern mixtures become metastable.
- **Zero Temperature** $T = 0$: All odd-n mixture states are stable. The even-n mixtures are always unstable saddle points. The pure states and their reversed counterparts are the ground states with well defined basin of attraction. The number of metastable odd mixture states grows exponentially with P [13, 15].

The proliferation of spurious states complicates the dynamics and reduces the effective capacity of the network especially in the high load regime that will be presented in the following Section 1.4. Instead, a deeper discussion of these fixed points will be exposed in Section 1.5 using the *signal to noise analysis*.

1.4 High load regime

Until now we have explored the Hopfield model scenario in the low load regime and in Sec. 1.3.1 the behaviour of the system increasing the temperature T. The

Hoplield model is now studied near its saturation so when the number P of stored patterns increases with the size of the network N [13, 15]. The model's behavior changes qualitatively when the number of stored patterns P scales extensively with the system size N, i.e., $P = \alpha N$, where $\alpha > 0$ is the storage capacity or load parameter in Eq. (1.32). In this high-load regime, the network can be still used as an associative memory, even if the stability of the embedded patterns is no longer guaranteed. While the original patterns may be destabilized, new stable states may develop, very near the original learnt ones.
The analytical study of the Hopfield model in the regime of an extensive number of patterns require the computation of the free energy per spin $f = -\lim_{N\to\infty} \frac{1}{\beta N} \ln Z_N$ in thermodynamic limit $N \to \infty$. To resolve this computation we need to average the logarithm of the partition function Z_N (Eq. 1.34) over the quenched disorder, represented by the random patterns $\{\xi_i^\mu = \pm 1\}$ (Eq. 1.20). The central technical difficulty is represented by the computation of the average $\mathbb{E}[\ln Z]$.
Before proceeding into the specific description of the Hopfield model in the high load regime, we briefly describe the **Replica Method** with the **Replica Symmetry (RS) Ansatz** and its physical failure in certain conditions that conducted to the famous Parisi's **Replica Symmetry Breaking** formulation. For a detailed description we remind to [97, 96, 102] while recent upgrade in the field of spin glasses are contained in [31].

The Replica Method

For the computation of the average $\mathbb{E}[\ln Z]$, the **replica trick** is employed, starting from the mathematical identity:

$$\mathbb{E}[\ln Z] = \lim_{n\to 0} \frac{\mathbb{E}[Z^n] - 1}{n}. \tag{1.59}$$

This transforms the problem into computing the moments $\mathbb{E}[Z^n]$ for integer n and analytically continuing the result to $n \to 0$. For the Hopfield model, $\mathbb{E}[Z^n]$ is expressed as the partition function of an effective n-replica system.

Replica Symmetry and its Failure

The initial, simplest assumption is that of **Replica Symmetry (RS)**, where the solution is invariant under any permutation of the n replicas. This leads to an ansatz for the structure of the overlap matrix $Q_{ab} = \frac{1}{N}\sum_i S_i^a S_i^b$, which is a key order parameter in the replicated theory. Under the RS ansatz, this matrix takes the form:

$$Q^{ab} = \delta_{ab} + q(1 - \delta_{ab}), \tag{1.60}$$

where q is the Edwards-Anderson order parameter [49], representing the typical overlap between two different equilibrium states.
However, as was famously discovered in the context of the Sherrington-Kirkpatrick (SK) model for spin glasses [128], the RS solution is often unstable and leads to unphysical results (e.g., a negative entropy at low temperatures) for a wide range of parameters. This indicates the emergence of a complex free energy landscape composed by many metastable states, a phenomenon known as **ergodicity breaking**.

Parisi's Replica Symmetry Breaking

The correct solution for such systems requires the breaking of the replica symmetry. The monumental breakthrough came from Giorgio Parisi, who introduced a hierarchical scheme for breaking the symmetry of the overlap matrix Q^{ab} in stages [108, 109]. This **Replica Symmetry Breaking (RSB)** ansatz is not a single approximation but an infinite sequence of ansatze (1RSB, 2RSB, ..., fullRSB) of increasing complexity, which ultimately describes a continuous distribution of overlaps between states. The full Parisi RSB solution for the SK model proved to be stable and physically consistent.

Guerra's Interpolation Method

While the replica method provides a powerful analytical framework, the mathematical intricacies of handling the $n \to 0$ limit, especially under RSB, are profound. A significant advancement in placing these results on a more rigorous footing came with the work of Francesco Guerra. He introduced an **interpolation method** [65] that allows one to prove the validity of the Parisi RSB free energy expression for the SK model as an upper bound (Guerra's bound) at all temperatures. This method interpolates between the original disordered system and a simpler, decoupled system, using convexity and integration inequalities. This was the first major rigorous result supporting the correctness of Parisi's solution followed by successive investigations [67, 66]. Later, Talagrand [135] completed the proof by establishing the corresponding lower bound, thus fully confirming Parisi's solution. This was the first complete proof that the Parisi formula gives the exact free energy for the SK model at all temperatures and external fields.

The Hopfield model in the extensive-load regime shares a profound connection with the SK model. Its effective Hamiltonian can be mapped to that of a Sherrington-Kirkpatrick-type model with additional, pattern-related order parameters. Consequently, we can derive the RS solution for the model that fails for low enough temperatures and certain values of α, necessitating a Parisi-style RSB treatment to obtain the correct thermodynamic description and understand the structure of its spin glass phase (Sec. 1.7.3). We want also to underline that rigorous mathematical interpolation technique analogous to that of Guerra [65] made on the SK model, have been developed also for the resolution of the Hopfield model [134, 9, 112, 133, 19, 4]. In our analysis of the Hopfield model, we follow the replica method approach to derive the averaged free energy. We then analyze the stability of the RS solution and, where it fails, we adopt the Parisi hierarchical RSB scheme, the structure of which is directly inspired by his solution for the SK model [108, 109].
In the high storage case even the second term in the Hamiltonian (1.26) is relevant since it is of order $O(N)$ and so contributes to the free energy. Here is reported again the Hopfield Hamiltonian presented in Eq. (1.26):

$$H(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{\mu=1}^{P}\sum_{i,j}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j + \frac{1}{2N}\sum_{\mu=1}^{P}\sum_{j=i}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j = -\frac{N}{2}\sum_{\mu=1}^{P}m_\mu^2(\boldsymbol{\sigma}) + \frac{P}{2}. \tag{1.61}$$

In this case, the saddle point integration does not apply since it involves integrals over an extensive number of variables. Indeed, when α becomes higher is no longer possible to evaluate the constrained free energy by saddle point integration since the dimension of the integral diverges at the same time as the exponent of the integrand. The number of ground states of the Hamiltonian will diverge in a spin glass configuration. When P reaches the $O(N)$, a spin glass solution emerges: the state of the system is randomly frozen and has no correlation with the patterns. If the load $\alpha = \frac{P}{N}$ exceeds a threshold, the spin glass state becomes the only stable one at low temperature.
The analysis begins with the calculation of the average replicated partition function, $\mathbb{E}[Z^n]$, where the expectation is taken over the disorder represented by the patterns $\boldsymbol{\xi}$. For simplicity, since all patterns are equivalent in the calculation, we assume a finite number l of patterns are "condensed," meaning they have macroscopic overlaps $m_\mu = \mathcal{O}(1)$, while the remaining $P - l$ patterns act as a source of quenched noise, with overlaps of order $\mathcal{O}(1/\sqrt{N})$. In order to do that, we rewrite the Hamiltonian (1.26) as:

$$H(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{\mu=1}^{l}\left(\sum_{i=1}^{N}\sigma_i\xi_i^\mu\right)^2 - \frac{1}{2N}\sum_{\mu>l}^{P}\left(\sum_{i=1}^{N}\sigma_i\xi_i^\mu\right)^2 + \frac{P}{2}. \tag{1.62}$$

After introducing the replica symmetric (RS) Ansatz and employing standard techniques—including Gaussian linearization, the Hubbard-Stratonovich transformation, and saddle-point integration in thermodynamic limit $N \to \infty$—the intensive free energy per spin is derived. The stable states of the system correspond to the saddle points of this free energy, found by extremizing with respect to the order parameters $\boldsymbol{m}$, $\boldsymbol{q}$ and $\boldsymbol{r}$. We report the details of the full calculation [11, 33, 13, 15] in Appendix A.3.
The resulting RS free energy is:

$$f(\boldsymbol{m}, q, r) = \frac{\alpha}{2} + \frac{\alpha}{2\beta}\left[\ln(1-\beta(1-q)) - \frac{\beta q}{1-\beta(1-q)}\right] + \frac{1}{2}\boldsymbol{m}^2 + \tag{1.63}$$

$$+\frac{1}{2}\alpha\beta r(1-q) - \frac{1}{\beta}\int Dz\ \mathbb{E}\Big[\ln\Big(2\cosh\Big(\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi})\Big)\Big)\Big] \tag{1.64}$$

where $\int Dz = \int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}$ denotes the Gaussian measure, and $\mathbb{E}_{\boldsymbol{\xi}}[\cdot]$ is the average over the condensed patterns.
The self-consistent equations in the replica-symmetric ansatz for the relevant order parameters are:

$$\boldsymbol{m} = \int Dz\ \mathbb{E}\Big[\boldsymbol{\xi}\tanh\Big(\beta\Big(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi}\Big)\Big)\Big], \tag{1.65}$$

$$q = \int Dz\ \mathbb{E}\Big[\tanh^2\Big(\Big(\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi})\Big)\Big)\Big], \tag{1.66}$$

$$r = \frac{q}{(1-\beta(1-q))^2}. \tag{1.67}$$

Equation (1.65) is a vector equation generalizing the mean-field equation for the magnetization from the low load case. Equation (1.66) defines the Edward-Anderson order parameter [49], which measures the alignment between different replicas and is non-zero in the spin glass phase. Equation (1.67) couples the spin glass parameters q and r. These are the equation in the case of zero external field $\boldsymbol{h}$ that enters just as an additive factor in the $(\boldsymbol{m} + \boldsymbol{h}) \cdot \boldsymbol{\xi}$ terms. These vectors have s components each corresponding to the patterns that have condensed. In conclusion, the free energy of the Hopfield model and the equations of state in the high-load regime consist of two competing contributions: a ferromagnetic part, governed by the order parameter $\boldsymbol{m}$ which results from the l condensed patterns, and a spin-glass part, generated by the term $\sqrt{r\alpha}z$ which represents the collective interference from the remaining $P - l$ patterns.

1.5 Signal to noise analysis

The stability of stored patterns can be analyzed directly at zero temperature ($T = 0$) using a signal-to-noise analysis. This approach gives an useful way to estimate the network's storage capacity by separating the local fiele acting on a neuron into a "signal" term that reinforces the desired pattern and a "noise" term caused by interference from all other stored patterns.
The signal to noise analysis is used to establish under which conditions a given network configuration is stable with respect to the *intrinsic noise*. Indeed it is necessary to distinguish thermal noise, the fast one that is unavoidable, from this contribution. The system does not move deterministically towards its ground state and does not stays permanently there due to intrinsic randomness related to the temperature. The slow noise is instead related to the load α: the number P of stored patterns constitute a source of noise. If we try to store too much information in the network, the retrieval is ineffective and the system goes in a *black-out scenario* [13].
In the signal to noise analysis the external thermal noise is set to zero $T = 0$.
We consider a network configuration perfectly aligned with a single stored pattern, without losing generality $\boldsymbol{\sigma} = \boldsymbol{\xi}^1$, and we analyze the stability of a specific spin, σ_i, by checking if the local field h_i aligns with it. The condition of dynamical stability is satisfied if $\xi_i^1 h_i > 0 \; \forall \, i$. The local field is given by:

$$h_i = \sum_{j \neq i} J_{ij} \sigma_j = \frac{1}{N} \sum_{j \neq i} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \xi_j^1. \tag{1.68}$$

Multiplying by ξ_i^1 and separating the contribution from the retrieved pattern $\mu = 1$ from all others $\mu > 1$, we get:

$$\xi_i^1 h_i = \underbrace{\frac{1}{N} \sum_{j \neq i} \xi_i^1 \xi_j^1 \xi_j^1 \xi_i^1}_{\text{Signal (S)}} + \underbrace{\frac{1}{N} \sum_{j \neq i} \sum_{\mu=2}^{P} \xi_i^\mu \xi_j^\mu \xi_j^1 \xi_i^1}_{\text{Noise (R)}} . \tag{1.69}$$

Observing the previous expression, we can define:

- **Signal Term S** Since $(\xi_i^1)^2 = (\xi_j^1)^2 = 1$, this simplifies to $S = \frac{N-1}{N} \to 1$ in thermodynamic limit $N \to \infty$. This term strongly favors stability since it tends to align the network configuration with the pattern $\boldsymbol{\xi^1}$. It is, therefore, related to the retrieval of the chosen pattern.
- **Noise Term R** This term arises from the interference of all other stored patterns $\mu = 2, ..., P$ and make difficult the retrieval since it destroys the correlations between the configuration $\boldsymbol{\sigma}$ and the $\boldsymbol{\xi^1}$. Each term in the sum, $\xi_i^\mu \xi_j^\mu \xi_j^1 \xi_i^1$, is a random variable taking values ± 1 with equal probability. The sum consists of $(N-1)(P-1) \approx NP$ such independent random variables.

By the Central Limit Theorem, the noise term R is a Gaussian random variable with:

$$\mathbb{E}[R] = 0, \quad \text{Var}(R) = \frac{NP}{N^2} = \frac{P}{N} = \alpha. \tag{1.70}$$

Thus, $R \sim \mathcal{N}(0, \sqrt{\alpha})$ [13, 15]. The pure attractor configurations are stable when the intrinsic noise of the network is negligible so when P is kept fixed and N is very large. This condition is no longer valid in the high storage regime.

1.5.1 Stability Criterion and Capacity Limit

A spin is stable if the total local field aligns with its desired state:

$$\xi_i^1 h_i = 1 + R > 0 \quad \Rightarrow \quad R > -1. \tag{1.71}$$

As $\alpha > 0$, that is the variance of R, the probaiblity that $h_i \xi_i^1 > 0$ is not tending to 1 anymore and the probability that a *single spin* is stable $p_{\text{spin}} = \mathbb{P}(h_i \xi_i^1 > 0)$ is given by:

$$p_{\text{spin}} = \mathbb{P}(R > -1) = \frac{1}{\sqrt{2\pi\alpha}} \int_{-1}^{\infty} e^{-x^2/(2\alpha)} dx = \frac{1}{2}\left[1 + \text{erf}\left(\frac{1}{\sqrt{2\alpha}}\right)\right]. \tag{1.72}$$

For the *entire pattern* $\boldsymbol{\xi}^1$ to be a stable fixed point, all N spins must be stable. If we treat these events as independent, the probability of full pattern stability is:

$$P_{\text{pattern}} = \prod_{i=1}^{N} \mathbb{P}(h_i \xi_i^1 > 0) = (p_{\text{spin}})^N. \tag{1.73}$$

For this probability to approach 1 (i.e., for patterns to be retrievable with high probability in the large-N limit), p_{spin} must be extremely close to 1. We examine the limit where α is small [11]. Using the asymptotic expansion of the error function for large arguments, $\text{erf}(z) \sim 1 - \frac{e^{-z^2}}{\sqrt{\pi} z}$, Eq. (1.72) becomes:

$$p_{\text{spin}} \sim 1 - \sqrt{\frac{\alpha}{2\pi}} e^{-1/(2\alpha)}. \tag{1.74}$$

The probability that the entire pattern $\boldsymbol{\xi}^1$ is stable is then:

$$P_{\text{pattern}} \sim \left(1 - \sqrt{\frac{\alpha}{2\pi}} e^{-1/(2\alpha)}\right)^N \approx 1 - N\sqrt{\frac{\alpha}{2\pi}} e^{-1/(2\alpha)}. \tag{1.75}$$

This quantity must be close to unity in order to have perfect retrieval; so the second term, that is the amount of error $\mathbb{P}_{\text{err}}$, must be negligible when $N \to \infty$. This requires the exponential factor to decay faster than N grows. The condition to be satisfied is:

$$e^{-1/(2\alpha)} \ll \frac{1}{N} \quad \Rightarrow \quad \frac{1}{2\alpha} > \ln N \quad \Rightarrow \quad \alpha \leq \alpha_1 \equiv \frac{1}{2\ln N}. \tag{1.76}$$

This relation represents an upper bound for the retrieval of one pattern without errors. If P grows faster than $N/\ln N$, the probability that a stored pattern remains a stable attractor vanishes in thermodynamic limit.
If we now require that *all* P patterns are retrieved without errors, we can compute the average number of errors in all pattern that is $P \cdot \mathbb{P}_{\text{err}}$ and see up until which α scaling it will tend to zero.

$$\mathbb{P}_{\text{err}}^{\text{tot}} = P \cdot \mathbb{P}_{\text{err}} = PN\sqrt{\frac{\alpha}{2\pi}}e^{-1/(2\alpha)} = N^2\sqrt{\frac{\alpha^3}{2\pi}}\exp\left(-\frac{1}{2\alpha}\right). \tag{1.77}$$

Now inserting the value of α_1 obtained from Eq. (1.76) we have:

$$\mathbb{P}_{\text{err}}^{\text{tot}} = \frac{N^2}{\sqrt{16\pi a^3(\ln N)^3}}(e^{\ln N})^{-a} = \frac{N^{2-a}}{\sqrt{16\pi a^3(\ln N)^3}} \tag{1.78}$$

that goes to zero as $a \geq 2$ and gives [95, 11]:

$$\alpha_P = \frac{1}{4\ln N}. \tag{1.79}$$

For a value of the load $\frac{1}{4\ln N} < \alpha < \frac{1}{2\ln N}$ then still most fundamental memories will be stable [95, 15] and each pattern has a non trivial radius of attraction. The memories are not retrievable exactly, but it is still possible to find stable states in their vicinity.

1.5.2 The "Black-Out" Scenario and Critical Load

The associative memory provided by the system is degraded only when $P > \alpha_c N$ with $\alpha_c \simeq 0.1 - 0.2$ [15]. When $P = \alpha N$ and, in particular, P becomes larger than a critical value $P_c = 0.138N$ [13], the network goes into a state of total confusion and a negligible amount of patterns are remembered.
This threshold represents the load above which retrieval states (correlated with a single pattern) cease to exist altogether. Indeed, despite the existence of a finite amount of noise which destabilizes the stored patterns, the modified stable states remain very near the original patterns, if $\alpha < \alpha_c \simeq 0.14$. In this case, no more than 1% of the neurons deviate from the memorized patterns [35]. At $\alpha \simeq \alpha_c$ the system saturates and the noise created by the random overlaps of every memorized pattern with all others overcomes the whole system; this means that all the patterns are no longer retrievable [13, 15].
As long as $\alpha < \alpha_c$ the overlap between them and the original pattern is $m \geq 0.97$, which is the value of the overlap at the critical load α_c. As said previously, the retrieval of memories is stable against a small amount of stochastic noise even if

α is finite but the maximum allowed level of noise decreases to zero as the system approaches its maximum capacity α_c.
The model exhibits a complex trade-off between storage capacity α, retrieval quality m, and noise robustness T. The maximum critical load for retrieval at $T = 0$ is $\alpha_c \approx 0.138$ [13, 15]. When the number of patterns P exceeds the critical value $P > \alpha_c N$, the network enters a **spin-glass phase**. In this phase:

- The noise from pattern interference completely overwhelms the signal.
- The original patterns are no longer stable attractors.
- The network's dynamics become dominated by a vast number of spurious states, leading to a failure of associative memory—a "black-out" scenario.

Thus, the Hopfield network exhibits a fundamental trade-off between storage capacity and retrieval quality. The maximum number of patterns that can be stored and reliably retrieved scales linearly with N, but with a proportionality constant α_c that is surprisingly small $\alpha_c \approx 0.14$, highlighting a fundamental limitation of this simple memory model.
With a finite noise T we have that the condition for retrieval is given by $P < \alpha_c(T)N$ with $\alpha_c(T) < \alpha_c$. The retrieval quality at $\alpha_c(T)$ is measured by the overlap $m_c(T)$ [13]. In the Section 1.7 we will describe the phase diagram of the Hopfield model in the parameters $(T - \alpha)$.

1.6 T=0 equations

Before proceeding into the description of the entire phase diagram $(T - \alpha)$ we briefly describe the behavior of the Hopfield model in the zero temperature limit $T = 0$ [15, 33]. In this case, the RS equation for the magnetization $\boldsymbol{m}$ in Eq. (1.65) is simplified since the Gaussian integral can be written as:

$$\int_{-\infty}^{-\infty} dz\, e^{-\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} \tanh[\beta(Az + x)] \to \sqrt{\frac{2}{\pi}} \int_0^{x/A} dz\, e^{-\frac{z^2}{2}} = \operatorname{erf}\left(\frac{x}{\sqrt{2}A}\right) \tag{1.80}$$

This means that the mean field equation for the magnetization $\boldsymbol{m}$ becomes:

$$\boldsymbol{m} = \mathbb{E}\left[\boldsymbol{\xi} \operatorname{erf}\left(\frac{\boldsymbol{m} \cdot \boldsymbol{\xi}}{\sqrt{2\alpha r}}\right)\right]. \tag{1.81}$$

This zero temperature expression should be compared with the corresponding equation of the finite-P limit (1.40). There is a resemblance between the erf and the tanh of the low case regime. If r remains finite, the zero temperature finite-α overlaps behave very much similar to the finite temperature, finite-P case. As $\alpha \to 0$, Eq. (1.81) reduces to $\boldsymbol{m} = \mathbb{E}[\boldsymbol{\xi} \operatorname{sign}(\boldsymbol{m} \cdot \boldsymbol{\xi})]$ that represents the zero noise limit of Eq. (1.40). The term $\sqrt{2\alpha r}$ is therefore interpretable as a noise caused by the interference of the pattern that hinders the retrieval.
We assume $\boldsymbol{m}$ has one single component different from zero m [15], then the Eq. (1.81) becomes:

$$m = \operatorname{erf}\left(\frac{m}{\sqrt{2\alpha r}}\right). \tag{1.82}$$

In the limit $\beta \to \infty$ the overlap $q \to 1$; we can define a parameter $C = \lim_{\beta\to\infty} \beta(1-q)$ that will be [15]:

$$C = \sqrt{\frac{2}{\pi\alpha r}} e^{-\frac{m^2}{2\alpha r}} \quad \text{with} \quad r = \lim_{\beta\to\infty} \frac{q}{(1-\beta(1-q))^2} = \frac{1}{(1-C)^2}. \tag{1.83}$$

If we rescale the variable defining $x = \frac{m}{\sqrt{2\alpha r}}$, we get:

$$m = \text{erf}(x), \quad C = \frac{2}{\sqrt{\pi}} \frac{1}{\sqrt{2\alpha r}} e^{-x^2} \quad \text{and} \quad \sqrt{r} = 1 + \frac{2}{\sqrt{\pi}} \frac{1}{\sqrt{2\alpha}} e^{-x^2}. \tag{1.84}$$

These equations reduce to a single equation for the variable x:

$$x = \frac{\text{erf}(x)}{\sqrt{2\alpha} + \frac{2}{\sqrt{\pi}} e^{-x^2}}. \tag{1.85}$$

This equation gives the relation between the retrieval quality m and the storage load α. We can solve Eq. (1.85) numerically finding the value of the critical capacity α_c; the graphical solution is in Fig. 1.3.

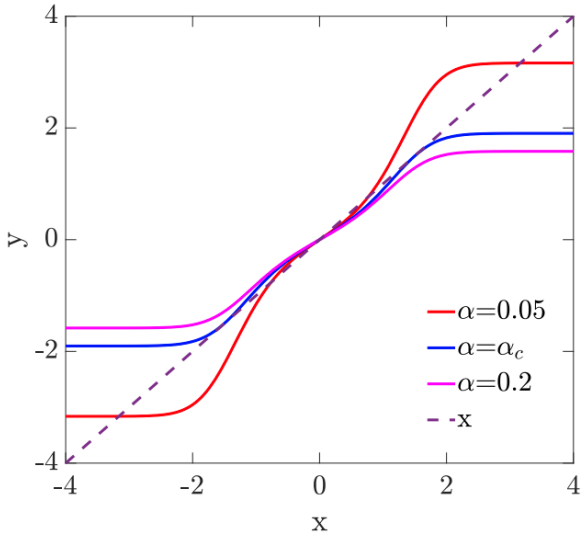


Figure 1.3 Graphical solution of Eq. (1.85). The continuous line refers to the right member of the equation while the dotted line is $y = x$. The equation is plotted for different values of α: $\alpha = 0.05 < \alpha_c$, $\alpha = 0.2 > \alpha_c$ and $\alpha = \alpha_c$.

The highest possible value for which we can find a non-zero x is the critical value for the retrieval regime, i.e., $m \neq 0$. Therefore, the critical capacity α_c is simply the value of α above which the equation has no solution, except $x = 0$ that is always a solution. This is the spin-glass (SG) solution $m = 0$, with no macroscopic overlaps with any of the patterns. For $\alpha > \alpha_c$ this is the only solution. For $\alpha < \alpha_c$, the equation admits three solutions $x = 0$, $x = \pm x^*$: $2P$ (FM) solutions with $m \neq 0$ appear. They represent the spin glass at $m = 0$ and the retrieval ones with higher $|m|$ both dynamically stable.

The change is abrupt in the value of the stable x solution and the retrieval/spin-glass transition is of the first order. In particular α_c is a spinodal point for the retrieval phase, whereas, in the context of thermodynamic phase transition, the first order critical point is the value of α at which the coexisting retrieval and black-out phases (or, equivalently, ferromagnetic and spin-glass phases) display the same free energy, i.e., they are probabilistically equivalent.
As $\alpha \to 0$ the expression in the Eq. (1.82) becomes:

$$m \sim 1 - \sqrt{\frac{2\alpha}{\pi}} e^{-\frac{1}{2\alpha}} \tag{1.86}$$

This is because in the region where FM retrieval state exists, $m \simeq 1$ so the error function has an asymptotic behavior for $x \to \infty$ [15]:

$$\text{erf}(x) \simeq 1 - (\sqrt{\pi}x)^{-1} e^{-x^2}. \tag{1.87}$$

Let's define N_{err} the average number of spins that do not align with the patterns:

$$N_{\text{err}} = \frac{N}{2}(1 - m). \tag{1.88}$$

This equation follows directly from the relation between the overlap and the Hamming distance [11]. Looking at the variable N_{err} from the perspective of the unstable entries, we can observe that the mean number of error in a pattern is the probability for a spin σ_i to be unstable times the total number of spins N. So it coincides exactly with the $\mathbb{P}_{\text{err}}$ in Eq. (1.77) when we insert the asymptotic expression of the magnetization m:

$$\frac{N_e}{N} \simeq \sqrt{\frac{\alpha}{2\pi}} e^{-\frac{1}{2\alpha}}. \tag{1.89}$$

The behavior of the average fraction of errors as a function of the load α at $T = 0$, is reported in Figure 1.4. As can be seen from the figure, there is a sharp critical value $\alpha_c \sim 0.138$ which is the threshold in α and the level of errors in retrieval before the collapse at a storage level of $0.138N$ patterns, is never higher then 1.5%.
In [35] an improvement for the average percentage of errors if found until a critical value of $\alpha_c^{1RB} = 0.144$ due to replica symmetry breaking that will be analyzed briefly in Sec. 1.7.3.
Looking at the expression in Eq. (1.89) we can observe that the average fraction of errors N_e/N vanishes for $\alpha \to 0$. For a finite value of α, N_e/N is finite. So, if a small fraction of misaligned spins is allowed, the retrieval is possible when $P < \alpha_c N$ in complete agreement with the signal to noise analysis done in Sec. 1.5.
The most important result of this Section is that there exists a value of α_c where the system provides an effective retrieval of memory. Indeed, despite the existence of a finite amount of noise which destabilizes the stored patterns, the modified stable states remain very near the original patterns if $\alpha < \alpha_c = 0.138$. As long as $\alpha < \alpha_c$, the overlap between the retrieval patterns and the original one is greater than $m \simeq 0.97$ [13, 15]. As α decreases to zero, m increases exponentially fast towards 1 at $T = 0$. For $\alpha > \alpha_c$ the overlap is $m \simeq 0.35$ that is much smaller than the value $m \simeq 0.97$ at α_c^- but it is anyway higher with respect to the overlap with a random initial state $m \simeq 0.08$ [15].

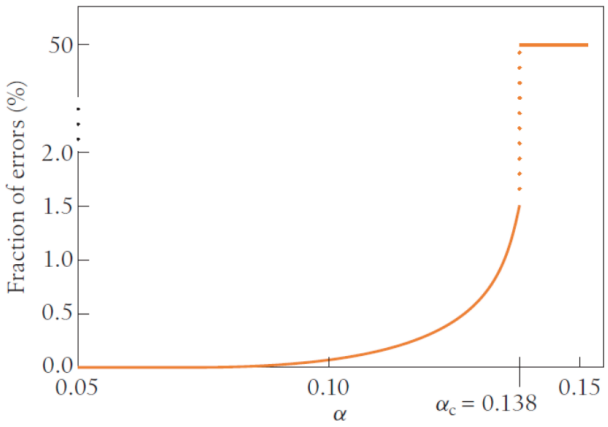


Figure 1.4 Average percentage of errors, N_e/N, in the retrieval states, as a function of α at $T = 0$. At $\alpha_c = 0.138$ (dashed curve) there is a jump from 1.57% to 50% errors. The figure is taken form [26].

1.7 Phase diagram

The solutions to RS equations in Eqs. (1.65, 1.66, 1.67) describe the different phases of the Hopfield model shown in the $(T - \alpha)$ phase diagram reported in Figure 1.5. We identify three primary phases:

1. **Paramagnetic Phase (PM):** At high temperature $T > T_g(\alpha)$, thermal noise dominates. The spins fluctuate randomly $(m = 0, q = 0)$ resulting in an ergodic dynamics; no pattern retrieval is possible.

2. **Spin-Glass Phase (SG):** For $T_M(\alpha) < T < T_g(\alpha)$, the network fails to retrieve any pattern $m = 0$ but exhibits frozen disorder $q > 0$, meaning it is trapped in spurious minima unrelated to the stored memories. The system is non-ergodic and frozen in spurious states uncorrelated with the patterns.

3. **Metastable Retrieval Phase (M):** For $T_c(\alpha) < T < T_M(\alpha)$ and $\alpha < \alpha_c$, $2P$ ferromagnetic retrieval states $\boldsymbol{m} \neq 0$ appear as *metastable* states. They coexist with the more stable SG state, which remains the global minimum of the free energy for $0.051 < \alpha < 0.138$. Pattern retrieval is possible if the network is well-initialized.

4. **Ferromagnetic (Retrieval) Phase (F):** For a small number of patterns $\alpha < \alpha_c$ and low temperature $T < T_c(\alpha)$, the system can successfully retrieve memories. Here, $m > 0$ and $q > 0$. The ferromagnetic retrieval states become the *global minima* of the free energy.

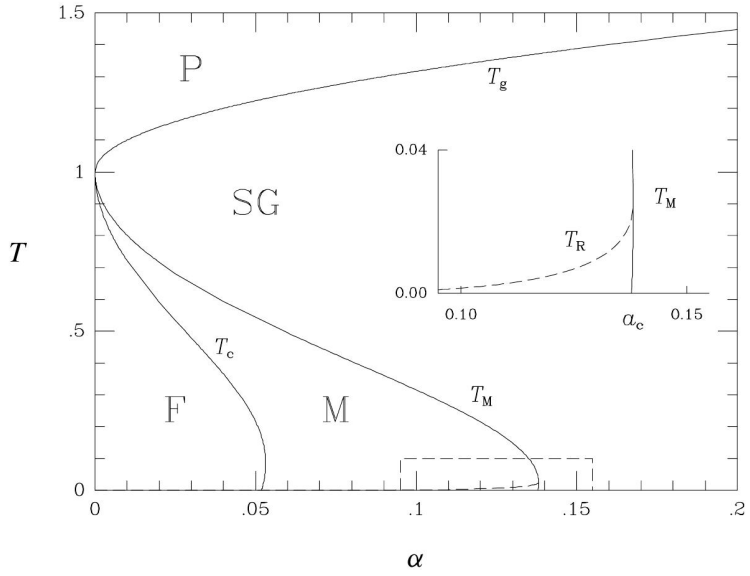


Figure 1.5 The phase diagram of the Hopfield model in the replica-symmetric approximation. The phases are characterized by the order parameters (m, q). PM: paramagnetic phase $(m = 0, q = 0)$; SG: spin-glass phase $(m = 0, q > 0)$; F: ferromagnetic (retrieval) phase $(m > 0, q > 0)$. The critical load is $\alpha_c \approx 0.138$. Inset: the AT instability for the retrieval solution $T_R(\alpha)$. Image taken from [33].

1.7.1 The Spin Glass Transition

The transition from the paramagnetic phase to the spin glass (SG) phase is a second-order transition. The critical temperature $T_g(\alpha)$ is found by expanding the expressions for q and r around $\boldsymbol{m} = 0$, leading to:

$$T_g(\alpha) = 1 + \sqrt{\alpha}. \tag{1.90}$$

For $T > T_g(\alpha)$, $q = 0$ and the network is paramagnetic. Below $T_g(\alpha)$, a spin glass solution emerges ($\boldsymbol{m} = 0$, $q > 0$, $r > 0$). The only parameter that distinguish the paramagnetic phase from this spin glass one is the overlap q that develops continuously from zero as T decreases below the ling T_g. These states exhibit frozen local disorder (hence $q > 0$) but have no macroscopic alignment with any stored pattern $\boldsymbol{m} = 0$, making them spurious states useless for associative memory. The system is non-ergodic but frustrated due to the combined effects of thermal noise and the high number of patterns.

This transition temperature is found by the expansion of the expression of q and r at $\boldsymbol{m} = 0$ using the relation $\tanh^2(x) \leq x^2$ and $0 \leq q \leq 1$ [15]. Since we have found from Eq. (1.44) that in the low load regime $\boldsymbol{m} = 0$ for $T > T_c = 1$, this SG state persists at least down to $T_c = 1$.

Linearization of Eqs. (1.65, 1.66) from small q and $\boldsymbol{m}$ shows the continuous bifurcations as T is decreased:

	at	from	to
$\alpha > 0$:	$T = 1 + \sqrt{\alpha}$	$\boldsymbol{m} = 0$, $q = 0$	$\boldsymbol{m} = 0$, $q > 0$
$\alpha = 0$:	$T = 1$	$\boldsymbol{m} = 0$, $q = 0$	$\boldsymbol{m} \neq 0$, $q > 0$

where the $\alpha = 0$ result corresponds to the one found in Sec (1.3.1). Moving down a vertical line $\alpha > 0.138$, the reduction of noise T does not bring about retrieval down to $T = 0$.

1.7.2 The Retrieval (Ferromagnetic) Phase and Metastability

For storage loads below the critical capacity, $\alpha < \alpha_c \approx 0.138$, a new type of solution emerges at a temperature $T_M(\alpha)$ lower than the spin glass transition $T_g(\alpha)$: the **ferromagnetic (F)** retrieval state, characterized by $\boldsymbol{m} \neq 0$ and $q \neq 0$. This phase exhibits a *first-order transition* from the spin glass (SG) phase, so the magnetization m jumps discontinuously to a non-zero value. The value of $T_M(\alpha)$ depends on α, and as the load decreases the system can retrieve at higher level of temperature T [13]. Just below $\alpha < \alpha_c = 0.138$ there are two types of attractors, those associated with one of the memorized pattern $\boldsymbol{\xi^\mu}$ and therefore with $m^\mu \simeq 1$ and a spin glass state with magnetization zero. The nature of the low-temperature energy landscape depends crucially on the value of α:

- **Coexistence and Metastability ($0.051 < \alpha < 0.138$):** Just below α_c, two types of attractors coexist: the $2P$ retrieval states (with $m^\mu \simeq 1$) and the spin glass state ($\boldsymbol{m} = 0$, $q > 0$). In this range, the spin glass state is the *global minimum* (lowest free energy), while the retrieval states are *metastable*. This defines the **metastable (M) phase**, bounded by the line $T_M(\alpha)$ in Figure 1.5. It is still a glassy phase (ferromagnetic states are metastable) but there is a possibility of retrieval if the state is initialized sufficiently close to a memory pattern.
  The retrieval overlap m goes abruptly from zero to a finite value, as soon as the line $T_M(\alpha)$ is crossed and is a decreasing function of α. It tends to zero as $\alpha \to 0$ matching the finite-P situation in Sec. 1.3 in which the magnetization vanish continuously at the transition temperature $T_c = 1$. The line $T_M(\alpha)$ is found numerically as the temperature at which the non-trivial $m \neq 0$ solution to the saddle-point equations ceases to exist for a given α. The behavior of the retrieval quality as a function of T fro several values of α is shown in Fig. 1.6(a).

- **Stable Retrieval:** As α decreases below approximately 0.051 at $T = 0$ and for $T < T_c(\alpha)$, the ferromagnetic retrieval states become the **global minima** (absolute minima) of the free energy. The spin glass state is no longer the most stable phase.

- **Appearance of Mixture States $T < T^{(3)}$:** At very low temperatures ($T < T^{(3)} \approx 0.461$), symmetric mixture states (e.g., 3-pattern mixtures) also become dynamically stable (metastable) attractors (see Sec. 1.3.1) and magnetization m_n shown in Fig. 1.6(b). However, *none* of these mixture states are ever global energy minima; the $2P$ pure ferromagnetic states remain the absolute minima of the landscape.

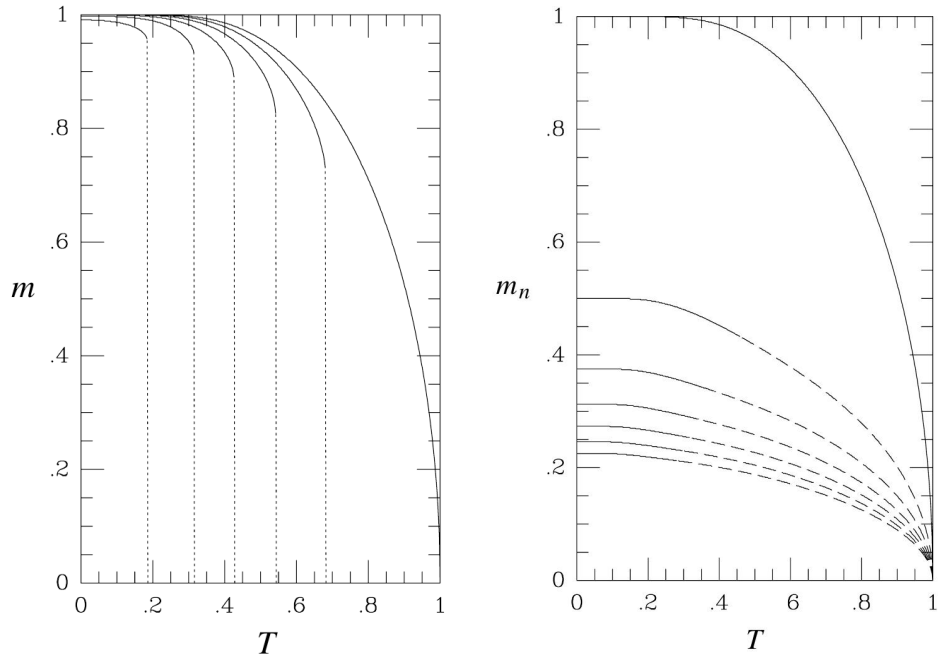


Figure 1.6 On the left (a): RS magnetization m of the pure states of the Hopfield model as a function of temperature T. From top to bottom: $\alpha = 0.000, ..., 0.125$ ($\Delta\alpha = 0.025$). Each curve ends at a value of m equal to the discontinuity at $T_M(\alpha)$. On the right (b): magnetization m_n of the mixture states presented in Sec. 1.3.1 as a function of temperature T. From top to bottom: $n = 1, 3, 5, 7, 9, 11, 13$. Solid: region where they are stable, local minima of $f(m)$. Dashed: region where they are unstable. Images taken from [33].

1.7.3 Replica Symmetry Breaking and its Effects

As in the case of the SK spin glass model [128], the replica-symmetric solution becomes unstable at low temperatures, as indicated by the emergence of negative entropies. The RS ansatz non longer corresponds to the correct physical saddle point and RS must be broken. This instability is analyzed via the *de Almeida-Thouless (dAT)* stability criterion [39], which involves studying small fluctuations (so-called 'replicon') of the free energy $f(\boldsymbol{m}, q, r)$ in Eq. (A.41) around the RS solution. The line $T_R(\alpha)$ in the phase diagram in Fig. 1.5 marks the boundary where the RS solution becomes unstable. AT-instability locates where the RS solution no longer corresponds to the correct extremal point for the free energy and indicates a second-order transition to a spin glass state where ergodicity is broken in the sense that the distribution $\overline{P(q)}$ is no longer a δ-function. In the paramagnetic regime ($\boldsymbol{m} = q = 0$) the AT condition reduces precisely to $T > T_g$ with T_g evaluated in Eq. (1.90): this means that the paramagnetic solution is stable. Numerical evaluation shows that the retrieval solution $\boldsymbol{m} \neq 0$ in unstable only for very low temperatures $T < T_R$ as visible from the inset in Fig. 1.5. For most of the interval $0 < \alpha < \alpha_c$ the FM phase is stable against RSB in a finite range of temperatures below $T_M(\alpha)$ ($T_M(\alpha) > T_R(\alpha)$ for $0 < \alpha < 0.137$). This supports the assertion that RSB effects are relatively weak [15]. Introducing Replica Symmetry Breaking, with replica symmetry broken

once (1RSB), has two major consequences:

1. It slightly increases the critical capacity from $\alpha_c^{\mathrm{RS}} = 0.138$ to $\alpha_c^{\mathrm{RSB}} \approx 0.144$ [35].
2. It improves the quality of retrieval; for example, at α_c, the magnetization m increases from 0.967 (RS) to 0.983 (RSB) [35].

For $0.138 < \alpha \leq 0.144$, retrieval states exist *only* within the RSB framework. At $\alpha = 0.145$ and beyond no one-step RSB solution with finite m is found.
For details and explicit calculation we remind to [15, 35]. Another rigorous method to detect the instability of the replica-symmetric solution, which does not require a full replica analysis was developed in [8].

1.8 Take home message: Properties and Limitations of the Hopfield Model

In this Chapter we presented a complete analytical treatment of the Hopfield model [73] for associative memory from its biological inspiration [94, 88, 113] to its analysis using the tools of statistical physics [12, 13, 14].
In this last section we want to underline the main characteristics and result of the model.

- **Theoretical Clarity:** The Hopfield model is a fundamental bridge connecting neuroscience, computer science, and statistical mechanics since it is formally equivalent to an Ising spin system with long-range, structured interactions. It provides a mathematically tractable framework for understanding how collective computation and emergent memory properties can arise from a network of simple, interacting units. The analytical mean-field and replica calculation consolidating the Hopfield network's role as both a historical model and a quantitative baseline for studying neural memory.
- **Memory as Attractors:** The Hopfield network encodes P random binary patterns as fixed-point attractors via the Hebbian learning rule (Eq. 1.22), with deterministic, zero-temperature dynamics (Eq. 1.5) designed to retrieve patterns robustly from degraded inputs.
- **Energy Landscape and Dynamics:** The core of the model's operation is the existence of a Lyapunov function—the energy landscape defined by the Hamiltonian $\mathcal{H}(\boldsymbol{\sigma})$ (Eq. 1.26). This landscape, defined by the Hebbian learning rule, guarantees convergence to fixed-point attractors, which are the physical representation of stored memory patterns.
- **Order parameter:** The fundamental order parameter is the Mattis magnetization $m_\mu(\boldsymbol{\sigma})$ (Eq. 1.27) which quantify the overlap between the network state and the stored patterns $\boldsymbol{\xi}^\mu$. The Hamiltonian can be expressed in terms of these overlaps (Eq. 1.28), explicitly linking energy minima to pattern retrieval.
- **The System Regimes:** The model's performance is governed by the load parameter $\alpha = \lim_{N\to\infty} P/N$ (Eq. 1.32). This reveals a fundamental trade-off:

  - **Vanishing load** $\alpha \to 0$ (Sec. 1.3): The network functions as a robust, high-quality associative memory. The stored patterns are stable attractors with large basins of attraction. For $\alpha \to 0$ (finite P), a mean-field analysis yields a continuous ferromagnetic transition at $T_c = 1$. Below T_c, the $2P$ pure states are the only stable attractors. Spurious states, symmetric mixtures of n patterns (Sec. 1.3.1), emerge at lower temperatures $T^{(n)}$ as local minima.
  - **High Load** $\alpha > 0$ (Sec. 1.4): Pattern interference, or "crosstalk noise," becomes significant. This leads to a finite **critical capacity** α_c [15], beyond which retrieval qualitatively breaks down. For $\alpha > 0$, the system exhibits a rich phase diagram in the (T, α) plane (Fig. 1.5), derived via the replica method under the Replica Symmetric (RS) ansatz.

- **Quantitative Storage Capacity:** The maximum number of retrievable patterns at $T = 0$ scales as $P_c = \alpha_c N$ with $\alpha_c \approx 0.138$ [15] within the Replica Symmetric (RS) theory. This transition is first-order, with retrieval quality m dropping discontinuously at α_c. For $\alpha > \alpha_c$, overlaps drop from $m \approx 0.97$ to $m \approx 0.35$ and retrieval ceases, signaling a catastrophic failure of the model as associative memory.

- **Signal-to-Noise Principle:** Retrieval succeeds as long as crosstalk noise from other patterns remains below the signal. The stability of a memory pattern can be understood intuitively through a zero-temperature signal-to-noise analysis (Sec. 1.5). The "signal" reinforcing a pattern is overwhelmed by Gaussian "noise" from other patterns when P exceeds $\alpha_c N$, leading to a catastrophic failure of memory retrieval—a "black-out" scenario. The stability condition $\xi_i^1 h_i > 0$ for all i leads to the conclusion that the average number of errors remains negligible ($< 2\%$) only for $\alpha < \alpha_c$ so the signal-to-noise analysis confirms the existence of a critical capacity.

- **The Role of Replica Theory and RSB:** The replica method is essential for analytically solving the model in the high-load regime (Sec. 1.4). The Replica Symmetric solution is stable for most of the retrieval phase (Fig. 1.5) but becomes unstable at very low temperatures. The stability of the RS solution is governed by the de Almeida-Thouless (AT) condition. Incorporating Replica Symmetry Breaking (RSB) (Sec. 1.7.3) refines the results, slightly increasing the critical capacity to $\alpha_c^{\text{1RSB}} \approx 0.144$ and improving retrieval quality near the limit [35].

- **Phase Structure:** The equilibrium phase diagram in the (T, α) plane shows the network operates in several regimes:

  - **Paramagnetic Phase**: Spins fluctuate randomly for $T > T_g(\alpha)$ (Eq. 1.90), no retrieval is possible. The line $T_g(\alpha)$ marks a *second-order* transition from the high-temperature paramagnetic (PM) phase ($\boldsymbol{m} = 0$, $q = 0$) to the spin glass (SG) phase ($\boldsymbol{m} = 0$, $q > 0$).
  - **Spin-Glass Phase** ($\boldsymbol{m} = 0, q > 0$): For $T_M(\alpha) < T < T_g(\alpha)$, the network is dominated by spurious minima with frozen disorder ($q > 0$)

and uncorrelated with the patterns ($m = 0$), losing memory capacity. The line $T_M(\alpha)$ signifies the *first-order* transition where retrieval states ($\boldsymbol{m} \neq 0$) appear. Above this line, no retrieval is possible and it terminates at $\alpha_c = 0.138$, the critical load beyond which retrieval states cease to exist.

  - **Metastable (M) Phase**: For $T < T_M(\alpha)$, retrieval states co-exist with spin-glass minima, and memory retrieval is possible but not globally stable.
  - **Retrieval / Ferromagnetic Phase** ($\boldsymbol{m} > 0, q > 0$): Below a certain line $T_c(\alpha)$ the $2P$ ferromagnetic states become the *global minima* of the free energy, are stable, the basins of attraction are well-defined, and retrieval occurs even from corrupted initial conditions.
  - **Spurious Mixture Attractors** $T^{(n)}(\alpha)$: Below $T^{(n)}(\alpha)$, spurious mixture states become additional metastable attractors, further complicating the energy landscape. The first are the symmetric 3-mixtures and decreasing T all the other odd mixtures appear. Anyway, the $2P$ ferromagnetic states remain the global minima of the energy.

- **Main Limitation:** The fundamental limitation of the Hopfield model is set by the exponential growth in interference as P increases, resulting in a strict bound on associative memory capacity far below theoretical maximum for a network of N neurons.

In summary, the Hopfield model offers a complete and analytically tractable illustration of the interplay between structure, dynamics, and disorder in neural networks. Its great power lies in its tractability: it is an exactly solvable framework that make possible the creation a precise phase diagram. This allows us to make rigorous statements about capacity limits and retrieval stability.
Paradoxically, the Hebbian rule (Eq. 1.22)—the source of the model's content-addressability—is also the origin of its fundamental limitation. The analysis rigorously establishes that robust retrieval is possible only for a load $\alpha = P/N$ below a critical threshold of $\alpha_c \simeq 0.138$. Beyond this point, the network's functionality does not gradually degrade but shatters at a sharp critical transition, plunging into a spin glass phase dominated by spurious states. This sharp phase transition causes catastrophic functional collapse and represents theoretical price paid for dense, symmetric connectivity.
This strict capacity limit presents a significant challenge for employing the standard Hopfield model as a high-performance associative memory. It consequently motivates the search for alternative learning rules and algorithms designed to overcome the performance of the simple Hebbian prescription. This search was motivated by synergy between computational theory and neuroscience, particularly inspired by biological hypotheses on memory consolidation.
In the following chapter we explore historical strategies, such as unlearning, or *dreaming*, algorithms implemented to suppress spurious minima in an attempt to enhance the network's storage capacity. This discussion will provide the foundation for Chapter 3, where we introduce and analyze our approach to a dreaming procedure.

Chapter 2

Extending the Hopfield Model: The Role of Dreaming

Enabling human beings to work together to build the impossible.

– *"Talkin' Hawkin"*, S. Hawking. *The Endless River*, Pink Floyd.

In this Chapter we would like to give an overview on the main historical proposed extensions of the traditional Hopfield model, all formulated with the intent to improve the network's performance as an associative memory, particularly in the high-load regime $\alpha > \alpha_c$. The original formulations were related to the concept of dreaming; this process involves a strategic modification of the standard Hebbian learning rule (Eq. 1.22) to obtain better performances in terms of storage capacity. We begin with a discussion of the historical motivations for such procedure and the actual algorithms with their limitations. We review the literature on alternative learning rules, arguing that rules that avoided the pitfall of *dreaming too much* showed other undesirable properties (the most common ones being the addition of non-local terms and the vanishing of the basins of attraction). For some of them we provide a more detailed description in order to make ore clear the comparison with the specific algorithm adopted in this thesis exposed in Chapter 3.

2.1 Historical Motivation: Unlearning and Forgetting

Hopfield Networks [73] are among the most well-known models for storing and retrieving memories in a neural network. The original Hopfield model, based on the Hebb rule [72], is both analytically tractable [13, 15] and was formulated to be, at least originally, biologically plausible in its dynamics [11].
Although the number of patterns P that can be stored and efficiently retrieved in the Hopfield model is extensive, i.e. linear in the number of neurons N, the storage capacity (or maximum load) α_c is rather limited. Indeed, memories can be efficiently retrieved only if $\alpha = P/N < \alpha_c \simeq 0.138$ [15], that is the number of stored memories is just a small fraction of the number of neurons.
Given this limitation of the Hebb rule, it is crucial to find other rules that can improve the storage capacity of the Hopfield model. After the Hopfield model was introduced, research interest focused on the necessity of improving its capabilities to

enable its use as an associative memory, even in the high-load regime. Following the formulation of the Hopfield model, research efforts focused on overcoming its fundamental limitation: the catastrophic collapse of retrieval quality beyond a critical load $\alpha > \alpha_c$.
An original hypothesis was proposed by Crick and Mitchison, who suggested that biological dreaming might serve an *unlearning* function. They postulated that during the REM sleep phase, the brain weakens spurious, parasitic synaptic connections, thereby consolidating genuine memories and improving cognitive performance. In [34], they presented an hypothesis about the benefits of a dream procedure inspired by sleeping natural mechanisms. In this context, the Hebb rule can be interpreted as the *day* phase where the memories to be stored are encoded in the synapses' strengths (i.e. in the couplings), while the dreaming procedure can be interpreted as the *night* phase where spurious memories are erased. Within the Hopfield framework, spurious states refer to stable configurations of the network that do not correspond to any single stored pattern. These include, as already explained in Sec. 1.7:

- **Mixture States:** Symmetric combinations of n multiple patterns that have non zero overlap with more than one memory (e.g., $\boldsymbol{\sigma} \sim \text{sign}(\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2 + \boldsymbol{\xi}^3)$) which become stable below a certain temperature $T < T^{(n)}$.
- **Spin Glass States:** States with frozen disorder that are completely uncorrelated with any stored pattern.

Mixture states are stable only under a certain temperature $T < T^{(n)}$ so normally it is sufficient to work at a temperature greater than the highest of the critical temperatures of the mixed states present to eliminate their influence.
But while thermal noise at $T > 0$ can destabilize mixture states, spin glass states persist, in the high load regime, even at lower temperatures as exposed in Chapter 1. In the $T = 0$ simulations performed in this work, these spurious states are prevalent and significantly hinder pattern retrieval. Once memories are learned, spurious ones are also created and, therefore, can be recovered. In this case, an unlearning process can be applied that starting from a noisy input, improves the performance of the network allowing to reinforce real memories and minimizing the spurious ones.
An initial approach to manage capacity was followed in [111] imposing *hard constraints* on synaptic strengths by bounding the connection $|J_{ij}| < A$. Introducing a maximum value A for the entries of the J_{ij} matrix, old patterns are forgotten and the network can recalls only the most recent ones and avoid the phenomenon of confusion typical of the model at high load. Indeed, when the network tries to learn a number of patterns that exceed a threshold, it entries in a state of confusion and it is unable to learn any pattern. Bounding the J_{ij} allows to forget old patterns and learns others. The value of A is chosen in order to continue to have a storing capability proportional to N.
The update rule for the couplings J_{ij} with a bounding function $f(x)$ is [92]:

$$J_{ij}^{\text{new}} = f\left(J_{ij}^{\text{old}} + \frac{1}{\sqrt{N}}\xi_i^\mu \xi_j^\mu\right), \quad \text{where} \quad f(x) = \begin{cases} x & \text{for } -A < x < A, \\ A & \text{for } x \geq A, \\ -A & \text{for } x \leq -A. \end{cases} \tag{2.1}$$

This mechanism ensures the network only retains the most recently learned patterns, preventing the total confusion that occurs when P vastly exceeds P_c [13]. Numerical simulations have shown that, alternating between learning and unlearning, a network only stores and retrieves the most recently learned data and forgets the older ones. When a noisy pattern is presented to the network, it converges to the original memory since Hebbian unlearning eliminated undesirable correlations [143] improving the network performances. If the network is initialized close enough to a memory, the dynamics rule makes the spin configuration to converge to this attractor, reducing the number of misaligned spins. Note that also in [140] a combination of repeated learning and unlearning phases is described, but is suffers the same problem of forgetting older memories since the learning phase is applied only to new examples and it never reinforces old ones.
The approach [111, 92] described by Eq. (2.1) but also the one followed in [32, 121] are examples of local dreaming procedures that can be iterated indefinitely long, constantly learning new examples at the cost of forgetting the earlier ones, which means that these methods does not improve the capacity of the Hebb rule. Other iterative learning rules are presented in the following section.

Iterative learning rules

The locality of the learning rule is a key property to make the rule biologically plausible and efficient to implement. Several modifications of the original dreaming procedure have been developed over the years to impose such a locality property. Unfortunately, they maintain the locality at the cost of other problems. An iterative and local rule that can be seen as the expansion of the Pseudo Inverse rule is the Storkey rule [131], presented in details in Sec. 2.4, that was shown [132] to realize a trade-off between the large basins of the Hebb rule and the high capacity of the Pseudo Inverse rule [115, 80], described in Sec. 2.5.
In [117, 118] the authors defined a dreaming procedure that uses local fields instead of spin configurations. This local rule converges to the Pseudo Inverse rule for a large number of iterations [80], which means that it has the same problem of vanishing basins of attraction. Moreover, in [104, 103] it was shown that even by modifying the rule by subtracting states sampled from the Gibbs measure at high temperature, the learning dynamics converge to a coupling matrix of the pseudo-inverse family. An example of an iterative but non-local rule can be found in [23], where the authors subtract the largest eigenvector of the coupling matrix at each step of dreaming. They find that this procedure has the same performance as the classical unlearning, but has the great advantage of being much more transparent and analytically accessible.
Recently, more advanced dreaming variations, inspired by the perceptron rule, have been studied in [25, 24]. They seem to achieve optimal results but still require the fine-tuning of some parameters (similarly to the fine-tuning of the number of dreaming iterations). Furthermore, in [5], the authors explicitly linked the number of dreaming iterations to regularization hyperparameters, interpreting the whole model through the lens of machine learning. They also discuss ways to fix these hyperparameters a priori.
We will describe the procedure adopted in [25] as example for the Hebbian Unlearning

and compare their results with the ones obtained using our approach discussed in the next Chapter 3.

Learning rules with analytic expressions

Several strategies to increase the storage capacity avoid completely the use of the iterative learning procedure and rely on analytical expressions for the coupling matrix $\mathbf{J}$, in the same spirit as the original Hebb rule.
The pseudo-inverse rule [80, 115], exposed in Sec. 2.5, was introduced to deal with correlated data: it introduces a non-local term that makes it possible to store all the examples without error, regardless of their correlation. It is non-local in the sense that it requires the inversion of the correlation matrix, a task that cannot be done locally on a single neuron, and requires the intervention of a *deus ex machina* that knows the status of the entire network. Another problem is that this rule requires looking at all the training examples at the same time, meaning that if we want to store an additional example we should repeat the procedure from scratch. This rule reaches the maximal theoretical capacity of symmetric networks, namely $\alpha_c^{PI} = 1$ [59], but it presents a severe drawback: as $\alpha \to 1$ the radius of the basins of attraction goes to zero, making retrieval possible only in the trivial case of a noiseless initial condition.
In [47, 48], the authors defined a learning rule that is an interpolation between the Hebb and the pseudo-inverse rule. This rule can also be seen as the continuous limit of a certain dreaming procedure: the idea is to stop before the limit where basins of attraction vanish, in the same spirit as stopping the dreaming iterations. This rule also has a local-update version that is inspired by the learning with noise strategy suggested by Gardner [62]. Note that this rule needs to be modified in case of correlated examples [43]. Another interesting case of correlated examples is [71], where the authors consider examples correlated within groups. They show that, for this specific setting, unlearning mixtures of correlated examples is an effective strategy to make the examples retrievable.
In [53], the authors deal with the problem of dreaming indefinitely long by adding a reinforcement term in the energy. They have a parameter that interpolates between the Hebb rule and a pseudo-inverse-like rule, in the same spirit as [48, 47], but it is not known if it is possible to converge to this rule with a local iterative update of the coupling matrix. The answer to this question was our initial motivation to the formulation of the algorithm elaborated in Chapter 3.

Multi-spin Interactions

One recent strategy involves changing the nature of the interactions by considering many body terms, multi-spin interactions. This approach significantly enhances the capacity, making it more than linear [58, 84, 2], and in some cases even exponential [42, 120, 90]. We give some details about the Modern Hopfield Network (MHN) to make clear the passage from pairwise to multi-spin interactions.
The **Modern Hopfield Network (MHN)**, also known as a Dense Associative Memory (DAM), was introduced by Krotov and Hopfield in 2016 [84]. It represents a generalization of the classical Hopfield model, extending its pairwise (quadratic)

interactions to high-order synaptic couplings. This extension enables the network to store and retrieve a significantly larger number of patterns, effectively overcoming the classical capacity limitations. In the standard Hopfield model, the energy function is quadratic in the neuron activation's (Eq. 1.26); in contrast, the Modern Hopfield Network replaces this quadratic form with a nonlinear interaction potential $F(x)$, a nonlinear activation function determining the "order" of the effective interaction. The resulting energy energy function is:

$$E(\boldsymbol{\sigma}) = -\sum_{\mu=1}^{M} F\left(\sum_{i=1}^{N} \xi_i^\mu \sigma_i\right). \tag{2.2}$$

For $F(x) = \frac{1}{2}x^2$ the model reduces to the classical Hopfield network with pairwise interactions. Higher-order choices of $F(x)$ such as $F(x) \propto x^n$ with $n > 2$, introduce polynomial or exponential nonlinearity, effectively generating high-order coupling terms among neurons. With appropriate nonlinearities, the storage capacity scales as a polynomial or even exponential function of the system size N, surpassing the classical limit $\alpha_c \simeq 0.138$ since $M \sim N^{n-1}$ [84]. The nonlinear potential $F(x)$ creates a denser set of stable minima (memories), allowing finer discrimination among patterns and by tuning the nonlinearity it is possible to interpolate smoothly between the classical Hopfield model and high-order associative memories. However the spin dynamics is non-local as can be seen form the update rule:

$$\sigma_i(t+1) = \operatorname{sgn}\left(\sum_{\mu=1}^{M}\left(F\left(\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j(t)\right) - F\left(-\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j(t)\right)\right)\right). \tag{2.3}$$

While these strategies have demonstrated powerful applications, we will not discuss them in this work, as we are primarily interested in maintaining the Hebbian-like pairwise interactions and implement an algorithm based on the unlearning point of view that is based solely on the patterns to be stored.

2.2 Hebbian Unlearning: A Framework For Dreaming

Hopfield et al. [74] translated the biological unlearning concept into an algorithmic procedure. Noting that the network dynamics often converge to these spurious minima, they proposed to *unlearn* these states by applying an anti-Hebbian rule: the more neurons are correlated across spurious states, the more their connection will be weakened by the algorithm. The unlearning mechanism [81, 141, 142, 140, 143] can be seen essentially as the same of the learning one but with reversed sign since it has to remove these spurious states. As outlined in [74], a single dreaming iteration consists of three steps:

1. **Initialization:** The network is initialized with neuronal states $\sigma_i = \pm 1$ chosen randomly and the connectivity matrix is initializing according to the Hebb's rule ($\mathbf{J}^{(0)} = \mathbf{J}^H$).

2. **Relaxation:** The network evolves under its asynchronous dynamics until it converges to a stable fixed point $\boldsymbol{\sigma}^*$:

$$\sigma_i^* = \operatorname{sgn}\left(\sum_j J_{ij}\sigma_j^*\right) = \operatorname{sgn}(h_i). \tag{2.4}$$

3. **Synaptic Modification:** The coupling matrix $\mathbf{J}$ is updated based on the fixed point $\boldsymbol{\sigma}^*$ or the local field $\boldsymbol{h}$ encountered. In this case [142], the local field h_i for the random initial configuration is calculated and then iteratively modified in order to upgrade the couplings. We obtain respectively:

$$\delta J_{ij}^{(d)} = -\frac{\lambda}{N}\sigma_i^*\sigma_j^* \quad \text{or} \quad \delta J_{ij}^{(d)} = -\frac{\lambda}{N}h_i h_j, \quad J_{ii} = 0 \ \forall i \tag{2.5}$$

where λ is the unlearning strength parameter that must be tuned and d the discrete algorithm time.

Iterating this process selectively destabilizes spurious attractors, effectively "cleaning" the energy landscape and improving the basins of attraction for the genuine memories. This algorithm was first introduced to prune the landscape of attractors from proliferating spurious states, i.e. fixed points of Eq. (1.7) not coinciding with the memories [57]. The common thread among these methods is the dreaming procedure, a post-learning algorithm designed to optimize the synaptic matrix $\mathbf{J}$ and the specific implementation of Step 3 defines the variant of dreaming. Despite characteristic differences, the overarching effect of any of the dreaming procedure is to smooth the energy landscape. It reinforces the basins of attraction around the genuine memory patterns (global minima) while destabilizing spurious local minima (mixture and spin glass states). This is because each memory that is a global minima, is surrounded by many spurious local minima that have to be removed.
However, most versions of the dreaming iterative procedure studied so far in the literature encounter significant challenges. Indeed, repeating the unlearning steps excessively often leads to a point where, depending on the load α, even desired memories start deteriorating. This phenomenon results in *catastrophic forgetting*, where eventually no memory can be retrieved anymore unless the number of dreaming iterations is fine-tuned [141, 142]. Additionally, the known dreaming procedure struggle to handle efficiently correlated examples [140]. In Chapter 3 we will provide a solution for these issues introducing a new unlearning algorithm that we call *Daydreaming*.

The Hebbian Unlearning is an *unsupervised* algorithm, as it does not require explicit knowledge of the memory patterns $\{\vec{\xi}^\mu\}_{\mu=1}^P$. Instead, it operates only with the information already encoded in the Hebbian-initialized synaptic matrix (Eq. 1.22). The total number of unlearning iterations, D, is a crucial meta-parameter of the algorithm. Its optimal value must be chosen to maximize retrieval performance at a given load $\alpha = P/N$. An insufficient number of iterations (D too small) leaves spurious states insufficiently suppressed, while excessive unlearning (D too large) can degrade the stability of the genuine memory attractors themselves, leading to a catastrophic loss of storage capacity.
As a representative example of Hebbian Unlearning algorithm, we report in this manuscript the procedure exposed in [25] that can be considered the current state-of-the-art procedure.
In Chapter 3 we will introduce the Daydreaming algorithm and a performance comparison reveals a clear advantage for our procedure over this one. Specifically, Daydreaming achieves superior memory capacity and retrieval robustness across a wider range of network loads α, establishing it as a more effective training procedure

for recurrent neural networks.
The performance of the algorithm in [25, 75] has been studied in terms of the stabilities Δ_i^μ defined as:

$$\Delta_i^\mu = \frac{\xi_i^\mu}{\sqrt{N}\sigma_i} \sum_{j=1}^{N} J_{ij}\xi_j^\mu, \quad \sigma_i = \sqrt{\sum_{j=1}^{N} \frac{J_{ij}^2}{N}}. \tag{2.6}$$

In Fig. 2.1a, we report the behavior of the minimum, the average and the maximum stability $(\Delta_{\min}, \Delta_{\text{av}}, \Delta_{\max})$ as a function of the number of unlearning steps d rescaled by a factor λ/N. The minimal stability grows to positive values and peaks at $d = D_{\text{top}}$; between D_{in} and D_{fin} every stability is positive or, equivalently, every memory is a fixed point. As they increase α, the interval $[D_{\text{in}}, D_{\text{fin}}]$ shrinks, and the height of the peak at D_{top} lowers until a critical load [25]

$$\alpha_C^{HU} = 0.589 \pm 0.03 \tag{2.7}$$

above which Δ_{min} never reaches positive values.
The authors test the iterative unlearning algorithm, using an optimal number of steps $D = \mathcal{O}(N/\lambda)$, by measuring the shape of the basins of attraction around each memory through the retrieval map that we will describe in Sec. 3.3. The resulting optimal retrieval map for a load $\alpha = 0.4$, exceeding the critical capacity of the standard Hopfield model $\alpha_c = 0.138$, is shown in Fig. 2.1b. Parameters were $N = 800$ and unlearning strength $\lambda = 10^{-2}$. The optimal curve is obtained at $d = D_{\text{in}}$. The graph also shows retrieval maps for the symmetric perceptron (SP), that we will describe in Sec. 2.3 and the original Hopfield model, which fails to retrieve memories as expected.
This figure serves as a comparison with our proposed algorithm (Sec. 3.5) where we show the retrieval map for the same load α obtained using the Daydreaming algorithm; in particular, the red curve of Fig 2.1b is compared with the last asymptotic retrieval of Fig. 3.4. The comparison demonstrates the improvement in retrieval performance achieved by our method with respect to the Hebbian Unlearning and the Symmetric Perceptron at maximum margin $k \simeq k_{\max}$.

2.3 Symmetric Perceptron

The linear perceptron algorithm [59, 98, 102, 83, 52, 145, 16] is an iterative procedure for fully stabilizing memories while tuning their robustness capabilities. In this architecture, an N-dimensional input layer is fully connected via synaptic weights $\mathbf{J}$ to an N-dimensional output layer and we want to find a set fo couplings that satisfy the constraints

$$\Delta_i^\mu > k, \quad \forall \mu, i \tag{2.8}$$

with k the *margin* parameter. An elegant interpretation of this problem reformulates it as a linear regression in the space of coupling matrices $\mathbf{J}$ [98]. From the perspective of network dynamics, the parameter $k \geq 0$ directly controls memory robustness: larger values of k correspond to greater stability against dynamical perturbations. Specifically, for a given loading capacity α, all memory patterns remain stable up to a maximum robustness threshold $k_{\max}(\alpha)$. The maximum capacity achievable by

the network is $\alpha_c^P = 2$ such that $k_{\max}(\alpha_c^P) = 0$. The constraints in Eq. (2.8) will be satisfied after a number of iterations of the following serial update for the couplings

$$\delta J_{ij}^{(d)} = +\lambda \sum_{\mu=1}^{P} \epsilon_i^\mu(d)\xi_i^\mu\xi_j^\mu, \qquad J_{ii} = 0, \quad \lambda > 0 \tag{2.9}$$

$$\epsilon_i^\mu(d) = \frac{1}{2}(1 + \mathrm{sgn}(k - \Delta_i^\mu(d))) = \theta(k - \Delta_i^\mu(d)) \tag{2.10}$$

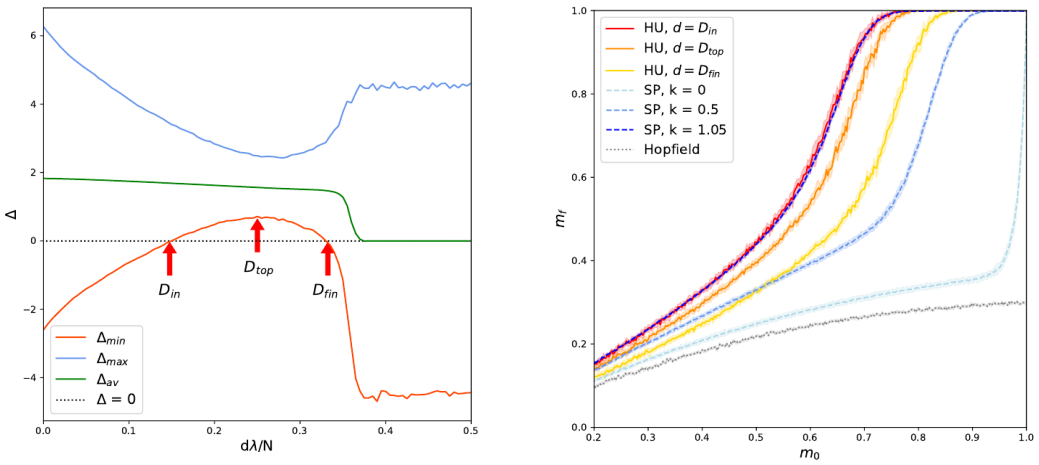
where λ is the learning rate of the perceptron and d again the discrete algorithm time. The standard perceptron learning rule, is inherently asymmetric because the update condition $\epsilon_i^\mu = \theta(k - \Delta_i^\mu)$ is computed solely from the perspective of the *postsynaptic* neuron i. This results in a directed update where J_{ij} is modified to correct the stability of neuron i, without regard for the stability requirements of neuron j. To enforce the symmetry condition $J_{ij} = J_{ji}$ required for energy-based models, the update rule must be identical for both J_{ij} and J_{ji}. The most natural approach is to define a symmetric update condition ϵ_{ij}^μ that incorporates information from both neurons:

$$\epsilon_i^\mu(d) \to \epsilon_{ij}^\mu(d) = \frac{1}{2}(\epsilon_i^\mu(d) + \epsilon_j^\mu(d)). \tag{2.11}$$

This leads to the learning rule of the Symmetric Percetron (SP):

$$\delta J_{ij}^{(d)} = \frac{\lambda}{2}(\epsilon_i^\mu(d) + \epsilon_j^\mu(d))\xi_i^\mu\xi_j^\mu. \tag{2.12}$$

Since $\epsilon_{ij}^\mu = \epsilon_{ji}^\mu$ by construction, the update satisfies $\delta J_{ij} = \delta J_{ji}$, guaranteeing that the coupling matrix $\mathbf{J}$ remains symmetric throughout the learning process when initialized symmetrically. The perceptron algorithm is *supervised*, as it requires explicit knowledge of the memory patterns $\{\vec{\xi}^\mu\}_{\mu=1}^P$ to perform the coupling updates.



(a) Stability evolution during unlearning

(b) Retrieval maps comparison

Figure 2.1 Hebbian Unlearning: Main Results of [25]. (a) Evolution of stabilities during unlearning. The minimal stability peaks at $d = D_{\text{top}}$, with all memories stable between D_{in} and D_{fin}. The stability window shrinks with increasing α until the critical capacity $\alpha_C^{HU} = 0.589 \pm 0.03$. (b) Retrieval maps for $\alpha = 0.4$, $N = 800$, $\lambda = 10^{-2}$, showing performance at different unlearning steps $(D_{\text{in}}, D_{\text{top}}, D_{\text{fin}})$, SP with varying margins k, and the original Hopfield model. Optimal retrieval occurs at D_{in} which coincides with the SP algorithm at $k \simeq k_{\max}$.

2.4 Storkey Rule

Another example of iterative and local learning algorithm is represented by the Storkey rule [131] that realizes an effective trade-off between the extensive basins of attraction characteristic of the Hebbian rule and the high storage capacity of the pseudo-inverse rule [132]. The key idea is to subtract out the influence of the current pattern on all the other patterns that have already been stored. This acts as a form of "local correction" that minimizes interference as you learn. The learning update at step t is defined as[1]:

$$J_{ij}^{t} = J_{ij}^{t-1} + \frac{1}{N}\xi_i^t\xi_j^t - \frac{1}{N}\xi_i^t h_{ji}^t - \frac{1}{N}h_{ij}^t\xi_j^t \tag{2.13}$$

where the local field h_{ij}^u is computed as:

$$h_{ij}^u = \sum_{k=1,k\neq i,j}^{N} J_{ik}^{u-1}\xi_k^u. \tag{2.14}$$

This represents the local field at neuron i excluding both self-connections ($k \neq i$) and the specific connection to neuron j ($k \neq j$), with ξ^t being the new pattern to be learned. The first two terms of Eq. (2.13) correspond to the standard Hebbian increment while the subsequent terms serve as corrective components that actively suppress interference from previously stored patterns since subtract the mutual predictions that neurons i and j have for each other based on their interactions with the rest of the network. The Storkey rule thus implements a form of "forgetting" that is proportional to the interference caused by new learning, allowing for incremental storage of patterns while dynamically managing cross-talk between memories. This results in both enhanced storage capacity and improved basin stability compared to pure Hebbian learning.

2.5 Pseudo-Inverse Rule

The Pseudo-Inverse rule [115, 80] represents a non-local learning algorithm that maximizes the storage capacity of associative memory networks, allowing for the exact retrieval of all stored patterns as stable fixed points. Unlike the Hebbian and Storkey rules, which are local and incremental, the pseudo-inverse rule requires global information about the entire set of patterns to be learned. The synaptic matrix is defined as:

$$J_{ij} = \frac{1}{N}\sum_{\mu,\nu=1}^{P}\xi_i^\mu (C^{-1})_{\mu\nu}\xi_j^\nu, \qquad C_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\xi_i^\nu \tag{2.15}$$

with $\mathbf{C}$ the $P \times P$ correlation matrix of the stored patterns. Each synaptic weight J_{ij} depends on all stored patterns, through $\mathbf{C}^{-1}$; thus, the rule cannot be implemented incrementally. This construction guarantees that each stored pattern is a fixed point of the network dynamics (Eq. 1.21), so the rule achieves theoretical maximum storage capacity of the Hopfield network, corresponding to $\alpha_c^{PI} = 1$: in other words, up to N

[1]The expression in Eq. (2.13) refers to the First Order Storkey Rule that is the formulation that will be compared with Daydreaming results in Sec. 3.5.2.

random and uncorrelated patterns can be stored and retrieved without errors. Since the rule explicitly decorrelates patterns, interference among memories is eliminated in the ideal noiseless limit. This rule is computationally challenging as it requires global access to all stored patterns and inversion of the full correlation matrix. The Pseudo Inverse rule thus provides an ideal reference model for associative memory performance: it achieves perfect recall and maximal storage capacity but at the cost of losing computational efficiency at large scale systems.

2.6 The Daydreaming Algorithm

Our work, detailed in Chapter 3, is inspired on the concept of reinforcing the memories [53]. We design a new iterative learning procedure that solves the major problems that limited unlearning algorithms so far. First, it can be iterated indefinitely, thus avoiding fine-tuning the early stopping. Moreover, it operates independently of any assumptions regarding the structure and correlation in the examples to be stored (at variance with methods like [43, 71]. Finally, the procedure only depends on a single parameter and keeps the conceptual simplicity of the original dreaming procedure. We call it *Daydreaming*, as it drops any distinction between night and day cycles. The key features of our learning procedure are:

1. We sample states to be unlearned as in the original procedure [74], by initializing randomly the model and converging to a spurious state;
2. We reinforce *multiple times* each example that we want to store. This is at variance with many examples from the literature (see for example [140, 32, 121]), where they reinforce *new* memories and forget the old ones, in the same spirit as [111, 92];
3. We find the right balance between these two operations, which consist in reinforcing a randomly chosen example each time we sample a spurious state to be unlearned.

It is worth stressing that the Daydreaming procedure is fully local, that is the learning of the coupling J_{ij} only depends on the values of neurons σ_i and σ_j (as well as the values of the patterns in i and j). This is an important observation, given that any biologically plausible learning procedure must be local, while methods based on the inversion of the pattern correlation matrix are not local [115, 80]. Note that the way in which we combine learning and unlearning is reminiscent of Hebbian and anti-Hebbian terms in moment-matching algorithms, with important differences in the sampling. We will discuss them in Sec. 3.9.2. We have reviewed in Sec. 2.2-2.4-2.5 three significative rules implemented in order to improve the storage capacity α_c of the Hopfield model that we will use as a comparison with our results with the Daydreaming algorithm in Chapter 3. For the Hebbian Unlearning, as said, we compare the retrieval map of uncorrelated pattern with the one obtained in [25] in Sec. 3.5 in Fig. 3.4; the comparison with our algorithm with the Storkey and Pseudo-Inverse rules is done in terms of basin of attraction of the network as exposed in Sec. 3.5.2 in Fig. 3.7.

Chapter 3

Daydreaming

Mankind's greatest achievements have come about by talking.

– *"Talkin' Hawkin"*, S. Hawking. *The Endless River*, Pink Floyd.

The severe storage limitation of the standard Hopfield model motivates the development of enhanced learning algorithms. In this Chapter, we address this challenge by developing the *Daydreaming* algorithm [127], a new procedure designed to overcome this constraint. Our approach implements a continuous, non-destructive process that simultaneously reinforces the original stored patterns (mimicking Hebbian consolidation) and suppresses spurious memories. This procedure converges asymptotically to stationary retrieval maps defined in Sec. 3.3.
The Chapter is structured as follows: first, we present the structure of the algorithm and it asymptotic convergence properties, see respectively Sec. 3.2 and Sec. 3.4. In Sec. 3.5 we demonstrate that when trained on random uncorrelated patterns, the Daydreaming algorithm achieves optimal performance, maximizing the size of the basins of attraction, defined in Sec. 3.3, and the quality of pattern reconstruction. In Sec. 3.7 we show that the algorithm spontaneously exploits correlations in structured data, generated here by a random-features model, leading to even greater increase in storage capacity and basin stability, while also stabilizing the hidden features. Finally, in Sec. 3.8 we validate the model's performance on the MNIST dataset, showing its ability to generate attractors proximate to both unseen examples and class prototypes.

3.1 The Model and the Dynamics

The starting point of our research is represented by the Hopfield model, the main characteristics of which are presented in detail in Chapter 1. We provide here a brief summary of the key concepts relevant to this chapter.
A Hopfield network [73] is a recurrent neural network made of N neurons (or spin) $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{N}$, where each spin can be in a state of ± 1. The state of each neuron is determined by the signal it receives from all other neurons in the network. At each

time step t, every neuron is updated according to the dynamics rule

$$\sigma_i^{(t+1)} = \text{sign}\left(\sum_{j=1}^{N} J_{ij}\sigma_j^{(t)}\right) , \tag{3.1}$$

where J_{ij} are the elements of coupling matrix and represent the synaptic weights from neuron j to neuron i. They are defined through the Hebbian rule:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu . \tag{3.2}$$

The dynamics defined by Eq. (3.1) minimizes the following energy function

$$H(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i,j} J_{ij}\sigma_i\sigma_j \, . \tag{3.3}$$

The formulation of this Hamiltonian made it possible to solve the Hopfield model in the context of statistical physics of disordered systems [13, 14, 15] where memories are interpreted as local minima of the energy landscape. When P is proportional to N, this landscape is populated by many spurious minima that can trap the dynamics defined by Eq. (3.1) and thus stop the retrieval of a memory if the dynamics is initialized too far from it.
We consider a symmetric matrix $J_{ij} = J_{ji}$ with zeros on the diagonal $J_{ii} = 0$, which means we do not consider the so-called self-energies. We perform asynchronous updates (see Sec. 1.1 for details), meaning that, at each time step, we update one neuron at a time in an arbitrary order. The dynamics stop when spin values reach a fixed point (or, equivalently, a local minimum of the energy, that can be either a memory or a spurious one).
This model works as an associative memory if, when initialized on a noisy version of one of the P stored patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$, the model converges to the clean version of that example. Our objective is to find an efficient method to determine the synaptic weights $\mathbf{J}$ that maximize both the number of retrievable patterns and the average initial noise level, defined through the mean initial distance from the memory, from which correct retrieval is still possible.

3.2 Daydreaming algorithm

Our algorithm is designed to enhance the storage capacity of the Hopfield network by implementing a dual-process mechanism: it simultaneously reinforces the stored memories and "dreams away" spurious memories—defined as local minima of the Hamiltonian $H(\boldsymbol{\sigma})$ that do not correspond to any intended pattern $\boldsymbol{\xi}^\mu$. The removal part component operates on the same principle as the original dreaming procedure [74].
At each step u in the learning process, we initialize the network to a random configuration, and we run the asynchronous update dynamics in Eq. (3.1) until we reach a fixed point $\tilde{\boldsymbol{\sigma}}^{(u)}$. Then we perform a dual update to the coupling matrix $\mathbf{J}$:

- we reinforce a randomly selected memory pattern $\boldsymbol{\xi}^{\mu(u)}$ by lowering its energy adding $\xi_i^{\mu(u)}\xi_j^{\mu(u)}$,
- we penalize the spontaneously recalled fixed point $\tilde{\boldsymbol{\sigma}}^{(u)}$ by raising its energy through the addition of the term $-\tilde{\sigma}_i^{(u)}\tilde{\sigma}_j^{(u)}$ to the coupling matrix.

Formally, the Daydreaming update rule is given by:

$$J_{ij}^{(u+1)} = J_{ij}^{(u)} + \frac{1}{\tau N}\left(\xi_i^{\mu(u)}\xi_j^{\mu(u)} - \tilde{\sigma}_i^{(u)}\tilde{\sigma}_j^{(u)}\right) , \tag{3.4}$$

where τ is a timescale parameter that acts as an inverse learning rate and the factor $1/N$ ensures the update scale is appropriate for the large-N limit. For the same reason, we define one *epoch* of training as N consecutive Daydreaming steps. This provides a scale-invariant measure of training time since we have a measure of training time that is fair at any N; the time t that appears in all figures measures the number of epochs. Finally, although the update rule in Eq. (3.1) is invariant under a global rescaling of $\mathbf{J}$, we prefer to keep it well bounded, and so we normalize the coupling matrix after each epoch.
The learning rule in Eq. (3.4) requires an initial condition for $\mathbf{J}^{(0)}$. We have tested several uninformed initializations—such as $J_{ij} = 0$ or $J_{ij} \sim \mathcal{N}(0,1)$— and found that all converge to a set of coupling matrices with statistically equivalent retrieval properties. Just to make the learning process faster and expedite convergence, we initialize J_{ij} according to the Hebbian rule in Eq. (3.2). We report the complete pseudo-code in the following Algorithm 1.

Algorithm 1 Daydreaming learning algorithm

Input examples $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$
Output coupling matrix J

```
J_ij ← (1/N) Σ_μ ξ_i^μ ξ_j^μ          ▷ Initialization with the Hebb rule
J_ii ← 0
for t = 1, ..., E do                    ▷ Do E epochs
    for u = 1, ..., N do                ▷ Do N steps in each epoch
        μ ← Unif({1, ..., P})           ▷ Pick an example at random
        σ_i ← Unif({+1, −1})            ▷ Initialize spins at random
        while not converged to some fixed point σ̃_i^(u) do
            σ_i ← sign(Σ_j J_ij σ_j)    ▷ Run the spin update dynamics
        end while
        J_ij ← J_ij + (1/τN)(ξ_i^μ(u) ξ_j^μ(u) − σ̃_i^(u) σ̃_j^(u))   ▷ Update the coupling matrix
        J_ii ← 0
    end for
end for
```

We designed this procedure as an iterative version of the rule described in [3]. The key innovation that enables indefinite iteration without catastrophic forgetting is the simultaneous application of learning and unlearning within a single update step. We found heuristically that this concurrent reinforcement and suppression is the

crucial factor for maintaining stability over long training times.
The decision to subtract a single state per update, as opposed to an average over multiple states, is motivated by the properties of uncorrelated data. For such patterns, as established in [15], random initializations converge with high probability to spin-glass states that are uncorrelated with the stored memories and with each other. Consequently, subtracting the contributions of multiple spurious states simultaneously—as performed in methods like [32, 121] is less effective. Their uncorrelated nature causes their contributions to average out, significantly reducing the efficiency of the unlearning while targeting one state at a time guarantees a maximally effective, non-zero update.
Interestingly, we obtained an update rule (Eq. (3.1)) that structurally resembles that of moment-matching algorithms, which can be derived from a maximum-likelihood principle. We will discuss this connection in more detail in Sec. 3.9.2. However, we note that there is a critical distinction between moment-matching algorithms and our sampling process. Our algorithm uses zero-temperature dynamics to sample states for unlearning. This methods finds spin-glass states with high probability and therefore is intrinsically out of thermodynamic equilibrium. In contrast, classical moment-matching algorithms, such as those used for training Restricted Boltzmann Machines, require equilibrium sampling from the Boltzmann distribution, a process that is much more computationally expensive.

3.3 Datasets and Observables

We test the Daydreaming algorithm on different datasets to quantify its impact on the storage capacity and evaluate its performances using as key metrics the average size of the basins of attraction obtained at the end of the training of the matrix $\mathbf{J}$ trough the Eq. (3.4).

Uncorrelated data

We start our investigation testing the capabilities of the Daydreaming algorithm on a dataset of random patterns. In this case, each component ξ_i^μ is an independent and identically distributed (i.i.d.) Rademacher variable, assuming the values ± 1 with equal probability:

$$\mathbb{P}(\xi_i^\mu) = \frac{1}{2}\left[\delta(\xi_i^\mu - 1) + \delta(\xi_i^\mu + 1)\right] \quad \forall i,\ \forall \mu. \tag{3.5}$$

Results obtained with these synthetic uncorrelated data are presented in Sec. 3.5. We also utilized these random patterns to study the algorithm's convergence towards stationary asymptotic states and to verify the correctness of the proposed update rule in Eq. (3.4). These convergence results are presented in Sec. 3.4.

Correlated data

In addition to a dataset of random patterns, we employ a structured dataset generated by the random features model [64]. This model produces P patterns $\boldsymbol{\xi}^\mu$ as superpositions of D underlying random features.

Specifically, each pattern is generated as:

$$\xi_i^\mu = \text{sign}\left(\sum_{k=1}^{D} c_k^\mu f_{ki}\right), \tag{3.6}$$

where the coefficients $c_k^\mu \sim \mathcal{N}(0,1)$ and the features $f_{ki} \sim \text{Unif}(-1,+1)$ are all independent.
This model introduces an additional control parameter $\alpha_D = D/N$, which governs the degree of correlation among the examples while parameter $\alpha = P/N$ remains the standard load parameter of the Hopfield model, reported in Eq. (1.32). In the limit $\alpha_D \gg 1$, the distribution of the examples converges to $\text{Unif}(\{+1,-1\})$ and we get back to the dataset of uncorrelated examples. In contrast, for $\alpha_D \lesssim 1$, the patterns are correlated. The Hopfield model equipped with the Hebbian rule composed by correlated patterns, presents a richer phase diagram beyond the usual storage transition at α_c. For loads α exceeding a critical threshold $\alpha^*(\alpha_D)$, the model enters a *learning phase*: the stored patterns $\boldsymbol{\xi}^\mu$ become unstable, while the underlying features $\boldsymbol{f}^k$ emerge as locally stable fixed points of the dynamics [101]. We collect the results obtained using correlated data in Sec. 3.6.
Finally, we test the performances of the Daydreaming algorithm on the MNIST dataset [86]; the results are presented in Sec. 3.8.

Retrieval map

With the different datasets established, we now define the principal observable employed to quantify algorithmic performance. The central metric for our analysis is the *retrieval map*, which gives a comprehensive characterization of memory storage and stability.
Given a generic Hopfield network defined by the coupling matrix $\mathbf{J}$, and assuming retrieval dynamics governed by the parallel update rule in Eq. (3.1), we quantify the network's performance through its storage capacity and the average size of its basins of attraction. We can do this by numerically computing the retrieval map using the following procedure.
First, we define the *magnetization* m^μ of a network configuration σ with a stored pattern $\boldsymbol{\xi}^\mu$ as:

$$m^\mu = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i. \tag{3.7}$$

This quantity measures the correlation between the network's state and a specific memory. An overlap $m^\mu = 1$ indicates the state is identical to the pattern $\boldsymbol{\xi}^\mu$, while $m^\mu = 0$ signifies orthogonality (i.e., the state contains no retrievable information about the pattern).
To construct the retrieval map, we initialize the network in a configuration that has a predefined initial overlap m_{I}^μ with a chosen pattern $\boldsymbol{\xi}^\mu$. The dynamics defined by Eq. (3.1) is then run until convergence to a fixed point $\tilde{\boldsymbol{\sigma}}$. The final overlap m_{F}^μ with the original pattern $\boldsymbol{\xi}^\mu$ is measured at this fixed point.
This process is repeated for many initial configurations (at a fixed m_{I}^μ) and across all stored patterns $\mu = 1, ..., P$. Finally, the final magnetization m_{F} is averaged over all these initial conditions and patterns. The resulting curve, plotting the average

m_{F} as a function of the initial mean magnetization m_{I}, is the retrieval map.
The shape of this map characterizes the quality of retrieval: a curve that remains near $m_{\mathrm{F}} = 1$ for low initial overlaps indicates large basins of attraction, while a sharp drop signifies smaller, more restricted basins.
The retrieval map serves as a rich source of information regarding network performance. A key insight it provides is the stability of the stored patterns. As the network approaches its capacity, stored patterns can become locally unstable, meaning they cease to be fixed points of the update dynamics. This instability is directly observable on the retrieval map: a pattern is locally stable if and only if an initial state with perfect overlap $m_{\mathrm{I}} = 1$ evolves to a state with near-perfect overlap $m_{\mathrm{F}} \simeq 1$.
Furthermore, the retrieval map provides a qualitative measure for the mean size of the basins of attraction. The extent of the plateau where $m_{\mathrm{F}} \simeq 1$ indicates the range of initial conditions from which the dynamics reliably converge to a stored pattern and can be considered as a primary indicator of large, robust basins of attraction.
Indeed we are interested in understanding how distant from a typical pattern the starting configuration can be placed such that the spin update dynamics converge to the chosen pattern. The width of this plateau corresponds to the maximum typical initial noise (or minimum initial overlap m_{I}) from which successful retrieval is still possible.

Basin of attraction

In order to give a more quantitative measure of the basin size of the stored patterns, we performed the following analysis. For each pattern μ, we constructed initial configurations σ by flipping each spin of $\boldsymbol{\xi}^{\mu}$ with probability $p \in [0, 0.5]$. The corresponding initial overlap is, following the definition of Mattis magnetization in Eq. (3.7):

$$m_{\mathrm{I}}^{\mu} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \xi_i^{\mu}, \tag{3.8}$$

The final overlap after the optimization dynamics is denoted m_{F}^{μ}, and retrieval is considered successful when $m_{\mathrm{F}}^{\mu} > 0.95$.
The distance between $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}^{\mu}$ is naturally expressed in terms of the Hamming distance, i.e. the number of mismatched spins [11]:

$$d_H(\boldsymbol{\sigma}, \boldsymbol{\xi}^{\mu}) = \#\{i : \sigma_i \neq \xi_i^{\mu}\}. \tag{3.9}$$

This is because we want to answer two fundamental questions regarding the network's associative capabilities: First, how stable is the retrieved configuration once it is reached? Second, what is the maximum initial distance—or conversely, the minimum initial overlap—from a stored pattern $\boldsymbol{\xi}^{\mu}$ from which the dynamics will still successfully converge back to it? In other words, how large is the region of configuration space that is attracted to each memory?
The initial overlap relates directly to the normalized Hamming distance $d = d_H/N$ since we can indicate with d_H the number of spins σ_i whose signal is inverted:

$$m_{\mathrm{I}}^{\mu} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i \, \xi_i^{\mu} = \frac{1}{N} \sum_{i=1}^{d_H} -\xi_i^{\mu} \, \xi_i^{\mu} + \frac{1}{N} \sum_{i=d_H+1}^{N} \xi_i^{\mu} \, \xi_i^{\mu} = -\frac{d_H}{N} + \frac{N - d_H}{N} = 1 - 2\frac{d_H}{N}. \tag{3.10}$$

In the previous expression we recognize the Hamming distance as:

$$d_H(\boldsymbol{\sigma}, \boldsymbol{\xi}^\mu) = \frac{N}{2}\,(1 - m_{\mathrm{I}}^\mu), \tag{3.11}$$

and the corresponding normalized distance

$$d = \frac{d_H}{N} = \frac{1 - m_{\mathrm{I}}^\mu}{2}. \tag{3.12}$$

In this way, the quantity $1 - m_{\mathrm{I}}^\mu = 2d$ directly measures twice the normalized Hamming distance between the perturbed configuration and the reference pattern. For each pattern μ, we scanned over increasing perturbations and determined the largest initial distance that still leads to retrieval, i.e.

$$d_{\mathrm{basin}}^\mu = \max\Big\{\, 1 - m_{\mathrm{I}}^\mu \;\Big|\; m_{\mathrm{F}}^\mu > 0.95 \Big\}. \tag{3.13}$$

The basin size is therefore expressed in terms of $1 - m_{\mathrm{I}}^\mu$, which is equivalently twice the normalized Hamming distance from the reference pattern.
Finally, we computed the mean basin size over all patterns,

$$\langle d_{\mathrm{basin}} \rangle = \frac{1}{P} \sum_{\mu=1}^{P} d_{\mathrm{basin}}^\mu, \tag{3.14}$$

together with its statistical error. The dependence of $\langle d_{\mathrm{basin}} \rangle$ on the load $\alpha = P/N$ is shown in Fig. 3.7 for our Daydreaming algorithm compared to the ones obtained through the Storkey rule and the PseudoInverse model.

3.4 Convergence

We begin our analysis of the Daydreaming algorithm by examining its convergence properties toward a stationary asymptotic state. Figure 3.1 illustrates the training dynamics of the synaptic matrix $\mathbf{J}$ on uncorrelated data, defined in the previous Sec. 3.1 for a load $\alpha = 0.2$, a value chosen specifically because it exceeds the critical capacity of the standard Hopfield model with Hebbian couplings.
A key feature of the Daydreaming algorithm is its minimalism; it is governed by a single parameter, τ, which controls the timescale of the training dynamics. Our analysis will demonstrate that the algorithm's performance is robust independently on the specific value of τ, t as long as it is large enough to ensure convergence.
Consequently, Daydreaming operates effectively without requiring delicate parameter fine-tuning, rendering it a virtually parameter-free algorithm for practical applications.
The training of the coupling matrix $\mathbf{J}$ is a stochastic dynamical process evolving in a very high-dimensional space[1]. To analyze its behavior, we must therefore project the dynamics onto lower-dimensional, informative observables.
In Fig. 3.1, we track the evolution of two such key observables:

[1]The number of independent elements in an $N \times N$ symmetric matrix with null diagonal is $N(N-1)/2$.

- **Panel (a)** The scaled norm of the matrix update, $\tau||\Delta\mathbf{J}||_2$, where the update for a single step is defined as:

$$\tau\Delta J_{ij}^{(u)} = \frac{1}{N}\left(\xi_i^{\mu(u)}\xi_j^{\mu(u)} - \tilde{\sigma}_i^{(u)}\tilde{\sigma}_j^{(u)}\right). \tag{3.15}$$

  This quantity measures the typical magnitude of the displacement per training step.

- **Panel (b)** The Euclidean distance between the current coupling matrix $\mathbf{J}$ and its initial Hebbian condition $\mathbf{J}_0$ (see Eq. (3.2)), defined as:

$$\|\mathbf{J} - \mathbf{J}_0\|_2 = \sqrt{\sum_{i<j}(J_{ij} - J_{0,ij})^2}. \tag{3.16}$$

  This quantity measures the total cumulative deviation induced by the learning dynamics, quantifying how far the system has traveled from its starting point in the space of coupling matrices.

Together, these observables provide complementary information: the distance from the initial condition tracks the overall *path traversed*, while the update norm characterizes the *instantaneous activity* of the learning process. We plot the scaled displacement $\tau\|\Delta\mathbf{J}\|_2$ because, for sufficiently large τ, this quantity becomes independent of the specific value of τ. In this regime, both the scaled displacement and the distance from $\mathbf{J}_0$ converge to a constant asymptotic value on timescales proportional to τ itself.

It is important to clarify that this convergence does not imply the coupling matrix $\mathbf{J}$ itself converges to a fixed point. Due to the stochastic nature of the learning rule, $\mathbf{J}$ continues to evolve slowly, as evidenced by the persistent, non-zero scaled displacement. However, the learning dynamics does reach a stationary *statistical* state. This is an ensemble of coupling matrices, all possessing similar statistical properties, which ultimately produce identical retrieval performance—as quantified by the retrieval maps.

The experiments in Fig. 3.1 strongly suggest that running the Daydreaming algorithm for times significantly longer than τ is unnecessary, as the system's relevant statistical properties have stabilized by that point.

Most importantly, the asymptotic state reached by the Daydreaming algorithm exhibits a significant degree of robustness. Its dependence on τ is only noticeable for very small values. As soon as τ is large enough (and a number of the order of a few tens of steps is enough), the results are τ-independent. This property renders the Daydreaming algorithm essentially parameter-free, eliminating the need for delicate fine-tuning. Consequently, a simple and effective practical rule is to choose a reasonably large τ (e.g., $\tau \sim 100$) and run the algorithm for a number of steps slightly exceeding τ. This protocol works robustly across a wide range of conditions, saving the considerable computational time typically dedicated to parameter optimization in other machine learning methods.

The significant advantage of the Daydreaming algorithm is its capacity for indefinite iteration. The synaptic matrix $\mathbf{J}$ evolves without degradation over time, and the corresponding Hopfield network consistently maintains its retrieval capabilities. This is empirically demonstrated by the convergence of the retrieval maps to a stationary

form, as shown in Fig. 3.4(a) and throughout Fig. 3.8. For this reason, the most robust and meaningful signature of the algorithm's convergence is the stationarity of the retrieval map over time. We will adopt this criterion to evaluate convergence across all datasets studied in this work—namely, uncorrelated data in Sec. 3.5, random features data in Sec. 3.7, and MNIST data in Sec. 3.8.

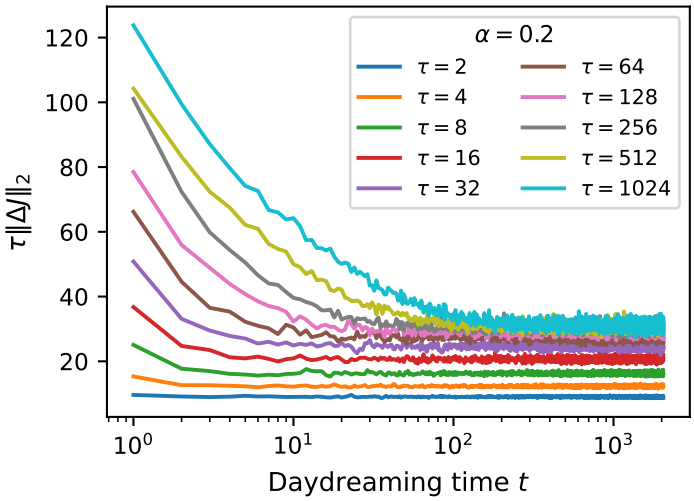


Figure 3.1 (a) The Daydreaming algorithm has a good asymptotic behaviour. The training process becomes stationary after a time $t \simeq \tau$ and, when $\tau \geq 64$, the asymptotic regime does not depend on the value of τ. We plot, as a function of training time t, the rescaled norm of the increment $\tau||\Delta J||_2$. We used $N = 500$ and $P = 100$ for this figure.

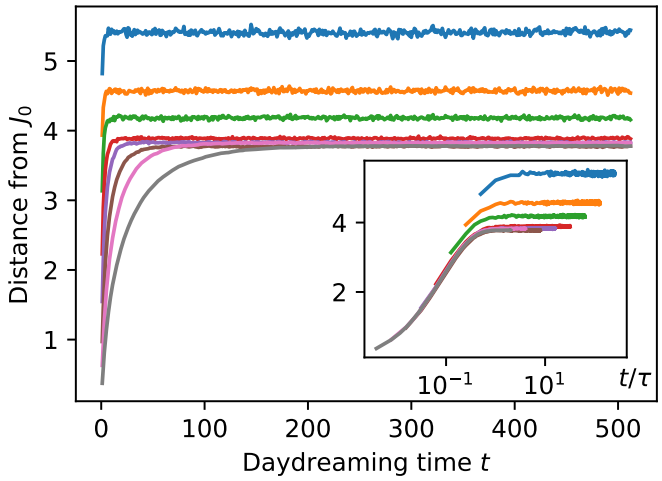


Figure 3.2 (b) The Daydreaming algorithm has a good asymptotic behaviour. The training process becomes stationary after a time $t \simeq \tau$ and, when $\tau \geq 64$, the asymptotic regime does not depend on the value of τ. We show the distance of the coupling matrix J from the Hebbian initial condition J_0. We show curves only for $\tau \leq 256$ to be more clear. *Inset*: we plot the same curves of the large panel, now as function of t/τ to make the collapse more visible. We used $N = 500$ and $P = 100$ for this figure. The colormap is the same of Fig. 3.1(a).

3.5 Results on uncorrelated data

As outlined previously, we begin our analysis by testing the Daydreaming algorithm on the dataset of uncorrelated patterns defined in Sec. 3.3. Having established the convergence of the algorithm in the preceding Sec. 3.4, we now focus on evaluating its performance.

Our objective is to determine whether this new procedure can overcome the catastrophic forgetting scenario inherent to the standard Hopfield model that occurs at $T = 0$ when the number of stored memories increases such that the network load surpasses the critical threshold $\alpha > \alpha_c \sim 0.138$. Therefore, our primary objective is to investigate the possibility of extending the retrieval phase of the Hopfield model at $T = 0$ through the implementation of the Daydreaming algorithm. To quantify this improvement, we compare the retrieval map of the standard Hopfield model, reported in Fig. 3.3a with that obtained from the Hopfield network after training with the Daydreaming algorithm, shown in Fig. 3.3b, at identical values of the load parameter α.

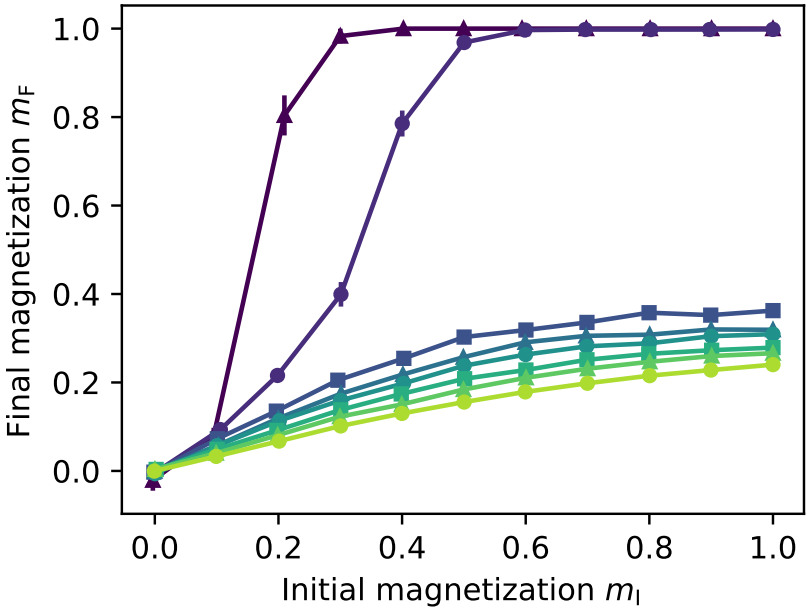
Our results confirm that this goal has been successfully achieved. As demonstrated in Fig. 3.3b, the proposed algorithm significantly increase the model's retrieval capabilities. The marked improvement in the final magnetization m_F and the extension of the plateau where $m_F \approx 1$ give clear evidence that the Daydreaming procedure successfully enlarges the basins of attraction and stabilizes the pattern retrieval for loads α that surpass the critical capacity α_c of the standard model.

In the low load regime $\alpha = 0.05, 0.1$, corresponding to the two darkest curves in both panels of Fig. 3.3, the Daydreaming allows an enhancement of the length of the plateau retrieval. This indicates that successful mean retrieval is achieved even from initial states with a mean magnetization as low as $m_I < 0.3$ and $m_I < 0.4$, respectively. Therefore, for $\alpha < \alpha_c$, Daydreaming further stabilizes retrieval that is already guaranteed by the Hebbian coupling, effectively widening the basins of attraction.

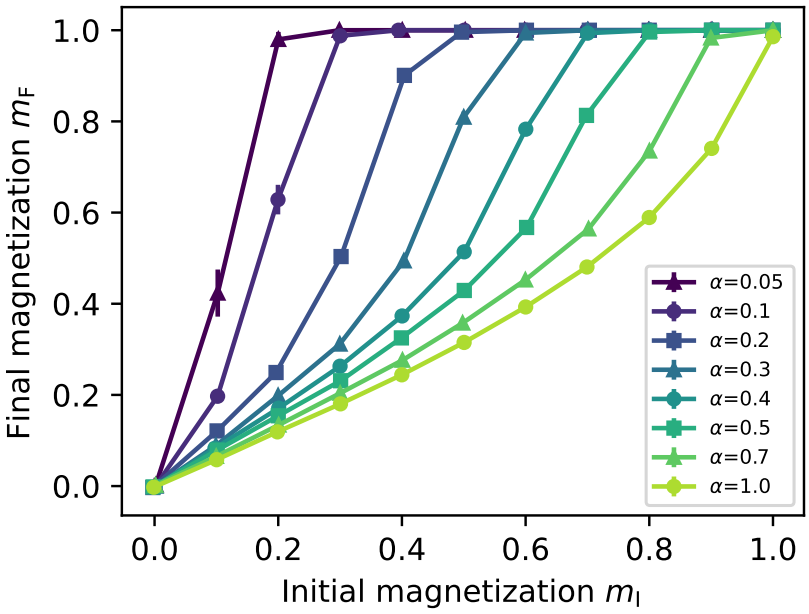
The most significant result result is observed in the high load regime $\alpha > \alpha_c$. In this case, the original model fails to recall stored memories even when initialized from the patterns themselves, as visible from the lighter curves in Fig. 3.3a where the average final overlap is only $m_F \sim 0.5$. Instead, the trained network shows a really improvement in the accuracy ($m_F = 1$ at $m_I = 1$). The stored patterns remain stable fixed points across this entire range, demonstrating that the Daydreaming algorithm can achieve the maximum theoretical capacity of $\alpha_c = 1$ for uncorrelated patterns [59]. As expected, the size of the basins of attraction, indicated by the extent of the $m_F \simeq 1$ plateau, shrinks as α increases and vanishes as α approaches the critical value of 1.

Moreover, the dual application of reinforcement and dreaming, implemented through the update rule in Eq. (3.1), also produces a appreciable extension of the retrieval plateau. This ensures stable retrieval even from initial configurations with significant noise $m_I < 1$, demonstrating the algorithm's robustness.

This means that the retrieval region at $T = 0$ in the $(T - \alpha)$ diagram phase of the Hopfield model in Fig. 1.5 is effectively enlarged, extending the ferromagnetic retrieval phase to higher values of α and moving the system out of the spin glass phase.



(a) Retrieval map of the original Hopfield model.



(b) Asymptotic retrieval map for the Daydreaming algorithm.

Figure 3.3 Comparison of retrieval performance. (**a**) Retrieval maps for a Hopfield network at different values of the load α. As the third curve (from top to bottom) demonstrates, for $\alpha > \alpha_c$ the network fails to retrieve the stored patterns $m_F < 1$, even when the dynamics is initialized from the patterns themselves $m_I = 1$. This indicates that the stored patterns are no longer fixed points of the dynamics beyond the critical capacity. (**b**) Asymptotic retrieval maps for the Hopfield network trained with the Daydreaming algorithm, for the same values of α. A plateau with $m_F \simeq 1$ exists for any $\alpha < 1$. We used $N = 10^3$ for these figures. Note that error bars are not always visible, since most of them are of order 10^{-2}.

The maximum theoretical capacity $\alpha_c = 1$, with vanishing basins of attraction, has been already obtained by other algorithms that implement a dreaming mechanism [53]. The key advantage of the Daydreaming algorithm is that it reaches these optimal performance levels through a local and biologically plausible learning rule. It does not require global operations, such as matrix inversions or computations involving the entire network state, which are necessary for other optimal algorithms. Furthermore, algorithms that rely on matrix inversions, (e.g. in the inversion of the patterns correlation matrix for the pseudo-inverse rule) often suffer from numerical instabilities near the maximum capacity. Notably, we observe no such numerical issues in the Daydreaming algorithm.

In Fig. 3.3b, we have shown the asymptotic retrieval maps for the Hopfield model trained with the Daydreaming algorithm, for a range of loads α from $\alpha = 0.05$ up to the maximum theoretical load $\alpha = 1$.
In Fig. 3.4, we report the results obtained by running the Daydreaming algorithm on uncorrelated data for a load of $\alpha = 0.4$, therefore we show the evolution of the retrieval map during the train until convergence. The final result corresponds to the fifth curve from the top in Fig. 3.3b. Also in Fig. 3.4 , different colors correspond to different training times, with lighter colors indicating earlier times and darker colors indicating later times.

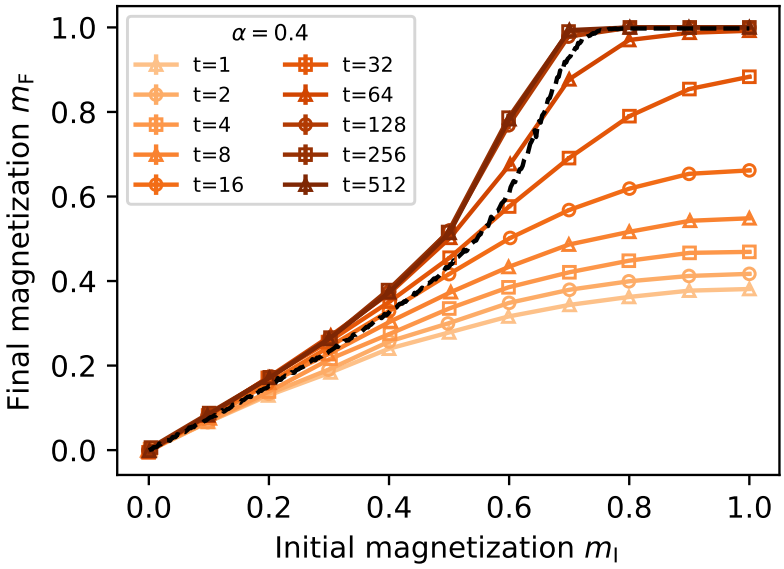


Figure 3.4 The Daydreaming algorithm improves state-of-the-art performances in retrieving random data. Evolution of the retrieval map for a Hopfield network trained on uncorrelated patterns with system size $N = 10^3$. Training was performed at load $\alpha = 0.4$ with characteristic time $\tau = 256$. Lighter to darker shades represent progression from initial ($t = 0$, Hebbian matrix) to later times. The training converges around $t \simeq 128 \lesssim \tau$ and the asymptotic curve outperforms the state-of-the-art results from [25] (dashed black line). Error bars are of order 10^{-2} and typically smaller than the plot markers.

At the initial time $t = 0$, the coupling matrix corresponds to the Hebbian matrix $\mathbf{J}_0$. As expected for a load $\alpha = 0.4$, which significantly exceeds the critical capacity of the Hebbian coupling matrix $\alpha_c \simeq 0.138$, the network fails to retrieve the stored patterns. This failure is evident from the retrieval map, which shows that an initial state with perfect overlap $m_{\mathrm{I}} = 1$ evolves to a final state with $m_{\mathrm{F}} < 1$, indicating that the stored patterns are not fixed points of the dynamics.
As training progresses, the retrieval map improves significantly, particularly for states near the stored patterns (i.e., for large m_{I}). The stored patterns themselves become locally stable fixed points by $t = 64$. By $t = 128 \lesssim \tau$, the retrieval map has converged to its asymptotic shape, which remains unchanged upon further training. The asymptotic retrieval map highlights two key improvements. First,it substantially outperforms the Hebbian baseline, exhibiting a wide plateau of successful retrieval $m_{\mathrm{F}} \simeq 1$ that extends down to initial overlaps as low as $m_{\mathrm{I}} \simeq 0.7$. Second, it slightly surpasses the current state-of-the-art performance achieved by the supervised algorithm described in [25] (shown by the black dashed line), which was previously considered optimal.
In summary, the Daydreaming algorithm spontaneously modifies the synaptic couplings to accommodate a number of patterns that exceeds the classical Hopfield capacity. It achieves this by maximizing the basins of attraction of the stored patterns without requiring external supervision or fine-tuning of hyperparameters, unlike the supervised approach [25], which depends sensitively on parameters such as the dreaming time or perceptron threshold.
To understand how rapidly the Daydreaming algorithm learns the optimal retrieval maps shown in Fig. 3.3b, we examine the data plotted in red and reddish shades in panels (a)-(d) of Fig. 3.8. The convergence to the asymptotic state is faster at smaller values of α and becomes progressively slower as the load approaches its maximum value. It is worth noticing, see panel (a) of Fig. 3.8, that, even for low loads, where the Hebbian rule is sufficient to store the patterns, the Daydreaming algorithm slightly improves over the Hebbian rule, by enlarging the basins of attraction. Overall, we find that the Daydreaming algorithm matches the same storage capacity for uncorrelated patterns as the state-of-the-art methods [3, 25] (which is the highest that is theoretically possible), while it has larger basins of attraction than those achieved in [25] as shown in Fig. 3.4.

3.5.1 Energy Landscape Evolution during Training

To quantify how the Daydreaming algorithm reshapes the energy landscape, we analyze the evolution of the energies associated with both the stored patterns and the states actually reached by the network dynamics defined in Eq. (3.4).
To analyze the energy landscape, we define the energy of a stored pattern $\boldsymbol{\xi^\mu}$ as the value of the Hamiltonian when the network configuration is exactly equal to that pattern, $\boldsymbol{\sigma} = \boldsymbol{\xi^\mu}$. The expression depends on the specific form of the coupling matrix $\mathbf{J}$. For the standard Hebbian coupling matrix, defined in Eq. (3.2) with the convention $J_{ii} = 0$, the energy of pattern $\boldsymbol{\xi^\nu}$ is given by:

$$H_0(\boldsymbol{\xi^\nu}) = -\frac{1}{2N} \sum_{i \neq j} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \xi_i^\nu \xi_j^\nu . \tag{3.17}$$

For uncorrelated, random patterns, the average (over patterns) of the quantity $\xi_i^\mu \xi_j^\mu \xi_i^\nu \xi_j^\nu$ is $\delta_{\mu\nu}$ for $i \neq j$ as specified in Eq. (1.23). Therefore the mean energy over all patterns is:

$$\overline{H_0(\boldsymbol{\xi}^{\boldsymbol{\nu}})} \sim -\frac{N}{2}. \tag{3.18}$$

During the Daydreaming process, the coupling matrix $\mathbf{J}(t)$ evolves over time. The energy of a pattern must therefore be computed using the current matrix:

$$H(\boldsymbol{\xi}^{\boldsymbol{\mu}}; t) = -\frac{1}{2} \sum_{i \neq j} J_{ij}(t)\, \xi_i^\mu \xi_j^\mu. \tag{3.19}$$

The mean energy of the patterns at time t is then:

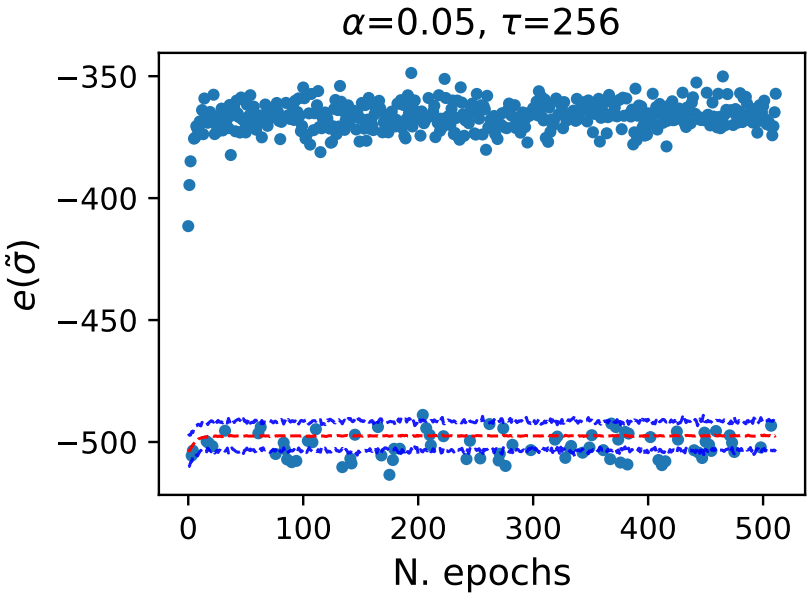
$$\overline{H(\boldsymbol{\xi}^{\boldsymbol{\mu}}; t)} = \frac{1}{P} \sum_{\mu=1}^{P} H(\boldsymbol{\xi}^{\boldsymbol{\mu}}; t). \tag{3.20}$$

This quantity is not static but evolves with the learning dynamics, reflecting how the energy of the stored patterns changes throughout the training process. After the asynchronous dynamics converges to a fixed point $\tilde{\boldsymbol{\sigma}}$, we compute its energy $H(\tilde{\boldsymbol{\sigma}})$[2] and compare it with the mean energy of the stored patterns, $\overline{H(\boldsymbol{\xi}^{\boldsymbol{\mu}}; t)}$. A noticeable energy gap emerges between these two values, which progressively narrows as the network load α increases.

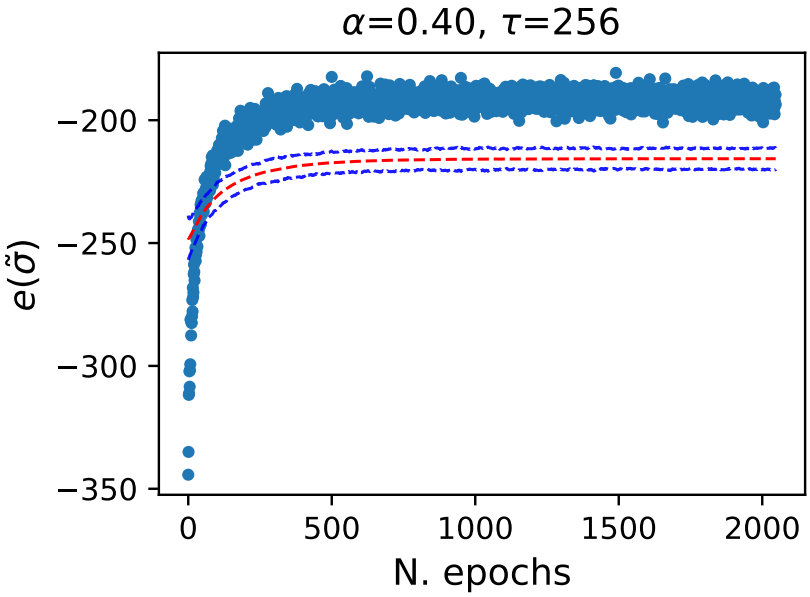
In the low load regime $\alpha < \alpha_c = 0.138$, the dynamics initialized from a random configuration can occasionally converge directly to a stored pattern, resulting in a fixed point with energy equal to the mean pattern energy as visible in Fig. 3.5a. Even in these cases, the subsequent coupling matrix update still performs a subtraction. However, since the initial matrix is Hebbian and the procedure is repeated for a number of steps much larger than P, the retrieval capability for all stored memories is ultimately guaranteed.

In the high-load regime $\alpha > \alpha_c$, the energy landscape becomes dominated by an exponential number of spurious states as visible in Fig. 3.5b-3.5c. Consequently, dynamics starting from a random initial condition almost always converge to a spurious state with an energy higher than that of the patterns. When the network load reaches the critical capacity $\alpha = 1$, the mean energy of the patterns equals the typical energy of the fixed points found by the dynamics, and the energy gap vanishes. This behavior is reported in Fig. 3.5d. For $\alpha > 1$, the gap reappears but is reversed; the energy of the patterns is now *higher* than the energy of the spurious minima found by the dynamics, confirming the complete loss of retrievability. As can be appreciated in Fig. 3.5e. This is directly reflected in the retrieval properties: even when starting from a pattern itself $m_I = 1$, the average final magnetization is less than one, $m_F < 1$, indicating that the patterns are no longer stable fixed points. The complete energy scenario is summarized in Figure 3.5.

[2] In the panels of Fig. 3.5 we have used the label $e(\tilde{\boldsymbol{\sigma}})$.

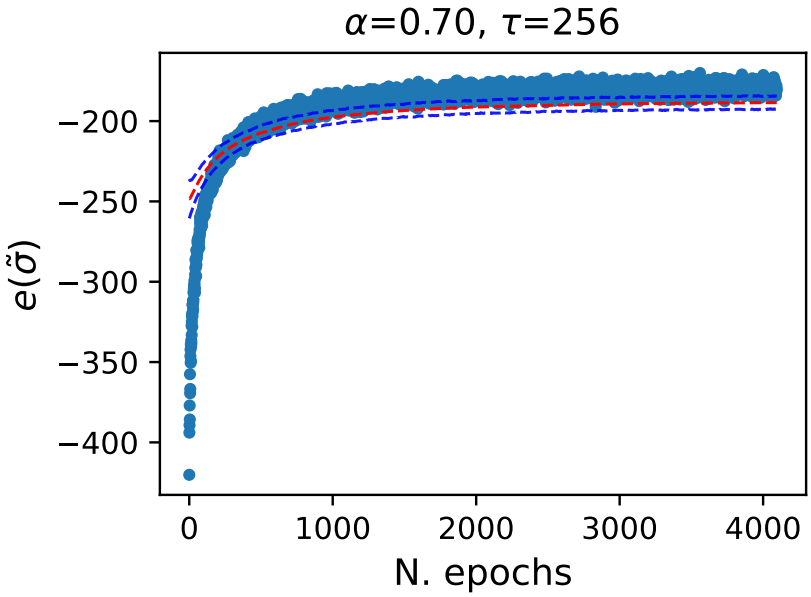


(a) Low load regime $\alpha = 0.05$. The energy gap is large and stable. The dynamics frequently converges to states with energy equal to the patterns.

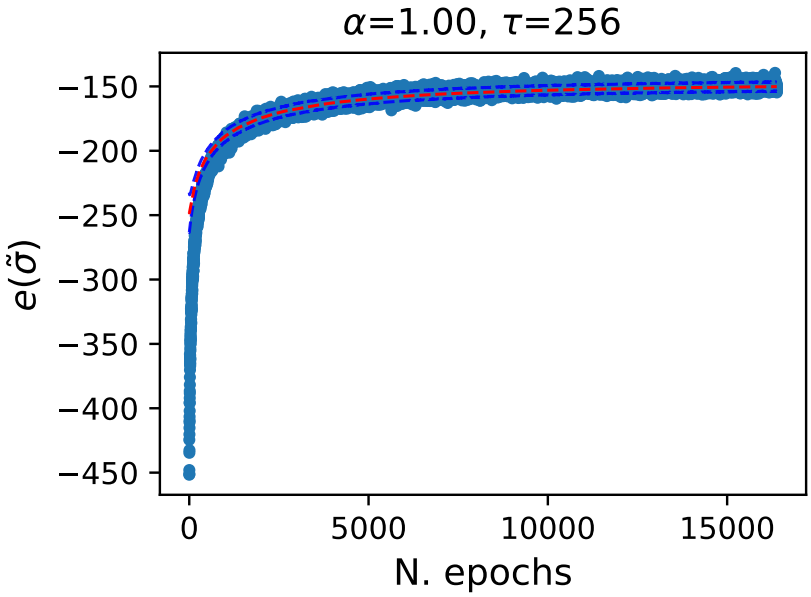


(b) Medium load $\alpha = 0.4$. The gap narrows as spurious states dominate the landscape.

Figure 3.5 Evolution of energies during Daydreaming training. Comparison between the energy $e(\tilde{\boldsymbol{\sigma}})$ of stationary states reached from random initial conditions and the mean energy of the stored patterns $\overline{H(\boldsymbol{\xi}^{\mu}; t)}$, for a network with $N = 1000$ in panel (a) and $N = 500$ in panel (b) as a function of the number of epochs in the Daydreaming process.

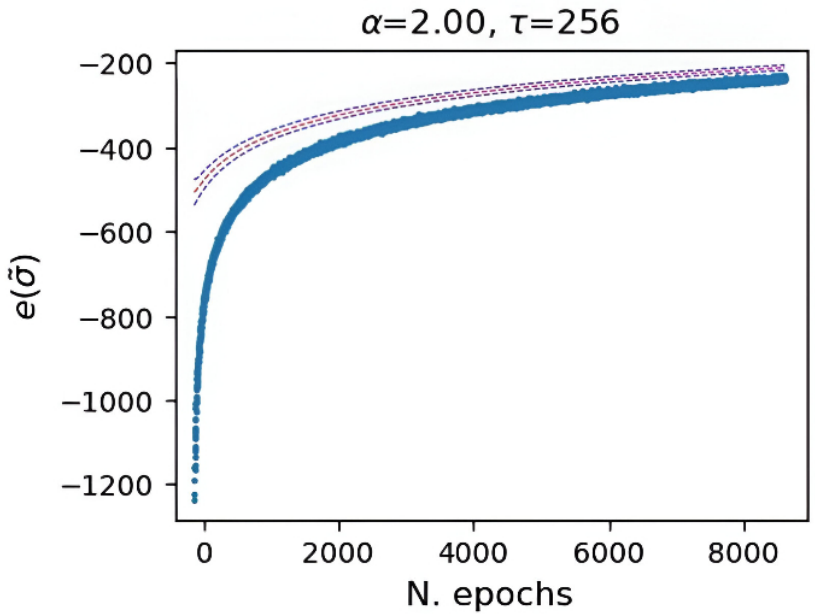


(c) High load $\alpha = 0.7$. The gap continues to narrow as the system approaches the maximum theoretical capacity for uncorrelated patterns and symmetric coupling matrix $\mathbf{J}$ [59].



(d) Critical load $\alpha = 1.0$. The energy gap vanishes.

Figure 3.5 Evolution of energies during Daydreaming training (continued). At the critical capacity $\alpha = 1$, the patterns become marginally stable as visible in the last curve on the right in Fig. 3.3b that corresponds to the darkest red in Fig. 3.8d. Results for $N = 500$.



(e) Super-critical load $\alpha = 2.0$.

Figure 3.5 Evolution of energies during Daydreaming training (continued). The energy gap reappears but is reversed: patterns have higher energy than the spurious minima found by the dynamics, confirming complete loss of retrievability. In this case $N = 1000$.

3.5.2 Comparison with the pseudo-inverse rule and the Storkey rule

A further benchmark to demonstrate the performance of the Daydreaming algorithm on uncorrelated data is a direct comparison with other established learning rules, namely the Pseudo Inverse [115] and Storkey [131] rules.
In this section we present a comparative analysis of the coupling matrices produced by the Daydreaming algorithm against those generated by these learning rules discussed in Sec. 2.4-2.5. The objective of this comparison is to quantitatively underscore the improvements in performance made by the Daydreaming procedure. We focus on two key metrics: the spectral properties of the asymptotic coupling matrix $\mathbf{J}$ and the resulting size of the basins of attraction, which directly quantify the quality of retrieval and whose definition is given in Sec. 3.3. As Daydreaming and Storkey are iterative rules, we first need them to converge to the stationary state before measuring the statistical properties of the asymptotic coupling matrices $\mathbf{J}$.
In Fig. 3.6 we compare the eigenvalue spectra of the coupling matrix $\mathbf{J}$ generated by the three different learning rules. Since the pseudo-inverse rule is constructed to enforce orthogonality among the pattern vectors, its eigenvalues concentrate on two well-separated values, with the number of positive eigenvalues being exactly equal to the number of stored patterns P. While these properties appear ideal in principle, we will show that Daydreaming gives superior functional performance. The spectrum

of the **J** matrix obtained by Daydreaming also exhibits a spectral gap, with the number of eigenvalues above this gap being exactly P, confirming the rule's perfect efficiency and its achievement of the maximum theoretical capacity $\alpha_c = 1$. However, this gap is smaller than that of the pseudo-inverse matrix. In contrast, the Storkey rule proves to be sub-optimal. Its spectral gap is already very small at $\alpha = 0.5$, and the number of eigenvalues above the gap exceeds the number of stored patterns. This spectral signature is consistent with its known sub-maximal capacity $\alpha_c < 1$. The spectral comparison reveals a fundamental distinction in how these algorithms encode memories. The pseudo-inverse rule constructs a mathematically ideal but rigid coupling matrix, strictly separating pattern-related eigenvalues from the bulk. In contrast, the Daydreaming algorithm produces a spectrum with a smaller gap, indicating a more distributed, robust, and potentially fault-tolerant encoding of memories. This suggests that Daydreaming can arrive to its high capacity not by enforcing strict orthogonality, but by finding a more efficient and smooth solution in the high-dimensional space of coupling matrices. The Storkey rule's spectrum reflects its inability to properly manage the interference between patterns, leading to its sub-optimal performance.

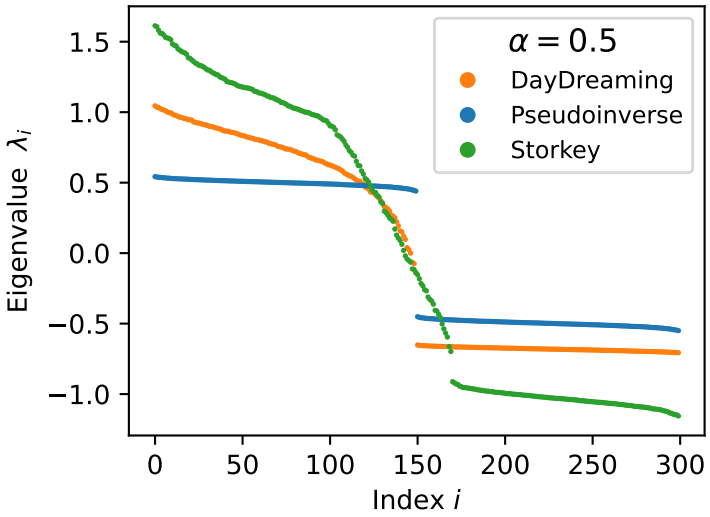


Figure 3.6 Eigenvalue spectra comparison. Eigenvalue spectra of the coupling matrices **J** for a network of size $N = 300$ at a load $\alpha = 0.5$. The spectra are shown for matrices obtained using the Daydreaming algorithm, the Pseudo-inverse rule, and the Storkey rule.

A more interesting comparison is shown in Fig. 3.7 where we report the average size of the basins of attraction across the entire range of $\alpha \in [0, 1]$. The Storkey rule produces the smallest, and therefore worst, basins of attraction. Furthermore, it becomes numerically unstable and fails to converge for $\alpha \gtrsim 0.7$. The pseudo-inverse rule, despite its "ideal" spectral properties, has basins of attraction that are consistently smaller than those generated by the Daydreaming algorithm for all values of α. This demonstrates that a mathematically optimal spectrum does not necessarily translate to superior retrieval robustness. We also emphasize a key

practical advantage: as α approaches 1, the pseudo-inverse rule fails due to numerical instabilities arising from the inversion of an increasingly singular pattern correlation matrix. The Daydreaming algorithm, operating via a local and iterative process, avoids this issue entirely and continues to perform smoothly across the entire range of α.

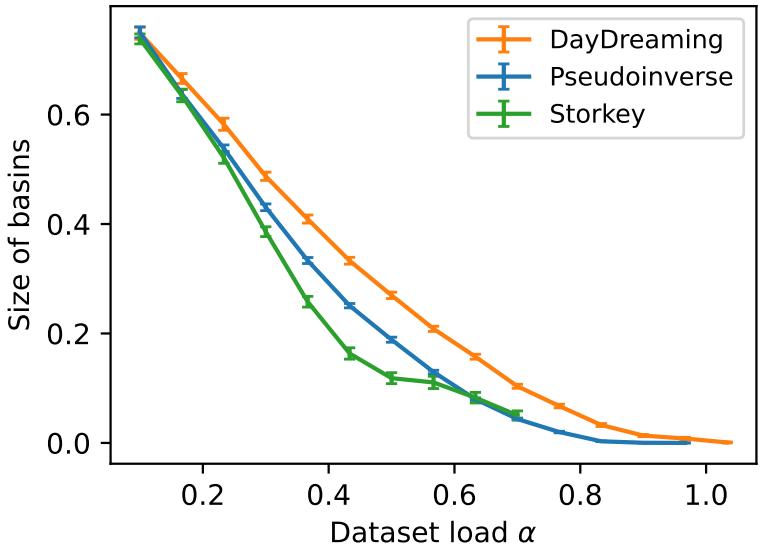
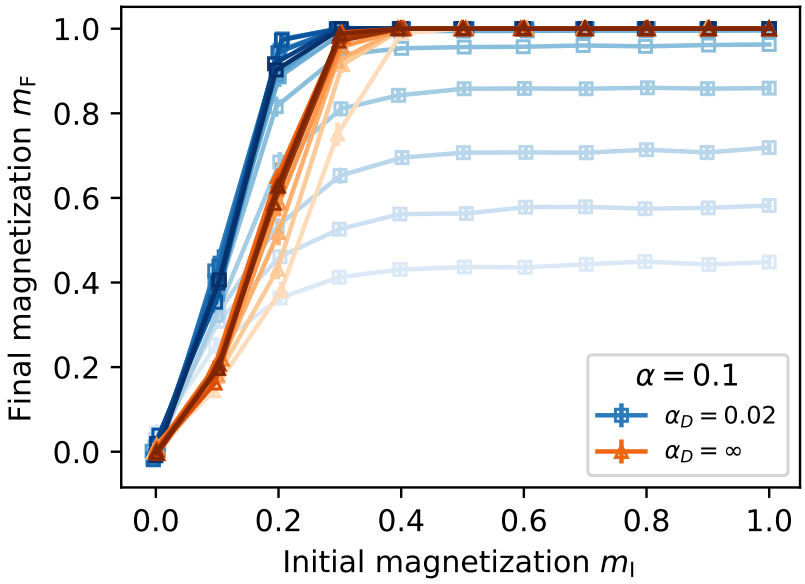


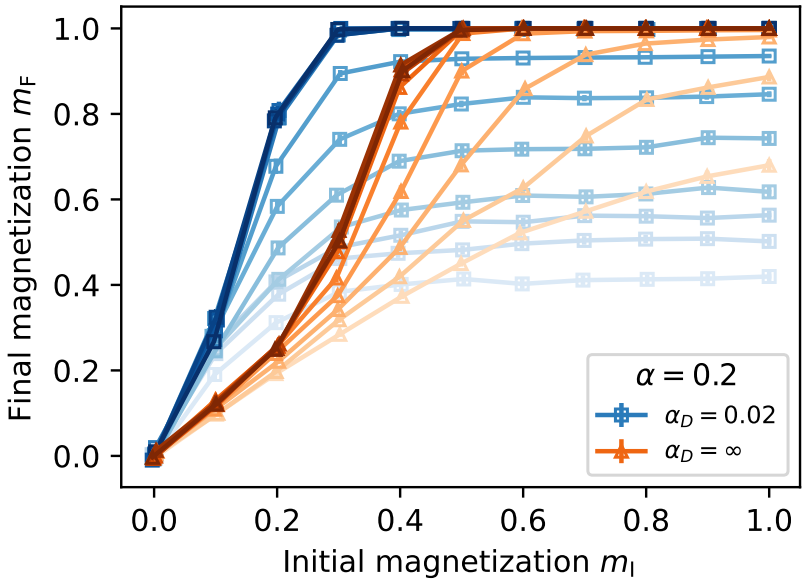
Figure 3.7 The Daydreaming coupling matrix is substantially different from other rules and performs better retrieval. Mean size of the basins of attraction for a $N = 300$ network as a function of the load α. The basin size for a pattern is determined by initializing the dynamics at a magnetization m_{I} and identifying the largest initial distance from which the dynamics converges to a final state with $m_F > 0.99$. The basin size is reported as the normalized Hamming distance $d = (1 - m_{\mathrm{I}})/2$ as described in Sec. 3.3. Data points are obtained by averaging over 30 independent samples.

3.6 Retrieval of correlated data

In this section, we present the results of applying the Daydreaming algorithm to the correlated datasets defined by Eq. (3.6).
In panels (a)-(d) of Fig. 3.8 we show the evolution of the retrieval map during training for both uncorrelated data (red triangles) and correlated data generated by the hidden features model (blue squares). The plotted results have been obtained with $N = 10^3$, and values of α and α_D, defined in Sec. 3.3, shown in the legend. Different shades of the colors represent different training times: the lightest color is $t = 1$ and the darkest is $t = 32768$, logarithmically spaced.
We observe that the convergence rate is load-dependent: it is faster for smaller values of α and slower for larger values. In any case, the Daydreaming algorithm converges without requiring any parameter fine-tuning, even when operating at the maximum theoretical load ($\alpha = 1$ for uncorrelated patterns).

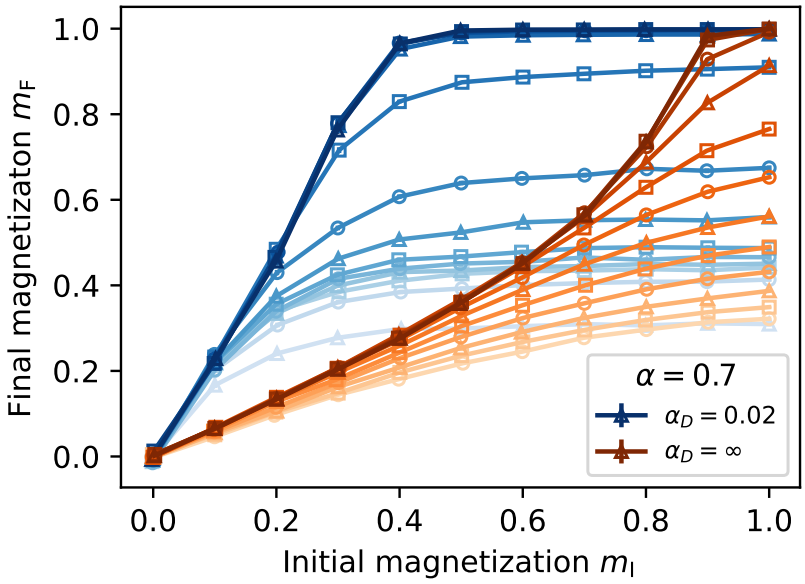


(a) $\alpha < 0.138$: Daydreaming enlarges the basin of attraction of uncorrelated examples and stabilizes the correlated examples.

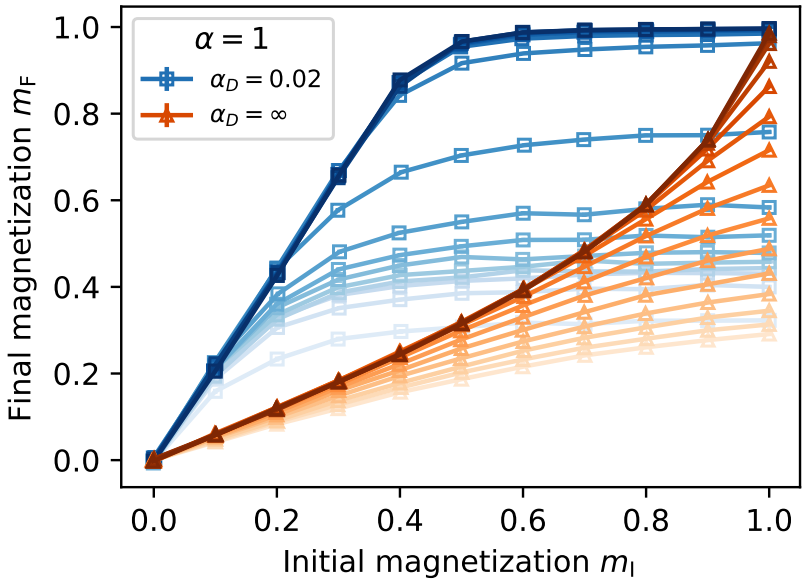


(b) $\alpha > 0.138$: Uncorrelated data become stable faster but correlated data end up with a larger basin of attraction.

Figure 3.8 Correlated examples are easier to store and retrieve than random ones with the Daydreaming algorithm. Evolution of retrieval maps during training. Blue squares: correlated data; red triangles: uncorrelated data. Shades represent training times from $t = 1$ (lightest) to $t = 32768$ (darkest, logarithmic spacing). Panel (a) shows performance below the Hopfield capacity α_c, (b) above capacity. Results for $N = 10^3$ in all panels.



(c) $\alpha = 0.7$: Analogous behavior of 3.8b with decreasing plateau for uncorrelaed data $\alpha_D = \infty$.



(d) $\alpha = 1$: Uncorrelated data become stable attractors but with vanishing basins; correlated data maintain large basins.

Figure 3.8 Correlated examples are easier to store and retrieve (continued). Panel (c) shows performance at $\alpha = 0.7$ and (d) at the maximum load $\alpha = 1$.

Even at low load, see panel in Fig. 3.8a, where $\alpha = 0.1$, we see using the Daydreaming algorithm a clear improvement over the standard retrieval obtained using the Hebbian rule. For uncorrelated data, we observe the formation of a slightly larger basin

of attraction. More significantly, for correlated data, the examples are unstable under the Hebbian coupling matrix but become spontaneously stabilized during Daydreaming training. At convergence, the correlated examples develop basins of attraction that are larger than those of uncorrelated patterns. Panel (b) in Fig. 3.8b shows results for a medium load $\alpha = 0.2$. At this load, the dreaming procedure is strictly required to have stable patterns. Although learning converges faster for uncorrelated examples, the coupling matrix $\mathbf{J}$ trained on correlated data produces larger basins of attraction. This result demonstrates the algorithm's inherent ability to extract and exploit useful correlations from a dataset autonomously, without explicit instruction. Fig. 3.8c shows the same behavior of the previous but with smaller basin of attraction for uncorrelated data due to the value of load $\alpha = 0.7$. Finally the last panel in Fig. 3.8d reports results for the case of $\alpha = 1$. Although the learning dynamic is slower, the Daydreaming algorithm converges to effective retrieval maps. For uncorrelated data, the retrieval map reaches the point $(m_{\mathrm{I}} = 1, m_{\mathrm{F}} = 1)$ with a vanishing plateau, meaning we are running at the maximum theoretical load $\alpha_c = 1$. Instead, the basin of attraction for correlated examples remains substantial even at $\alpha = 1$, meaning that the retrieval will also be possible for $\alpha > 1$. This implies that the Hopfield model, when trained with the Daydreaming algorithm, can exploit data correlations to overcome the storage milits imposed by uncorrelated examples. Daydreaming extends the storage limit of correlated examples described in [101], given that all the values of α and α_D reported in panels (a-d) of Fig. 3.8 are outside the storage phase identified in that work. The basin of attraction for correlated examples only vanishes around $\alpha = 2$ as visible in Fig. 3.9 where, as previously, blue squares are for correlated data and red triangles are for uncorrelated data.

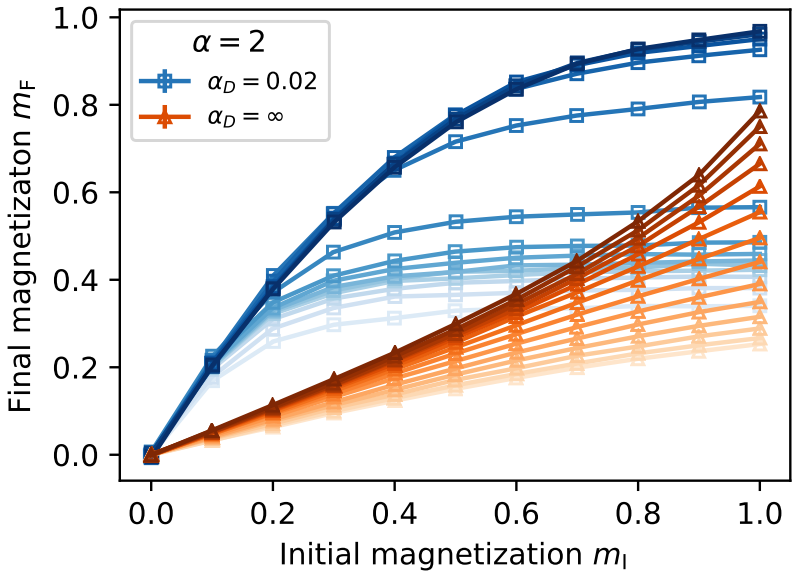


Figure 3.9 The basin of attraction of correlated examples vanish around $\alpha = 2$. Results for $N = 10^3$.

In conclusion, the results presented in Fig. 3.8 demonstrate that the Daydreaming

algorithm fundamentally improves the capabilities of the Hopfield model in two critical ways. First, it stabilizes memory patterns and enlarges their basins of attraction across all load regimes, effectively pushing the retrieval phase to the maximum theoretical load of $\alpha_c = 1$ for uncorrelated data. Second, and more importantly, it reveals a remarkable ability to autonomously exploit correlations within the data. The algorithm not only stabilizes correlated patterns that are unstable under Hebbian learning but also endows them with significantly larger basins of attraction than their uncorrelated counterparts. This intrinsic capability to extract internal structure allows the model to overcome the classical storage limits, allowing robust retrieval for correlated data at loads ($\alpha > 1$).

3.7 Retrieval of features

Previous work [101] demonstrated that the Hebbian rule applied to correlated data can stabilize hidden features within a specific *learning phase* region of the (α, α_D) plane (characterized by $\alpha_D < 0.138$ and sufficiently large α). A natural question is whether this capability extends to coupling matrices $\mathbf{J}$ learned via the Daydreaming algorithm. To investigate this, we define a *feature magnetization*:

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} f_{ki} \sigma_i, \tag{3.21}$$

which quantifies the overlap between the network state and a hidden feature. The corresponding feature retrieval map, where the quantities on the axes represent averages over all D features, is plotted in Fig. 3.10 for different training times.

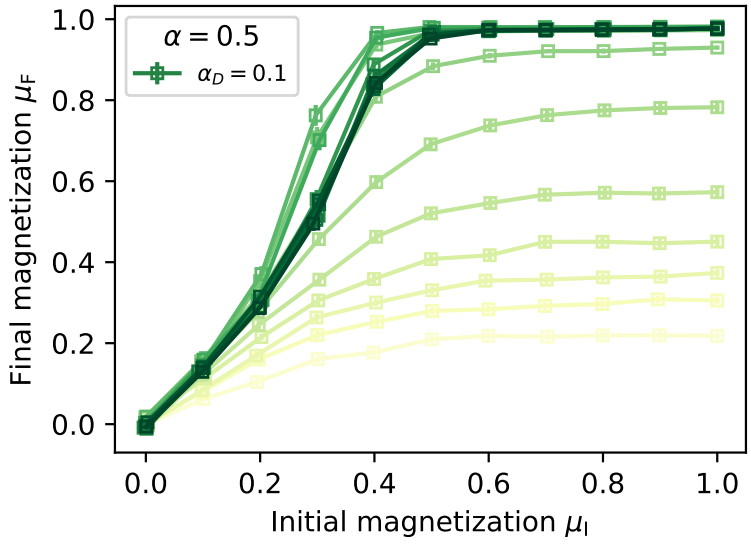


Figure 3.10 The Daydreaming algorithm stabilizes hidden features. Evolution of the feature retrieval map during training. The Daydreaming algorithm spontaneously stabilizes the features, creating well-defined basins of attraction. Results for a network of size $N = 10^3$ with $\alpha = 0.5$ and $\alpha_D = 0.1$. Different shades represent training times from $t = 1$ (lightest) to $t = 32768$ (darkest).

We notice that, for the parameters $\alpha = 0.5$ and $\alpha_D = 0.1$ used, the standard Hebbian rule fails to make the features locally stable ($\mu_{\mathrm{F}} < 1$ even when $\mu_{\mathrm{I}} = 1$). In contrast, the Daydreaming algorithm spontaneously stabilizes the features, producing large, well-defined basins of attraction around them. An intriguing observation is that the optimal shape of the retrieval map appears to be achieved at an intermediate training time, rather than at the final asymptotic state. The reason for this non-monotonic behavior is not yet understood and warrants further investigation.

In summary, learning the synaptic matrix $\mathbf{J}$ via the Daydreaming algorithm strongly enhances the stability and retrievability of both the stored correlated examples and the underlying hidden features that generated them. This extends the model's associative capabilities beyond the constraints of the standard Hebbian phase.

3.8 Results on the MNIST Dataset

Given the excellent performance of the Daydreaming algorithm in training the coupling matrix $\mathbf{J}$ on correlated data, we next explore its behavior when applied to highly structured, real-world data. For this purpose, we employ the MNIST dataset [86], the most widely dataset used in machine learning.
We train the model using a preprocessed, deskewed version of the MNIST dataset. The preprocessing pipeline, suggested in [22], involves the following steps: images are centered and vertically aligned, cropped to a central region of 14×14 pixels, and binarized to values ± 1 using a threshold of 86. Finally, the dataset is split into a balanced training set, from which we choose examples to feed the Daydreaming algorithm, and a test set containing unseen examples used exclusively for evaluation and therefore never presented to the network during the training.
We begin by discussing a caveat of the Daydreaming algorithm when applied to the MNIST dataset at very high loads ($\alpha \geq 100$) and a simple modification to address it. Consider a pair of pixels i and j for which $\xi_i^\mu \xi_j^\mu = 1$ holds for all training examples. This occurs frequently for pixels on the image boundaries, which consistently share the background color. Such a perfect correlation in the data induces a correspondingly strong coupling in the trained matrix J_{ij}. In principle, the Daydreaming algorithm correctly tries to assign $J_{ij} = \infty$ to enforce perfect alignment between these pixels (force pixels in i and j to take always the same value). However, a practical issue arises: if certain matrix elements become extremely large, subsequent normalization of the coupling matrix can cause the remaining elements to become vanishingly small. This effectively erases the relevant information stored in the coupling matrix. We resolved this issue by introducing a simple threshold $J_{\max}$ on the maximum absolute value any single coupling J_{ij} can assume. When imposing a threshold $J_{\max}$ on the coupling values, it becomes crucial to normalize the coupling matrix such that its overall scale is independent of the load α. Thus we chose to normalize the matrix obtained with the Hebb rule (i.e. the initial matrix of Daydreaming) by $P = \alpha N$ instead of N. This ensures a consistent norm across different values of α. We set the threshold to $J_{\max} = 0.5$ for all values of α used in the experiments. This value proved sufficient to prevent the dominance of a few couplings and the consequent collapse of the matrix's dynamic range. Furthermore, we verified that the results are robust and do not depend sensitively on the precise

value of $J_{\max}$ within a range of $[0.25, 1.25]$.
We summarize the performance of the Daydreaming algorithm trained on the MNIST dataset in Fig. 3.11.

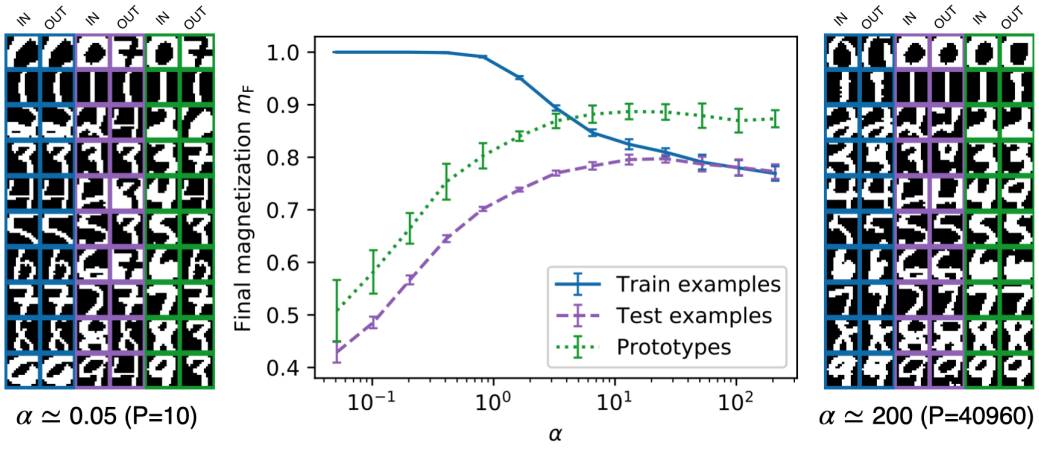


Figure 3.11 Attractors close to prototypes emerge at high α when training on MNIST with the Daydreaming algorithm. Central panel: Evolution of pattern stability as a function of the load α. Stability is quantified by the final overlap m_F between the asymptotic state of the dynamics and the initial pattern, for dynamics initialized at $m_I = 1$. Three types of patterns are shown: training examples (blue solid line), test examples (purple dashed line), and prototypes (green dotted line), where a prototype is the average of all training images for a given digit. At low α, only training images are stable. At high α, the curves for training and test data coincide, indicating generalization to unseen examples. In this regime, prototypes exhibit the highest stability, becoming the dominant attractors. **Lateral panels:** Visual examples of initial (left) and final (right) configurations for low load (left panel) and high load (right panel). The color coding matches the central panel. Parameters: $t = 2^{20}$ training steps, $\tau = 64$, coupling threshold $J_{\max} = 0.5$.

The central panel displays the final magnetization m_F obtained when the dynamics is initialized from configurations with $m_I = 1$ relative to different types of patterns:

- Training set examples (blue solid line)
- Test set examples (purple dashed line)
- Prototypes (green dotted line), where a prototype is defined as the average of all training images for a given digit (more details are provided below).

The lateral panels provide visual examples of the initial (IN) and final (OUT) configurations of the Hopfield dynamics. The color coding corresponds to that used in the central panel. The left and right panels illustrate results for low and high load conditions, respectively. A detailed discussion of these results follows.

3.8.1 Low Load Regime

We begin the analysis of Fig. 3.11 by examining the low load regime ($\alpha < \alpha_c$). Training examples remain stable ($m_F \simeq 1$) for loads up to $\alpha_c \simeq 0.8$, where the

critical load is identified as the point where the final magnetization for training examples begins to decrease (solid blue curve). This result is supported by Fig. 3.12, where we show the retrieval maps obtained for the MNIST dataset at several values of α and we observe that the size of the basins of attraction shrinks to zero around $\alpha \simeq 0.8$.

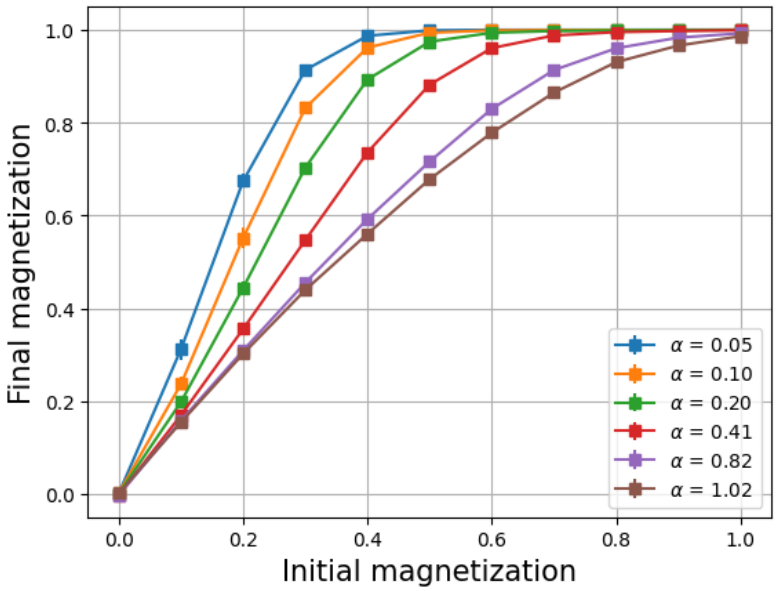


Figure 3.12 Evolution of retrieval basins for MNIST patterns under Daydreaming training. Retrieval maps after training completion for various loads α on the MNIST dataset. For $\alpha \leq 0.4$, the model develops finite basins of attraction around the stored memories. As α increases beyond this point, the patterns progressively lose stability, with their basins shrinking until the patterns become unstable in the high-load regime ($\alpha \gtrsim 0.8$). Results obtained using 14×14 pixel images and $t = 2^{20}$ training steps for each α value.

The model perfectly stores all training examples until it reaches a threshold between $P = 80$ and 160 (corresponding to α between 0.4 and 0.8). This capacity is comparable to recent results reported in [105], where a more complex attractor network was used to store approximately 100 MNIST examples. Note that they use full-resolution images and therefore more parameters than our model and also the system size N is larger. This means that the Daydreaming algorithm manages to store examples at a higher effective load α.

The performance of the Daydreaming algorithm is not only comparable to the state-of-the-art [105] but also challenges a common assumption in the field. Several recent modifications to the Hopfield model have been motivated by the premise that the standard model is inherently inefficient at storing and retrieving structured data such as the MNIST dataset. Our results demonstrate that this limitation is not fundamental; when equipped with an effective learning algorithm like Daydreaming, the standard Hopfield model can successfully handle such complex data.

This behavior is consistent with the expected properties of the Hopfield model in its storage phase. As shown in the left panel of Fig. 3.11, the network at low load acts as

a precise associative memory for the training set: it correctly retrieves stored training examples (blue boxes). However, it makes errors when presented with unseen test examples (purple boxes) or prototypes (green boxes), as these configurations lie outside the trained attractor landscape.

Classification

In the low load regime, when the network is presented with images not seen during training, the dynamics converge to the closest stored memory, a training image, rather than becoming trapped in spurious minima. This behavior is evident in the left panel of Fig. 3.11, where all output configurations correspond to one of the training images (blue columns).
Inspired by this reliable retrieval property and by related work applying attractor networks to classification tasks [22], we investigate the use of the Hopfield model trained with the Daydreaming algorithm for digit classification. When image labels are available, the Hebbian learning rule can be implemented in two primary ways, as summarized in [10]:

- **Supervised Hebbian learning:** Label information is used to group data points into their respective classes. The Hebb rule is then applied to the class means (prototypes), effectively storing these aggregated representations as attractors.
- **Unsupervised Hebbian learning:** Label information is disregarded. The Hebb rule is applied directly to the individual examples, storing each data point separately without exploiting class structure.

Similarly, we define supervised and unsupervised versions of the Daydreaming algorithm, distinguished by whether data regrouping is performed prior to training. We emphasize that both versions utilize the entire training dataset; however, only the supervised approach incorporates the additional information provided by the labels. For this reason, we decide to proceed with this approach.
We begin by defining a prototype $\boldsymbol{\pi}^C$ for each digit class C as the binarized average of the training images within that class:

$$\boldsymbol{\pi}^C = \operatorname{sign}\left(\mathbb{E}_{\mu \in C}\, \boldsymbol{\xi}^\mu\right). \tag{3.22}$$

We then proceed to train the model exclusively on these ten prototype pattern and evaluate its performance on the test set. For classification, a test image is presented to the network, and the dynamics are run until convergence. The predicted label is the class whose prototype matches the final state. Instead, if the dynamics converge to a spurious state not corresponding to any prototype, the image is left unclassified and then the label is assigned to that image.
On the MNIST dataset, we obtain with this supervised Daydreaming approach a test accuracy, defined as the fraction of correctly classified test images (i.e. fraction of correctly assigned label) of approximately 67.5%, surpassing the 61.5% accuracy reported in [22].
For completeness, we also evaluated an unsupervised version, where Daydreaming is

applied directly to the individual training images without using label information for regrouping. Classification is still performed by associating the final state with the nearest prototype. This method gives at the end an accuracy of approximately 66.2%. Although slightly lower than the supervised result, it still constitutes an improvement over the results reported in [22]. More in-depth results are reported in Table 3.1.

Digit	Correct (%)	Incorrect (%)	MCE	Spurious (%)
0	73.0 ± 3.8	24.8 ± 3.9	6	2.2 ± 0.8
1	96.1 ± 1.4	3.6 ± 1.1	0	0.3 ± 0.6
2	51.6 ± 5.2	45.4 ± 5.2	0	3.0 ± 1.1
3	66.0 ± 5.4	31.8 ± 5.0	1	2.3 ± 0.7
4	54.7 ± 8.1	43.8 ± 7.9	9	1.5 ± 0.8
5	60.2 ± 4.7	38.1 ± 4.5	3	1.7 ± 0.7
6	68.2 ± 3.6	29.2 ± 3.5	0	2.6 ± 0.8
7	79.8 ± 4.2	19.0 ± 4.2	1	1.2 ± 0.9
8	54.9 ± 5.9	42.9 ± 5.4	9	2.1 ± 0.9
9	65.7 ± 7.1	33.4 ± 6.9	7	0.9 ± 0.6

Table 3.1 Detailed classification results for MNIST using the supervised Daydreaming algorithm. For each digit, the table shows the percentage of test images classified correctly, incorrectly, and those converging to a spurious state (not matching any prototype). MCE (Most Common Error) indicates the digit most frequently confused with the target. Uncertainties represent the standard deviation over 30 independent runs.

3.8.2 High Load Regime

If we run the Daydreaming algorithm on the MNIST database with a large number of examples (high α regime), we observe a qualitative change in behavior. The training examples themselves lose local stability: dynamics initialized from a training image converge to a fixed point with a final magnetization m_F that decreases as P (and thus α) increases. At the same time, if we initialize the model in an example taken from the test set, we find that the final magnetization increases with α, as shown by the purple dashed line in Fig. 3.11. Notably, for $P \geq 10240$ ($\alpha \geq 52.2$), the final magnetization for training and test examples become equal. This indicates that the network treats seen and unseen examples identically, demonstrating a form of generalization. To identify the nature of the attractors that emerge in the high-load regime, we analyze the dynamics initialized from the class prototypes. The final magnetization when starting from a prototype follows the same increasing trend with α as for the test examples (green dotted line in Fig. 3.11), but is systematically higher. This indicates that the dominant attractors governing the network's dynamics are configurations that are closer to the prototypes than to any individual example. In the right panel of Fig. 3.11 we see that the network successfully reconstructs both training examples (blue boxes) and unseen test examples (purple and green boxes), converging to similar output patterns. This behavior is consistent with a generalized representation, where the network has learned to categorize inputs based on the underlying class prototypes rather than memorizing specific instances.

3.9 Take home message of the Chapter

In this Chapter we have collected and analyzed the Daydreaming algorithm, a new learning procedure that we designed to overcome fundamental limitations of previous dreaming algorithms for Hopfield networks, developed to increase the storage capacity of the original model. The algorithm modifies the synaptic coupling matrix $\mathbf{J}$ according to Eq. (3.4), implementing a continuous process that simultaneously stabilizes stored patterns, systematically enlarging their basins of attraction, while destabilizing spurious attractors.

3.9.1 Advantages and Key Contributions

The Daydreaming algorithm presents different significant advantages over existing methods:

- **Parameter-Free Operation:** Daydreaming is a compact and remarkably streamlined algorithm. It depends essentially on a single timescale parameter τ whose value requires no fine-tuning. Indeed, the convergence rate τ can be fixed with a large degree of freedom since the results depend on it very mildly. See Sec. 3.2-3.4. This resolves the critical "dreaming too much" problem that plagued previous approaches thus solving the main limitation of previously known dreaming algorithms exposed in Chapter 2.
- **Universal Data Handling:** The algorithm demonstrates robust performance across diverse data types, from uncorrelated patterns (analyzed in Sec. 3.5) to highly structured real-world data like MNIST (presented in Sec. 3.8), without requiring assumptions about data structure.
- **Exploitation of Correlations:** Daydreaming not only handles correlated data effectively but actually *benefits* from correlations. Correlated patterns, defined in Eq. (3.6), develop larger basins of attraction than uncorrelated ones and can be retrieved beyond theoretical limit of $\alpha = 1$ that constrains random patterns. The analysis is presented in Sec. 3.6.
- **Automatic Feature Extraction:** In both synthetic random-features data and MNIST, Daydreaming spontaneously discovers and stabilizes the underlying latent structure (hidden features and class prototypes), without explicit supervision or information about this structure during training.
- **Generalization Capability:** At high loads ($\alpha \gtrsim 50$ for MNIST), the network transitions from memorizing training examples to developing attractors that represent generalized concepts, treating seen and unseen examples identically, a hallmark of true generalization. Details are in Sec. 3.8.2.

Effectiveness on synthetic correlated data

The fact that Daydreaming finds large basins of attraction even for highly correlated random-features examples is somewhat surprising, as the classical picture in the literature is that correlation hinders retrieval [14, 55, 43, 89, 140]. We observe that

correlated patterns develop larger basins of attraction than uncorrelated ones and can be successfully retrieved even beyond theoretical limit of $\alpha = 1$ that constrains random patterns [59].
The above observations lead us to conjecture that the Daydreaming algorithm can automatically detect and exploit the correlation in the data thus improving the storage performances.
Strong evidence for this comes from experiments with synthetic data generated via the random-features model: in this case, despite receiving no explicit information about the underlying features, the algorithm improves also the retrieval of both the training examples *and* the hidden features that generated them. This indicates that the learning process spontaneously discovers the latent structure within the data. Note that this fact is again highly non-trivial, since the update rule in Eq. (3.1) ignores the structure of the hidden features that generated the examples. However, in the process of storing and reinforcing the examples the Daydreaming algorithm also learns the underlying hidden structure.

Non-trivial attractors on MNIST

By testing the Daydreaming algorithm on the MNIST dataset we have shown a proof of concept of how the exploitation of the hidden features might be relevant also for realistic datasets: new attractors emerge that are not exactly related to the training examples but rather to their underlying structure.
Application of Daydreaming to the MNIST dataset reveals a striking transition between two functional regimes. In the *storage phase* (low-to-medium load), the network acts as a high-capacity memory, perfectly storing a large number of correlated examples, surprisingly large compared to recent results [5, 105].
More significantly, in the *high-load regime*, a fundamental shift occurs. Rather than suffering catastrophic forgetting, induced by the loss of the local stability of the training examples, the network develops a new landscape of non-trivial attractors. While it is impossible for a Hopfield network to have attractors on each of them when the training examples are too many, the model trained with the Daydreaming algorithm develops a new non-trivial structure of basins of attraction. These attractors exhibit substantial overlap ($\simeq 0.75$) with both training and *unseen* test examples (continuous blue and dashed purple curves in Fig. 3.11). The fact that the network responds in the same way when initialized on both training and testing examples is noteworthy, suggesting that the model approaches some sort of generalization. Crucially, these emergent attractors are even closer ($\simeq 0.85$) to the class prototypes, green dotted curve in Fig. 3.11), demonstrating that Daydreaming has automatically discovered the underlying categorical structure of the data without any label supervision and indicates that the emergent attractors outside the storage phase are related to the hidden structure in the data (similarly to the random-feature case). If we interpret the training examples as noisy versions of their class prototype, the high-load energy landscape produced by the Daydreaming algorithm is reminiscent of what has been found in [5, 10]. A noteworthy difference between those works and the Daydreaming algorithm is that the latter is completely unsupervised, as it finds the class prototypes without using the class labels, at variance to the so-called supervised Hebb rule that instead groups data according to the class label.

Also note that, in [5], the authors need to select a finite, fine-tuned value for the dreaming time. Instead, the Daydreaming procedure can be iterated indefinitely even when training on the MNIST dataset. We believe that the reason behind this difference might be related to the sampling procedures in the two algorithms: Daydreaming is essentially a non-equilibrium process. Further investigations on this point are left for the future.
To understand qualitatively how good the emergent attractors are, in the right panel of Fig. 3.11 we have shown some examples: the output images appear as slight deformations of the inputs, but the nature of the digits appears to be preserved in most cases – even for test examples and prototypes, which never appeared explicitly in the training phase. This is in stark contrast with the low load regime, where the model perfectly stores train examples but fails on test examples and prototypes. More importantly, we do not observe the dynamics converging to spurious states, since these have been mostly removed by the Daydreaming procedure. So most of the test images do converge to meaningful attractors that can be eventually used for memory retrieval and classification. Unfortunately, we are not aware of any benchmark using the Hopfield model in this regime to compare with our results.

3.9.2 Relations and Distinctiveness to Moment-matching rules

Daydreaming is closely related to the maximum likelihood principle, as its update rule resembles a way to satisfy a moment-matching condition: see for example [91], where the "day" and "night" terms are explicitly identified, and also algorithms of the contrastive divergence family [30], that are commonly used to train Restricted Boltzmann Machines. For some reviews, see for example [40] or [85]. In this spirit, some learning rules related to ours, even if less effective, have been discussed in [82, 18] for a fully-connected symmetric model and generalized to sparse and asymmetric models in [27, 125].
The key difference between Daydreaming and all the mentioned moment-matching algorithms is how we sample the states to be unlearned: initializing the system to a random configuration and using Eq. (3.1) corresponds to a zero-temperature dynamics, whereas moment-marching requires sampling from the model at thermodynamic equilibrium (which is hard to sample for multimodal distributions) or some out-of-equilibrium stochastic process like those used in contrastive divergence. Whether our zero-temperature sampling can be related to a form of contrastive diverge is an interesting question that we leave for the future. Note that a zero-temperature sampling was also studied in [119], but instead of initializing the dynamics at random they use noisy versions of the training examples. This choice appears to produce smaller basins of attraction than our Daydreaming algorithm.

3.9.3 Limitations and Computational Considerations

The primary limitation of Daydreaming is its computational cost which it shares with other unlearning procedures and of pseudo-inverse rules. This translates into a long training time, which can be a limitation at high values of N and α (the number of epochs necessary to converge grows with α). Each step of Daydreaming requires the dynamics to converge from a random initial condition to a local minimum

of the Hamiltonian. The cost for this dynamics can be computed as follows: a single spin update costs $O(N)$, a global update of all spins costs $O(N^2)$, and the number of global updates to converge slightly grows with the problem size as $O(N^\delta)$ with the small exponent δ depending on both the α value and the coupling matrix J. Furthermore, we need N steps of Daydreaming to perform an epoch and the convergence of the learning process is achieved with a finite number of epochs (of the order of τ). Therefore, the computational complexity of Daydreaming is $O(N^{3+\delta})$. While this restricts application to very large-scale problems, the algorithm remains practical for moderate system sizes. We compare this complexity with the one of the rules discussed in the introduction: the iterative ones have the same complexity as Daydreaming, barring the fact that if they are run too much they destroy retrieval; the analytical rules instead typically involve an inversion of a matrix, which is an operation with complexity $O(N^3)$.
The other main limitation of Daydreaming is that a regularization is needed in the extreme case of very large loads (e.g. $\alpha > 100$ for MNIST). However, a simple coupling threshold $J_{\max}$ suffices to ensure stability without affecting the computational complexity of the algorithm.

3.9.4 Perspectives

We report here in conclusion some promising research future directions.

- **How Daydreaming Discovers Hidden Structure:** Given that at high loads we found indications of a rich and meaningful structure of basins of attraction, an extensive study of such structure seems promising. A deeper theoretical analysis of how Daydreaming extracts hidden structure could reveal general principles for unsupervised learning in attractor networks in the same spirit of [79].

- **Extension to Complex Datasets and Non-Equilibrium Analysis:** Given the results of Daydreaming on correlated data and its conceptual similarity to Boltzmann Machine training, it would be interesting to test it on challenging datasets, potentially requiring higher-order interactions. Furthermore, a formal interpretation of Daydreaming as a non-equilibrium process, analogous to analyses of Restricted Boltzmann Machines [41, 6], represents a promising theoretical direction.

- **Finite-Temperature Generalization:** The analogy with Boltzmann Machines also opens the door for supervised schemes such as the one presented in [126], where some of the nodes are interpreted as output, and the training is performed by clamping them to the desired labels. Overall, to compare our results to Boltzmann machines we would first need to extend our Daydreaming procedure to finite-temperature sampling. This extension to finite-temperature sampling could strengthen connections to Boltzmann machine literature.

- **Continual Learning:** The Daydreaming algorithm naturally operates as a continuous, ensemble-based learning process, where the coupling matrix $\mathbf{J}(t)$ evolves within a statistical ensemble of effective solutions. This dynamic nature suggests strong potential applicability to continual learning scenarios with

non-stationary data distributions. Noticing that the Daydreaming learning rule produces a coupling matrix which is not fixed, but changes in time taking values inside an ensemble of very good (in statistical sense) matrices, a key question could be how rapidly and efficiently this asymptotic ensemble can adapt to temporal changes in the dataset. We expect the Daydreaming learning rule could be a valid algorithm for the continual learning problem [70, 144].

- **Biological Extensions:** Another interesting possibility is to extend Daydreaming to sparse and non-symmetric coupling matrices, which are more biologically plausibly [125, 18]. The sparse case looks particularly promising from the perspective of the computational cost as it could reduce the cost per a single spin update from $O(N)$ to $O(1)$ enabling application of the algorithm to larger problems-scale problems.

This last direction naturally leads to the core questions explored in the second part of this thesis, where we investigate the dynamics of neural networks with biologically inspired, asymmetric, and diluted connectivity matrix to understand how such more biologically constrained architectures shape the attractor landscape and influence the long-term behavior of the system.

Part II

Asymmetric and Diluted Coupling Matrix

Chapter 4

Extending Neural Networks Toward Biology

Our greatest hopes could become reality in the future.

– *"Talkin' Hawkin"*, S. Hawking. *The Endless River*, Pink Floyd.

Now we move on the second part of this work which abandon the Hebbian paradigm typical of the Hopfield model and explore a more general setup for the construction of the couplings matrix more related to a biological inspiration since it does not represent a realistic model for neurobiological systems. The intent of this part is to define a framework for describing and understanding biologically inspired neural networks in order to investigate their properties and characteristics to improve our theoretical understanding of non-Hamiltonian systems.
Experimental studies consistently reveal that synaptic connectivity in both neocortex and hippocampus is highly structured rather than homogeneous or random. For instance, reciprocal connections and specific small-circuit motifs occur at rates far exceeding chance in neocortical microcircuits [129]. Similarly, recurrent connections in hippocampal CA3 networks, while sparse, are non-randomly organized, favoring structured motifs [69]. Theoretically, work by [29] demonstrates that optimizing memory basins under biological constraints naturally leads to sparse, moderately symmetric connectivity. These converging lines of evidence strongly motivate moving beyond fully connected, symmetric models toward architectures that incorporate structured heterogeneity, asymmetry, and dilution.
Indeed, moving toward biologically plausible network formulations, it becomes necessary to consider asymmetry and dilution in the network connections. Beyond the biological motivation, investigating disordered spin systems with asymmetric interactions is also interesting from a non-equilibrium statistical mechanics perspective. These models do not have an Hamiltonian and a detailed balance condition, which means that the fluctuation-dissipation theorem does not apply [37]. This implies that long-term dynamics cannot be derived using equilibrium statistical methods; instead, they require a complete analysis of the dynamic behavior of the system over time, as equilibrium averages no longer yield valid results.
Instead, under the assumption of symmetry in the synaptic connections between neurons i and j ($J_{ij} = J_{ji}$), network dynamics can be described as relaxation toward a

global energy function where fixed points represent local energy minima, surrounded by well-defined basins of attraction, and therefore the long-term behavior of this kind of neural networks model can be explored by linking them into a statistical mechanics framework as done for Hopfield model [73, 13, 15, 14].
However, synaptic connections in biological neural systems are asymmetric, making it important to examine how this asymmetry influences the long-term behavior of the model. Asymmetric neural networks, in a fully connected scenario, gained significant attention in the late 1980s: previous investigations [37, 20, 110, 38] have already shown that random asymmetry could lead to the destruction of spin glass states: a weak asymmetry can destroy spurious spin glass freezing in both SK-type and Hebbian fashion systems. This is because random dynamics fields replace the original static ones, stabilizing the paramagnetic phase at low temperature T.
Instead, ferromagnetic phases (obtained when J is chosen with a non-zero mean J_0) or the retrieval states for the Hebbian formulation, are described by global order parameters and, therefore, remain stable in presence of weak asymmetry since highly correlated and with local fields that cannot be generated by Gaussian random noise introduced by asymmetry. It is relevant to underline that the disappearance of the spin glass phase does not enlarge the basins of attraction of the memories since the paramagnetic phase attracts as much volume of the configuration space as the original SG one [38]. Recent related analysis have been conducted in [56, 93, 17].
In models with asymmetric couplings, two types of attractors emerge: ordered attractors (short cycles or fixed points) and chaotic attractors (whose length grows exponentially with the system size). Initial numerical observations revealing this dual nature of the attractors were made in [21, 68], where the relation between the mean number of fixed points and the size of the system of random ensembles of fully connected Ising spin glass models at thermodynamic limit, found in [136, 28], was extended to the asymmetric dense scenario.
The exponential dependence of the average cycle on the system size in this case was investigated in [106, 107], the behavior of the transient time, the number of steps needed to reach the attractors, was described in [36], while the description of the basin structure in [45]. These analyses have demonstrated that in asymmetric neural networks with deterministic parallel dynamics, a dynamic transition occurs from a chaotic regime, where the average cycle length exponentially increases with system size, to an oscillatory regime at high symmetry levels of the synaptic matrix J, where the average cycle length is 2. Recent studies have resumed the analysis for fully connected matrices with different degrees of symmetry [76].
In addition, the effects of dilution have been explored within discrete-time recurrent neural networks featuring McCulloch-Pitts neurons [54, 77, 87] considering the effect of dilution on fully asymmetric synaptic matrices. There is substantial variation in dilution: a diluted yet asymmetric connectivity matrix shows elevated capabilities in terms of attractors for the network. This regime is remarkably close to observations made previously in mammalian brain regions such as the cortex and hippocampus.
The connectivity matrix plays a key role in neural networks, as a schematic representation of *connectome*: the set of chemical synapses and electrical junctions between neurons. In developing neural network models as analogs of the brain, two fundamental points of view emerge. The first perspective is the *Connectionist Perspective* (Hebbian Learning) that emphasizes the importance of synaptic plasticity, which

refers to the ability of the connections between neurons to change their strength based on neural activity. In this perspective, the initial connective matrix, which represents the connections between neurons (synaptic weights), starts empty.
This means that in the early stages of brain development or network creation, there are no connections between neurons. As sensory experiences occur, new connections are gradually established. Each new experience contributes to filling this matrix, effectively encoding memories. These connections are formed based on the repeated co-activation of neurons during sensory experiences. The resulting connective matrix consists of pairs of neurons with associated connection strengths. These connections are modified through synaptic plasticity mechanisms.
One characteristic of this perspective is that the storage capacity of a neural network scales linearly with the number of neurons, leading to a limited maximum capacity α_c, where α is defined as the ratio between the number of stored memory P and the number of neurons N in thermodynamic limit. In this formulation, memories are stored in the network through the update of the synaptic weights using learning rules such as the Hebb's rule or other modifications.
Conversely, the second viewpoint, the *Innate Perspective* (Random Connectome), diminishes the role of synaptic plasticity, favoring a pre-existing random connectome meaning that synaptic weights are extracted from a defined distribution that identify a brain's cell born with that random connectome.
In this scenario, sensory inputs evolve towards attractor states driven by the innate properties of the random connectome, subsequently representing associated memory bits. The neural network starts with a set of connections between neurons that are random and not influenced by specific sensory experiences. Essentially, the connections are already in place before any sensory input is encountered. When a sensory input is presented, it dynamically evolves towards a specific attractor state based on the inherent structure of the random connectome. This attractor state represents the memory bit associated with that particular input.
The notable characteristic of this perspective is that it offers the potential for exponentially large storage capacity, closely resembling the biological reality. Since the network can encode memories based on the vast number of possible attractor states within the random connectome, the storage capacity is not constrained by the number of neurons.
In summary, the key difference between these two viewpoints lies in how they approach memory formation and storage in neural networks. The connectionist perspective emphasizes learning through experience and gradual connection formation, while the innate perspective relies on a pre-existing random network structure for memory storage, offering the potential for much larger storage capacity. The first perspective gives efficient results since can be mapped into an associative memory model but is far from a real biological system, which is more represented by the second model in particular, characterized by asymmetric and diluted connections.
The two points of view also offer a difference in the definition of the critical capacity of the resulting network. The first one is related to the maximum number of planted memories that the network can recover and distinguishes the retrieval phase from the spin glass phase where the dynamics does not converge to them. In the second formulation, it is related to the maximum number of partially correlated random memories for which the system admits solutions, as suggested in [7].

Initial analytical attempts to define diluted and asymmetric networks have been made in [61] in which the calculation done in [59, 60] to find the fractional volume of solutions was extended in the case of systems with defined symmetry η and in which each neuron is connected on average to $C \sim \ln N$ other neurons.
In thermodynamic limit $N \to \infty$, the volume is peaked around a typical correlation value $\eta(\alpha, K)$ that depends on the load α and the stability margin K. The analytical computation was performed in the symmetrically diluted scenario, which means that the existence of the link $i \to j$ implies the presence of the opposite link $j \to i$. This situation in the present work is obtained in the case $a_1 = 0$ and, regarding the dilution, for $\rho < 1$.
Under these constraints, the authors found that at zero margin $K = 0$ the maximum critical capacity for uncorrelated random memories $\alpha_c(K = 0) = 2$, is achieved not in the symmetric case but for a value of correlation $\eta = 1/\pi$. In the completely uncorrelated case $\eta = 0$, the critical capacity $\alpha_c(K = 0, \eta = 0) = 1.94$ and in the symmetric case $\alpha_c(K = 0, \eta = 1) = 1.278$.
Non symmetrically dilution was analyzed in [46] where was proposed a modified version of the Hopfield model using random independent parameters to represent the dilution and the asymmetry in the Hebbian matrix. The model constructs the connectivity matrix **J** by modulating a Hebbian prescription with random structural variables:

$$J_{ij} = C_{ij} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \tag{4.1}$$

where C_{ij} are independent, random parameters that simultaneously control dilution and asymmetry, and ξ^μ are the P stored memory patterns. For each pair i, j, C_{ij} is chosen at random according to the distribution:

$$\rho(C_{ij}) = \frac{C}{N}\delta(C_{ij} - 1) + \left(1 - \frac{C}{N}\right)\delta(C_{ij}). \tag{4.2}$$

Each connection exists with a probability $p = C/N$ with C the mean number of connection per neuron. Therefore, the definition of the load is $\alpha = (P - 1)/C$. This model was solved exactly for both parallel and random sequential dynamics. Similar to fully-connected models, a second order transition at zero temperature $T = 0$ was found at a value of the critical loading capacity $\alpha_c = 2/\pi$ beyond which retrieval fails. Below this capacity, so even when the system remembers, two initial states close to a stored pattern remain in its vicinity but never converge to an identical configuration. This behavior reveals the complex, non-trivial structure of the attractor basin surrounding a stored memory. In [139, 138] the authors studied a variant of the Hopfield model storing sparse $0, 1$ patterns and showed that the storage capacity diverges in the limit of vanishing activity. They demonstrated that in this sparse coding limit the capacity scaling coincides with the Gardner bound, implying that simple Hebbian learning rules become asymptotically optimal.

In this second part of thesis, we chose to abandon the Hebbian approach in the definition of the synaptic matrix keeping the formulation of the couplings completely general.

In an attempt to improve this scenario, we examine the dynamics of all possible inputs across numerous randomly connected synaptic matrices obtained with varying dilution and symmetry, continuing the results achieved in [76, 54, 77, 87].
These two main parameters, dilution, which represents the number of connections per node, and symmetry, which indicates the count of bidirectional connections with equal weights, represent two fundamental attributes that characterize the structure of the coupling matrix **J**.
Our goal is to move beyond traditional models and explore non-Hamiltonian models at different levels of dilution. Instead of maximizing the number of attractors, as in a Hebbian approach, we hypothesize that a network which has reached its storing capacity limit has certain characteristic optimal structures.
To find these structures, we sample sets of connectivity matrices with different levels of asymmetry and dilution, investigating the properties of random **J** matrices.
Our purpose is to improve previous results in a more general framework that includes the asymmetries in synaptic interactions, but even the possible symmetries in the topology of the interaction since these two aspects have been studied, until now, together.
We want to better understand the effect of each one introducing a general model that considers synaptic symmetry and network topology independently, allowing us to identify optimal models and understand how each aspect, joined with the dilution, contributes to network functionality.
We explore the behavior of systems with varying degrees of dilution and symmetry in the coupling matrix **J**, analyzing the effects of the structure of the synaptic matrix on the primary characteristics of the network.
Once this statistical study of the main properties of the model is completed optimization of the synapses matrix **J** can be pursued.
In Chapter 5 we define our model, the structure of the coupling matrix **J** thorough the main parameters (η, ρ, a_1) and the key observables of the dynamics. In Chapter 6 we collect and analyze all the numerical results.

Chapter 5

Dynamics in Asymmetric Diluted Spin Systems

With the technology at our disposal, the possibilities are unbounded.

– *"Talkin' Hawkin"*, S. Hawking. *The Endless River*, Pink Floyd.

In this Chapter we explore how key dynamical properties and primary characteristics of the network, such as the number, type, and stability of attractors, are influenced by the structural parameters of the synaptic matrix $\mathbf{J}$. Specifically, we investigate the effects of its degree of dilution and the asymmetry in both its couplings and topology. We present a generalized framework that distinguishes between topological asymmetry (a_1) and synaptic symmetry (η) providing a complete characterization of network architectures that goes beyond actual studies limited to specific parameter regimes.
The chapter is structured as follows. In Sec. 5.1, we introduce the model and its deterministic dynamics. In Sec. 5.2 we describe the procedure for constructing the synaptic matrix $\mathbf{J}$, defining the key parameters (ρ, η, a_1) that control its density, synaptic correlation, and topological asymmetry. In Sec. 5.3, we define the key observables used to characterize the system's attractor landscape and dynamical behavior. Finally, in the following chapter, we will report and discuss the numerical results obtained from a comprehensive exploration of this parameter space.

5.1 The model and dynamics

We consider a network of N binary spins, represented by Ising variables $\sigma_i = \pm 1$[1] interacting through a coupling matrix $\mathbf{J}$, where each element J_{ij} quantifies the connection strength from neuron j to neuron i. The system evolves in discrete time via synchronous (parallel) updates governed by the zero-temperature Glauber dynamics:

$$\sigma_i(t+1) = \operatorname{sgn}\left(\sum_{j\neq i} J_{ij}\,\sigma_j(t)\right). \tag{5.1}$$

[1] We're using Ising variables instead of McCulloch-neurons [94].

The synaptic couplings J_{ij} are quenched random variables, drawn from a specified distribution, we will standard Gaussian variables, and kept fixed thereafter. This results in a deterministic, disordered dynamical system whose evolution, defined by Eq. 5.1, can be described by the mapping [20]:

$$\boldsymbol{\sigma}(t+1) = f_{\mathbf{J}}(\boldsymbol{\sigma}(t)), \tag{5.2}$$

where $\boldsymbol{\sigma}(t) = (\sigma_1(t), \ldots, \sigma_N(t))$ denotes the full configuration of the system at time t, and $f_{\mathbf{J}}$ represents a random realization of the deterministic map defined by Eq. (5.1). Given the finite number of states (comprising 2^N distinct configurations) and the deterministic dynamics, each initial spin configuration $\boldsymbol{\sigma}(0)$ evolves towards a definite state. The long-time behavior of the system is characterized by attractors and the dynamics eventually converges to either:

- A **fixed point**, where $\boldsymbol{\sigma}(t+1) = \boldsymbol{\sigma}(t)$ for all subsequent times.
- A **limit cycle** of length L (where $2 \leq L \leq 2^N$), satisfying $\boldsymbol{\sigma}(t+l) = \boldsymbol{\sigma}(t)$, with L being the smallest positive integer for which this holds:

$$L = \min_l(\boldsymbol{\sigma}(t+l) = \boldsymbol{\sigma}(t)). \tag{5.3}$$

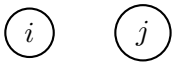
The evolution described in Eq. (5.1) connects the states of the system, enabling the description of dynamics in terms of oriented graphs on the state space. Each configuration, or node, has precisely one outgoing arrow, though different arrows may terminate at the same state. This structure falls within the framework of a **Random Map model** (see Appendix C.2 for a brief overview) in the fully connected case.
A crucial distinction from the model discussed in the previous part of this thesis, is the relaxation of symmetry constraints on the coupling matrix $\mathbf{J}$. In neural networks, the presence of asymmetry ($J_{ij} \neq J_{ji}$) implies the non-existence of a global energy function (Hamiltonian) that decreases monotonically under the dynamics, leading to the presence of limit cycles. For parallel dynamics, that it the rule update that we have used in our analysis, this phenomenon occurs even in symmetric cases.

5.2 Construction of the synaptic matrix J

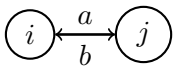
We now describe the procedure for generating the synaptic coupling matrix $\mathbf{J}$. The construction begins by defining a binary adjacency matrix $\mathbf{A}$ that encodes the existence of synaptic connections between neurons. We define the elements of A based on the following relations:

$$a_{ij} = \begin{cases} 1 & \text{if } \exists\, i \to j \\ 0 & \text{if } \nexists\, i \to j \end{cases}. \tag{5.4}$$

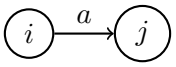
To systematically control the topological properties of the network, specifically its density and symmetry, we introduce three fundamental parameters (a_0, a_1, a_2) that define the joint probability distribution for each pair of matrix elements (a_{ij}, a_{ji}):

- $\boldsymbol{a_0 = \mathbb{P}[(a_{ij}, a_{ji}) = (0,0)]}$: Probability of missing link 

- $\boldsymbol{a_2} = \mathbb{P}[\boldsymbol{(a_{ij}, a_{ji}) = (1, 1)}]$: Probability of bidirectional link[2]



- $\boldsymbol{a_1} = \mathbb{P}[\boldsymbol{a_{ij} + a_{ji} = 1}]$: Probability of direct link



The parameters (a_0, a_1, a_2) must satisfy the normalization condition: $a_0 + a_1 + a_2 = 1$. Furthermore, defining ρ as the density of non-zero elements in the adjacency matrix (excluding the diagonal), we obtain the relation:

$$\frac{1}{2}a_1 + a_2 = \rho. \tag{5.5}$$

Since in our formulation, self-interactions are excluded ($a_{ii} = 0\ \forall i$), the density ρ is given by:

$$\rho = \frac{1}{N(N-1)} \sum_{i \neq j} \delta_{a_{ij},1} = \frac{1}{N(N-1)} \sum_{i<j} \left(\delta_{a_{ij},1} + \delta_{a_{ji},1} \right). \tag{5.6}$$

The parameters a_1, a_2, and a_0 can be explicitly written as:

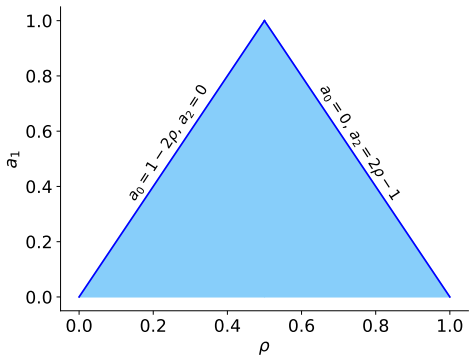
$$a_1 = \frac{2}{N(N-1)} \sum_{i<j} \left(\delta_{a_{ij},1}\delta_{a_{ji},0} + \delta_{a_{ij},0}\delta_{a_{ji},1} \right), \tag{5.7}$$

$$a_2 = \frac{2}{N(N-1)} \sum_{i<j} \delta_{a_{ij},1}\delta_{a_{ji},1}, \tag{5.8}$$

$$a_0 = \frac{2}{N(N-1)} \sum_{i<j} \delta_{a_{ij},0}\delta_{a_{ji},0}. \tag{5.9}$$

The relation in Eq. (5.5) is verified by noting that $\delta_{a_{ij},0} + \delta_{a_{ij},1} = 1$ for all a_{ij}.
In general, the probability a_1 is not independent but is constrained by the density ρ through Eq. (5.5). For the special case of an upper triangular adjacency matrix (containing only directed links from $i \to j$ with $i < j$), we achieve the maximum possible asymmetry since $\rho = \frac{1}{2}$ and $a_2 = 0$. This maximum corresponds to configurations where all existing connections are unidirectional.
More generally, the maximum achievable asymmetry $a_1^{\max}$ for a given connection density ρ is calculated through the relation in Eq. (5.10) and reported in Fig. 5.1. Considering now as initial parameters (ρ, a_1) with the condition $a_1 \leq a_1^{max}(\rho)$, we can construct the adjacency matrix A opportunely. We want to stress that this construction method is different from the random formulation in which the values of the adjacency matrix are randomly chosen. These specific topological constrains are imposed because we want to build an entire ensemble of synaptic matrices **J** with systematically controlled levels of asymmetry in the couplings, density and topological symmetry defined through simple and general parameters.

[2] In particular $a = b \iff \eta = 1$ that is the symmetric case.



$$a_1^{\max}(\rho) = \begin{cases} 2\rho & \text{if } \rho \leq 1/2 \\ 2(1-\rho) & \text{if } \rho \geq 1/2 \end{cases} \tag{5.10}$$

Figure 5.1 Maximum asymmetry parameter $a_1^{\max}(\rho)$ as a function of the density ρ.

For pairs of neurons connected by double links ($a_{ij} = a_{ji} = 1$), we introduce a new parameter η to quantify the correlation between the synaptic strengths J_{ij} and J_{ji}. Following [21, 68, 106, 36], we define:

$$\eta = \frac{\mathbb{E}[J_{ij}J_{ji}]}{\mathbb{E}[J_{ij}^2]} \in [-1, 1]. \tag{5.11}$$

where the expectation is taken over the distribution of non-zero couplings. This parameter interpolates between three fundamental regimes:

- $\eta = 1$: Perfect correlation (symmetric couplings, $J_{ij} = J_{ji}$)
- $\eta = 0$: Independence (asymmetric couplings)
- $\eta = -1$: Perfect anti-correlation (anti-symmetric couplings, $J_{ij} = -J_{ji}$)

The non-zero elements of $\mathbf{J}$ are drawn from a Gaussian distribution $\mathcal{N}(0, 1)$, imposing the desired correlation η for double links. We present two equivalent methods for generating correlated pairs (J_{ij}, J_{ji}):

Method 1: Linear Combination of Independent Variables We define:

$$\begin{cases} J_{ij} = c_1 z_1 + c_2 z_2 \\ J_{ji} = c_1 z_1 - c_2 z_2 \end{cases} \tag{5.12}$$

with $z_1, z_2 \sim \mathcal{N}(0, 1)$ independent and $c_1, c_2 \in [0, 1]$. To satisfy $\mathbb{E}[J_{ij}^2] = 1$ and $\mathbb{E}[J_{ij}J_{ji}] = \eta$, we require:

$$c_1^2 + c_2^2 = 1 \quad \text{and} \quad c_1^2 - c_2^2 = \eta. \tag{5.13}$$

Solving we get $c_1 = \sqrt{(1+\eta)/2}$ and $c_2 = \sqrt{(1-\eta)/2}$, giving:

$$J_{ij} = \frac{1}{\sqrt{2}}\left(\sqrt{1+\eta}\, z_1 + \sqrt{1-\eta}\, z_2\right), \quad J_{ji} = \frac{1}{\sqrt{2}}\left(\sqrt{1+\eta}\, z_1 - \sqrt{1-\eta}\, z_2\right). \tag{5.14}$$

Method 2: Cholesky Decomposition The second approach consists in using the Cholesky decomposition. We consider $\Sigma_0 = \mathbb{1}$ the initial covariance matrix between to random independent Gaussian variables $z_1, z_2 \sim \mathcal{N}(0,1)$. If we want now to correlate them with correlation η we apply the Cholesky decomposition to Σ defined as:

$$\Sigma = \begin{pmatrix} 1 & \eta \\ \eta & 1 \end{pmatrix} = LL^\top \tag{5.15}$$

with L is the Cholesky factor of the target covariance matrix Σ. We generate correlated variables through the transformation:

$$\begin{pmatrix} J_{ij} \\ J_{ji} \end{pmatrix} = L \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \tag{5.16}$$

The lower triangular matrix L contains the transformation coefficients that relate the original z_1 and z_2 to the correlated variables J_{ij}, J_{ji} that we want. The entries of the L matrix are obtained through the following relations:

$$L_{ii} = \sqrt{\Sigma_{ii} - \sum_{j<i} L_{ij} L_{ij}^T} \quad \text{and} \quad L_{ij} = \frac{1}{L_{jj}} \left(\Sigma_{ij} - \sum_{k<j} L_{ik} L_{jk}^T \right) \quad \text{for } i > j. \tag{5.17}$$

Therefore, the Cholesky decomposition gives:

$$L = \begin{pmatrix} 1 & 0 \\ \eta & \sqrt{1-\eta^2} \end{pmatrix} \tag{5.18}$$

resulting in:

$$\begin{cases} J_{ij} = z_1 \\ J_{ji} = \eta z_1 + \sqrt{1-\eta^2} z_2 \end{cases} . \tag{5.19}$$

With both the procedures we end, when double links are present, with two variables J_{ij} and J_{ji} extracted from a Gaussian distribution $\mathcal{N}(0,1)$ and correlation η as requested. The probability that $J_{ij} = J_{ji}$ is given by $\mathbb{P}[J_{ij} = J_{ji}] = \delta_{\eta,1} a_2 + a_0$.
The complete algorithm used for constructing the synaptic matrix can be found in App. B.1.4 (we have used this second method for the entries of the synaptic matrix in our implementation).
To completely explore the space of network architectures we vary the parameters (ρ, η, a_1) across their admissible ranges, analyzing the resulting dynamical properties of the systems generated by this statistical ensemble of synaptic matrices with controlled and tunable properties. An intuitive geometric pictorial map of the parameter space is given in Fig. 5.2. Only the triplets (ρ, a_1, η) lying within (or on the boundary of) the prism define mathematically valid synaptic matrices. This framework makes it possible to span the entire range of possible realization of the coupling matrix $\mathbf{J}$, from fully symmetric ($\eta = 1$) to fully anti-symmetric ($\eta = -1$) to maximally sparse or maximally asymmetric networks in a complete and clear manner. The prism encompasses all valid parameter combinations (ρ, a_1, η), The triangular base reflects the constraint $a_1 \leq a_1^{\max}(\rho)$ intrinsic to directed networks as explained in Eq. (5.10) and Fig. 5.1. This framework provides a complete characterization

of network architectures, distinguishing between topological asymmetry (a_1) and synaptic symmetry (η), a distinction often overlooked in previous literature. Indeed, we want underline that until now no clear and complete description of the parameters was made especially in the distinction between the asymmetry in coupling and the asymmetry in the topology.

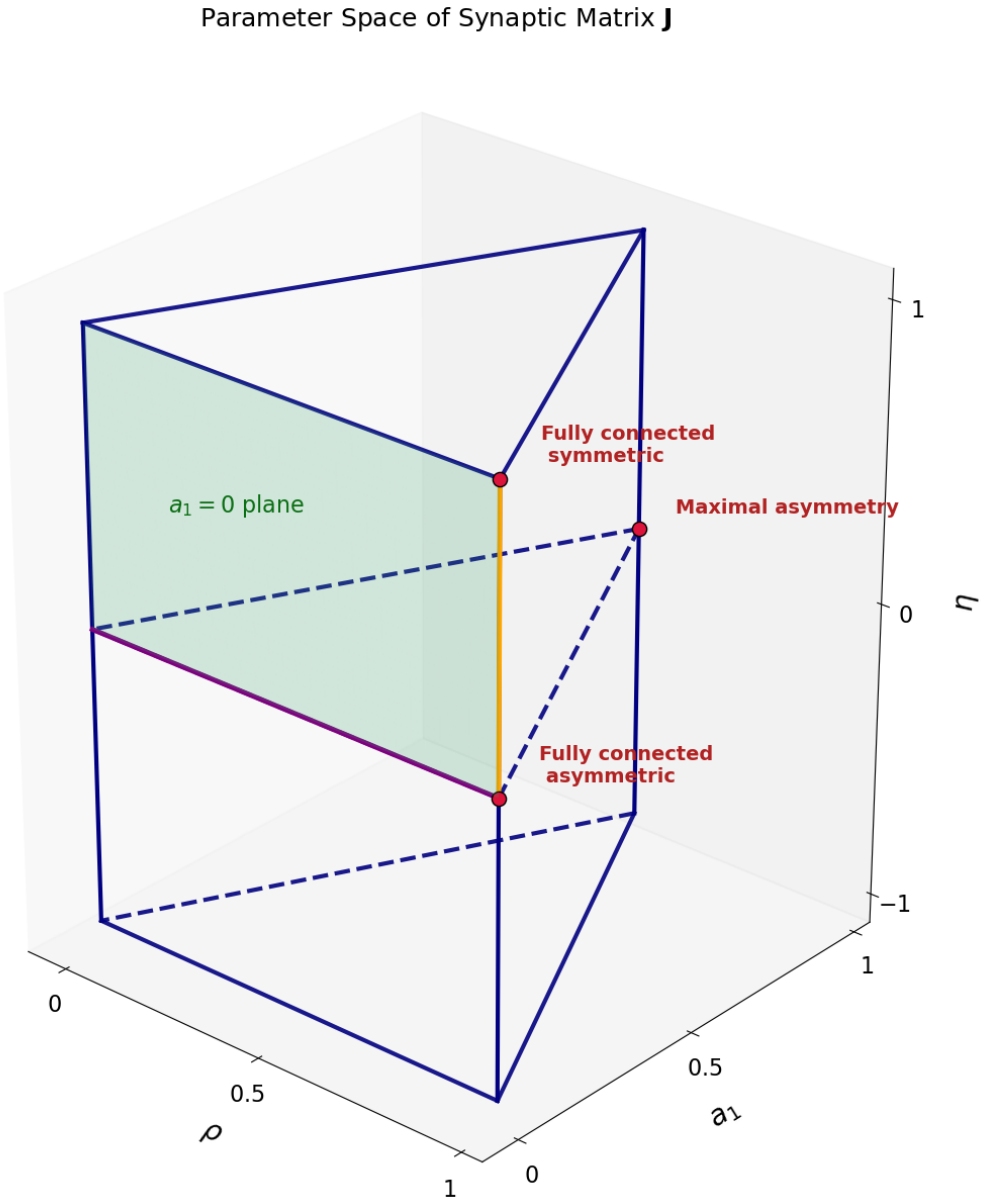


Figure 5.2 Parameter Space of Synaptic Matrix J. Geometric representation of the admissible parameter space for the synaptic matrix **J**. Each point in the prism corresponds to a network architecture specified by the triplet (ρ, a_1, η) where ρ represents connection density, a_1 controls topological asymmetry, and η governs synaptic correlation. The allowed region of parameter space forms a prism with triangular base bounded by the maximum asymmetry $a_1^{\max}(\rho)$ (see Eq. (5.10) and Fig. 5.1). Special points highlight distinguished architectures.

5.3 Key Observables

Our work characterizes the fundamental properties of finite systems and their dependence on the structural parameters of the coupling matrix **J**. For each triplet of parameters (ρ, η, a_1), we generate the synaptic matrix **J** following the methodology discussed in Sec. 5.2. The system's dynamics are then simulated by evolving all 2^N possible initial configurations according to the deterministic update rule in Eq. (5.1). This process generates a directed graph representation of the state space, where each vertex corresponds to a specific configuration $\mathcal{C}(t) \equiv \boldsymbol{\sigma}(t)$, and directed edges connect states to their subsequent images under the dynamics, $\boldsymbol{\sigma}(t) \to \boldsymbol{\sigma}(t+1)$. For a comprehensive view of the system's long-term behavior we refer to Fig. 5.3.
To completely characterize the dynamical landscape induced by the deterministic dynamics of Eq. (5.1), we introduce and analyze the following fundamental observables derived from the state transition graph. For statistical tractability, we generate S independent realizations ($\alpha = 1, \ldots, S$) of the connectivity matrix **J** for each parameter combination (ρ, η, a_1). So we will distinct to average procedures:

- $\langle \mathcal{O} \rangle$: Sample average over attractors within a single realization of **J**
- $\overline{\langle \mathcal{O} \rangle}$: Disorder average over different realizations of **J**

Due to substantial sample-to-sample fluctuations, we also compute the typical value $e^{\overline{\langle \ln \mathcal{O} \rangle}}$, which better represents the median of the distribution than the simple arithmetic mean.
The Observables analyzed are:

- **Number of Attractors ($C(\alpha)$):** Each attractor corresponds to a disjoint component (fixed point or limit cycle) in the transition graph, reflecting the system's long-term asymptotic states.
- **Cycle Lengths (L_i):** For each attractor $i = 1, \ldots, C(\alpha)$ in a given sample α, the integer L_i denotes the period of the cycle, i.e., the number of unique states in the i-th attractor.
- **Basin Sizes (Ω_i):** The basin size Ω_i quantifies the number of initial conditions whose evolution leads to the i-th attractor, thus measuring the domain of influence of each attractor.
- **Mean Distances (MD_i):** For each attractor, MD_i represents the average transient length (number of time steps) required for trajectories in its basin to reach the cycle for the first time.

Here, the index $i = 1, \ldots, C(\alpha)$ ranges over all attractors in a given realization of **J**. For a more global characterization of the system, we also consider the following derived quantities:

- **Effective Number of Attractors (N_{eff}):** This measures attractor diversity, defined as the inverse participation ratio over basin sizes:

$$N_{\text{eff}} = \frac{1}{\sum_{i=1}^{C(\alpha)} \left(\frac{\Omega_i}{2^N} \right)^2}, \tag{5.20}$$

- **Effective Cycle Length (L_{eff}):** The mean cycle length, weighted by each basin's relative size:
$$L_{\text{eff}} = \sum_{i=1}^{C(\alpha)} \frac{\Omega_i}{2^N} L_i. \tag{5.21}$$
with N the total number of states.

- **Chaoticity Parameter (p_{chaos}):** The fraction of pairs of neighboring initial states (differing by a single bit) that evolve toward distinct attractors. This variable quantifies the system's sensitivity to initial conditions. Its formulation is derived using the algorithm proposed in Sec. B.1.3.

The effective number of attractors N_{eff} and effective cycle length L_{eff} are calculated using basin weights $W_i = \Omega_i/2^N$, which represent the fraction of configuration space, including initial conditions, transient states and cycle configurations, flowing to the i-th attractor. Following [68], we define the participation ratio:

$$Y = \sum_{i=1}^{C(\alpha)} W_i^2, \tag{5.22}$$

which quantifies the probability that two randomly chosen initial configurations converge to the same basin. Consequently, the effective number of attractors is then given by $N_{\text{eff}} = 1/Y$. If $Y \to 0$ as $N \to \infty$, it indicates that in thermodynamic limit, the landscape is dominated by increasingly many basins with diminishing weights. Conversely, if Y remains finite as $N \to \infty$, it implies that there a finite number of relevant attractors with definite and non-vanishing weight (basin) covering almost the entire configuration space. The number of effective attractors N_{eff} is related to the localization of configurations since it takes into account the fraction of the total number of them that ends in a particular basin. If all basins are equal[3], then $N_{\text{eff}} \equiv C(\alpha)$. Also for the cycle length, if in a sample α all basins are equal then $L_{\text{eff}} \equiv \langle L \rangle$. The extreme values of Y, and consequently of N_{eff}, represent the situation with only one relevant attractor ($Y = 1$) and the case with an infinite number of relevant disjoint components ($Y = 0$).
A generalization for large systems proposed in [20, 68, 45] involves calculating the moments of the basin weights defined by:

$$Y_n = \sum_i W_i^n, \tag{5.23}$$

by counting the number of times that, in a given sample (for a given $\mathbf{J}$), n randomly chosen initial configurations fall onto the same attractor[4].
In Table 5.1 we summary of all observables analyzed in this study.

[3] $\sum_{i=1}^{C(\alpha)} W_i^2 = C(\alpha)W^2 = C(\alpha)\left(\frac{\Omega_i}{2^N}\right)^2 = C(\alpha)\left(\frac{2^N}{2^N C(\alpha)}\right)^2 = \frac{1}{C(\alpha)}$.

[4] The case $n = 1$ is the normalization, while for simplicity Y without subscript refers to $n = 2$.

Quantity	Description
$C(\alpha)$	Number of attractors (fixed points or cycles) per sample α
L_i	Length of the i-th attractor cycle
Ω_i	Basin size of the i-th attractor (number of converging ICs)
MD_i	Mean distance to the i-th attractor (average transient time)
N_{eff}	Effective number of attractors, weighted by basin sizes
L_{eff}	Effective cycle length, averaged over basin weights
p_{chaos}	Fraction of neighboring ICs converging to different attractors

Table 5.1 Key observables characterizing the dynamics of finite neural networks.

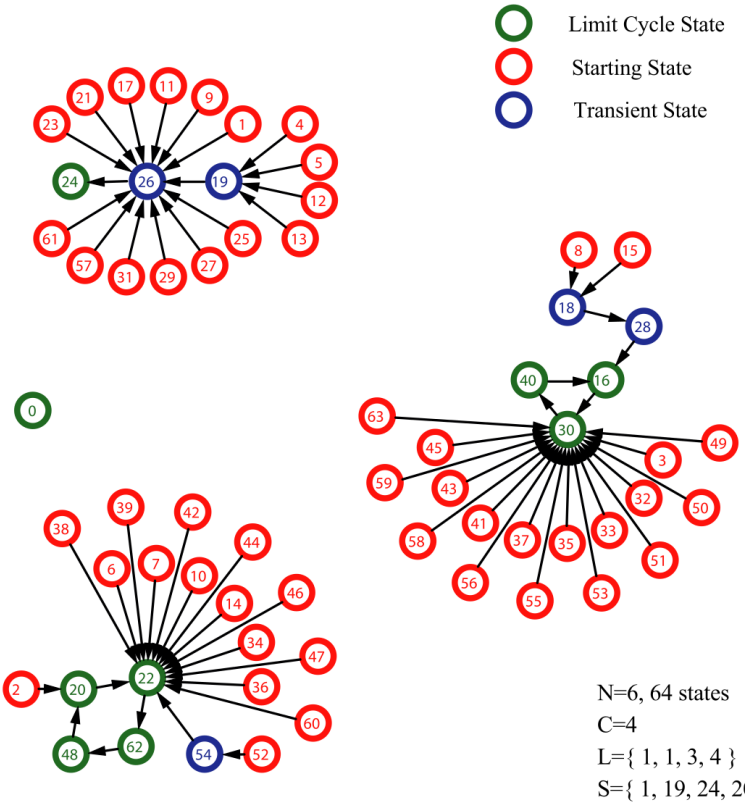


Figure 5.3 State transition graph for $N = 6$ nodes, illustrating attractor cycles (green), transient states (blue), and starting states (red). The network decomposes into several basins of attraction whose properties are quantitatively summarized by the key observables described in Table 5.1. Figure from [54].

If we define:

- Event A: "Two configurations end in different basins"
- Event B: "Two configurations end in the same basin" (the complement A)
- Condition C: "The two configurations are nearest neighbors" ($q_{\sigma\tilde{\sigma}} \sim 1$)
- Condition D: "The two configurations are randomly picked"

we have:

$$Y = P(B|D) = 1 - P(A|D) \quad \& \quad p_{\text{chaos}} = P(A|C). \tag{5.24}$$

Ending up in the same attractor does not necessarily imply a high overlap between the two configurations, as the dynamics is random.

5.4 Comparison with Previous Work

Our work builds upon, but also methodologically diverges from, previous key studies in the literature. It is therefore important to highlight these distinctions, as they explain quantitative discrepancies and underscore the comprehensiveness of our approach:

- **Finite-Size vs. Thermodynamic Limit:** We emphasize that our analysis focuses on finite system sizes, where the deterministic nature of the dynamics and finite configuration space guarantee that all trajectories converge to periodic attractors. This contrasts with thermodynamic limit approach (the limit $N \to \infty$ is taken before the limit $t \to \infty$), where single-spin dynamics in infinite systems can be studied via Monte Carlo simulations [130, 116, 51, 50]. Such methods have revealed, for instance, a phase transition in the Sherrington-Kirkpatrick model at $T = 0$ for asymmetric couplings, occurring at a critical symmetry parameter $\eta^* = 0.825$. For $\eta > \eta^*$, a state initially magnetized never forgets the initial configuration and has a non-zero remanent magnetization, while for lower symmetry ($\eta < \eta^*$), the dynamics becomes chaotic, the initial condition is completely forgotten and there is a complete separation of trajectories with high initial overlap. Our work instead characterizes the fundamental properties of finite systems and their dependence on the structural parameters of $\mathbf{J}$.

- **Spin Representation:** We use Ising spin variables ($\sigma_i = \pm 1$) rather than the McCulloch-Pitts neuronal states ($S_i = 0, 1$) used in works like [54, 87]. As noted in [106], this choice leads to quantitative differences in results, a crucial factor when comparing our findings with prior literature.

- **Exhaustive Enumeration vs. Statistical Sampling:** Our computational strategy is based on exhaustive enumeration: for each system size N and for each realization of $\mathbf{J}$, we simulate the deterministic dynamics from all 2^N initial configurations. This is complemented by extensive disorder averaging over $S \sim \mathcal{O}(10^3 - 10^4)$ independent matrix realizations. While computationally intensive, this provides a complete and unbiased characterization of the entire attractor landscape. This contrasts with studies that use statistical sampling by evolving only a subset of random initial conditions, which is necessary for much larger systems but may miss rare attractors [37, 21, 106, 36].

- **Finite-Size Focus:** We explicitly focus on the properties of finite systems, where finite-size effects are significant for $N \lesssim 16$. Consequently, our quantitative analysis is restricted to $N \geq 18$ to ensure robust conclusions.

5.5 Organization of the Results of Chapter 6

Having established theoretical framework and key observables, we now present the roadmap for our numerical investigation of the (ρ, η, a_1) parameter space (Fig. 5.2). Our exploration is finalized to understand how dynamics are influenced by dilution, synaptic symmetry, and topological asymmetry. The following Chapter contains a complete exploration of the entire parameter space, in Fig. 5.2, defined by the synaptic matrix $\mathbf{J}$. Here we briefly anticipate and describe the content of each Section.

- **Section 6.1: Benchmarking Against Known Regimes** We first validate our approach by examining two previously studied limits, which serve as benchmark cases for comparison with existing literature [87, 54, 76, 77]:
  - **The Dense Symmetry Axis (Yellow trajectory in Fig. 5.2):** Fully connected networks ($\rho = 1$) across the entire synaptic symmetry spectrum, from asymmetric ($\eta = 0$) to symmetric ($\eta = 1$). Details in Sec. 6.1.2.
  - **The Diluted Asymmetric Axis (Purple trajectory Fig. 5.2):** Networks with maximally asymmetric couplings ($\eta = 0$) across a wide range of connection densities ρ, from sparse to dense. Analyzed in Sec. 6.1.3.

  Revisiting these limits allows us to confirm the consistency of our methodology and provides a baseline for understanding the more complex interplay of parameters.

- **Section 6.2: Absence of direct links $a_1 = 0$ (Green plane in Fig. 5.2)** We then isolate the effect of synaptic correlations by analyzing the constrained case $a_1 = 0$. In this regime, connections are either bidirectional or absent, eliminating topological asymmetry. We investigate how the dynamics evolve as the synaptic couplings vary from independent ($\eta = 0$) to perfectly correlated ($\eta = 1$).

- **Section 6.3: The Full Parameter Space** Finally, we present the complete exploration of the three-dimensional parameter space. By varying (ρ, η, a_1) simultaneously across their admissible ranges, we reveal the rich and often non-trivial interplay between connection density, synaptic symmetry, and topological asymmetry. This comprehensive analysis extends our understanding beyond the specific architectures studied previously, providing a unified picture of dynamics in asymmetric, diluted spin systems and expanding our understanding to the full complexity of network architectures achievable through this generalized framework.

Chapter 6

Numerical Results in Asymmetric Diluted Networks

All we need to do is make sure we keep talking.»

– *"Talkin' Hawkin"*, S. Hawking. *The Endless River*, Pink Floyd.

As discussed in the previous Chapter, the characteristics of the network output, such as the number and length of cycles, transient time, mean distance from the attractor, and basin of attraction dimension, are primarily influenced by the structure of the synaptic coupling matrix $\mathbf{J}$. The generalized framework introduced in Sec. 5.2, parameterized by the triplet (ρ, η, a_1), allows a systematic exploration of network architectures across the entire spectrum of connection density, synaptic symmetry, and topological asymmetry. In this Chapter we present a comprehensive analysis of how these structural parameters shape the attractor landscape and dynamical behavior of finite neural networks. We fully analyze the networks properties exploring the full parameter space prism of Fig. 5.2. Following the roadmap outlined in Sec. 5.5, we begin by validating our approach against established results in previously studied parameter regimes (Sec. 6.1), then progressively expand our investigation to the full parameter space (Sec. 6.2 for the case $a_1 = 0$ and in Sec. 6.3 for the general case). Our analysis reveals how the interplay between dilution (ρ), synaptic correlation (η), and topological asymmetry (a_1) gives rise to distinct dynamical phases, from ordered regimes dominated by stable attractors to complex landscapes with extensive cycling behavior.

6.1 Numerical Results in Previously Studied Regimes

We report here the numerical results obtained in the regions already explored in the literature serving both to validate our implementation and to establish baseline behaviors. We collect in this section the results for the topologically symmetric situation $a_1 = 0$, where connections are either bidirectional or absent. In particular, we analyze the fully connected scenario ad different values of correlation ($\rho = 1$, $\eta \in [0, 1]$) in Sec. 6.1.2. Then, we explore the fully asymmetric case at different values of density ($\eta = 0$, $\rho \in (0, 1]$) in Sec. 6.1.3. These two regions correspond to the highlighted purple and yellow lines in Fig. 5.2. The special case of a fully connected

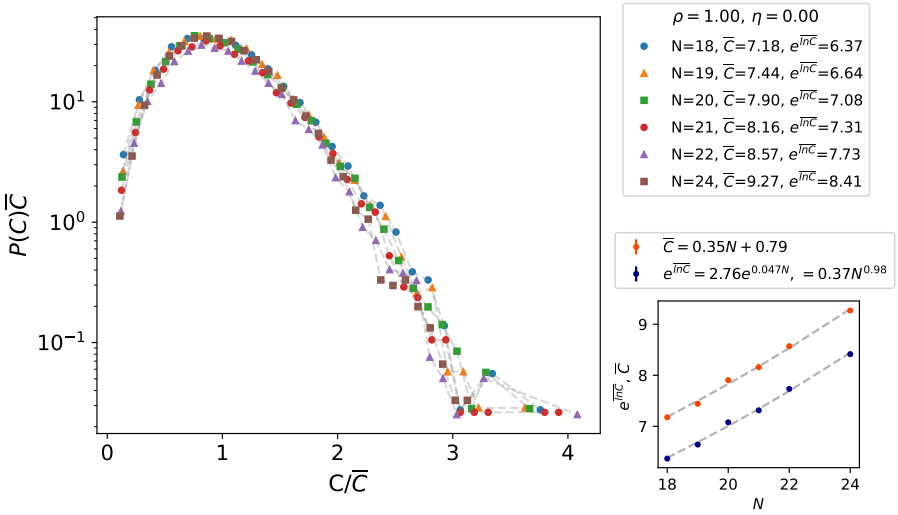
and fully asymmetric case, corresponding to the point $(\rho = 1, \eta = 0, a_1 = 0)$ in the space parameter prism is exposed in Sec. 6.1.1. This entire section wants to became a sort of brief collections of all the results in the preexisting explored parameters space in order to make more visible the state of art of knowledge and provide new analyses and comprehensive characterization of these established regimes.

6.1.1 Fully Connected and Fully Asymmetric Neural Networks

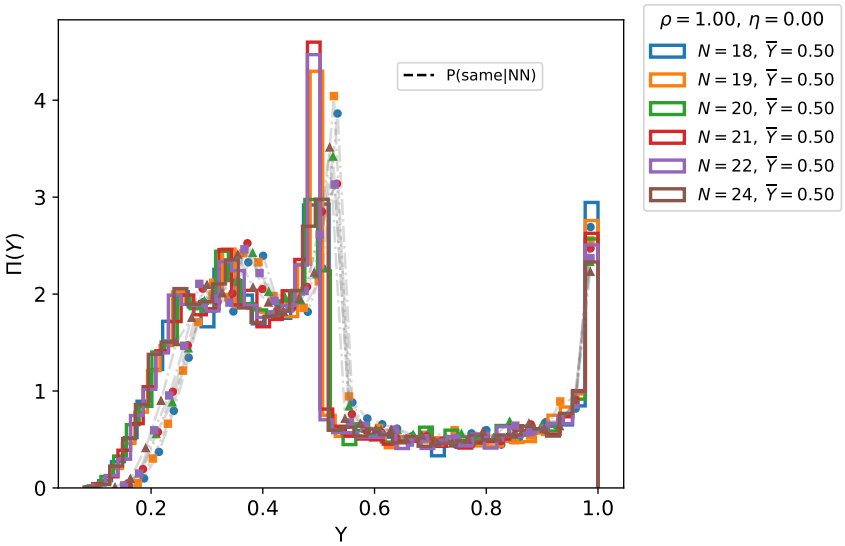
We begin with the paradigmatic case of fully connected, fully asymmetric networks, corresponding to parameters $\rho = 1$, $\eta = 0$, $a_1 = 0$. This case represents a fundamental reference point in this study.

Attractor Statistics and Basin Structure

Figure 6.1a displays the scaled distribution of attractor counts $P(C/\overline{C})$, where $\overline{C}$ denotes the mean number of attractors averaged over S independent realizations of the coupling matrix. The inset demonstrates the linear scaling of $\overline{C}$ with system size N, consistent with predictions from random map theory [45, 44]. This linear growth aligns with the hypothesis of effectively random dynamics in this parameter regime. However, as previously observed in [20], the basin statistics deviate from the random map prediction. The mean value $\overline{Y} = 1/2$ and all moments of the distribution $\Pi(Y)$ are significantly smaller than the random map expectation of $\overline{Y} = 2/3$. The histogram in Fig. 6.1b reveals that the system decomposes into relatively few disjoint components, with the distribution of the effective number of attractors $\overline{N_{\text{eff}}}$ being size-independent and concentrated around values 1 and 2. We find $\overline{N_{\text{eff}}} = 2.5$ consistently across all simulated system sizes. The probability distribution $\Pi(Y)$ exhibits distinctive singularities at $Y = 1/n$ for $n \geq 1$, characteristic of this dynamical regime. This singular structure reflects the hierarchical organization of basin sizes. Regarding the number of fixed points C_1, we have found that $\overline{C_1} = 1$ for all sizes N as already observed in [68].



(a) Scaled distribution of attractor counts $P(C/\overline{C})$. Inset: linear scaling of $\overline{C}$ with system size N.

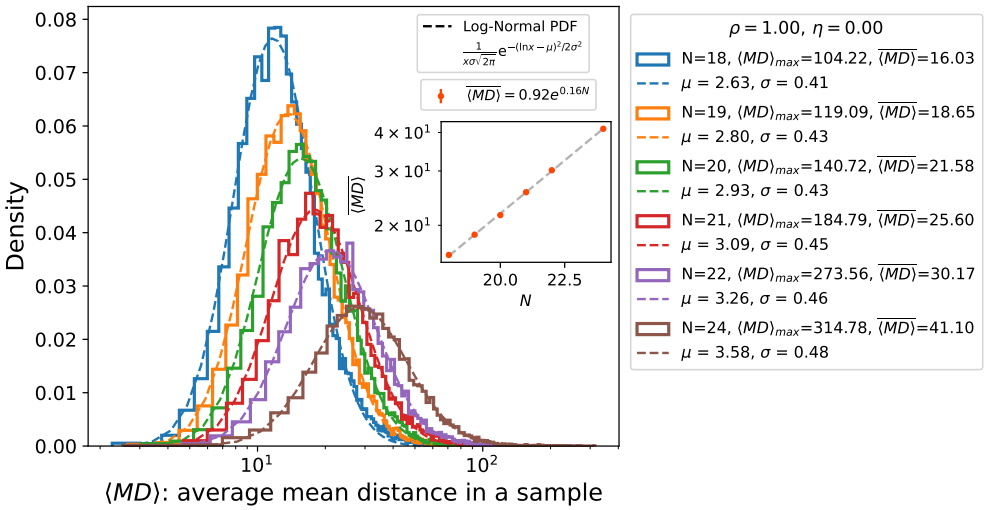


(b) Distribution $\Pi(Y)$ compared to the probability $P(\text{same}|\text{NN})$ that neighboring initial configurations converge to the same attractor.

Figure 6.1 Attractor statistics for fully connected, fully asymmetric networks (a) The linear scaling of $\overline{C}$ with N agrees with random map predictions. (b) The distribution $\Pi(Y)$ shows distinctive singularities with the mean value $\overline{Y} = 0.5$ different from the expected value of a Random Map.

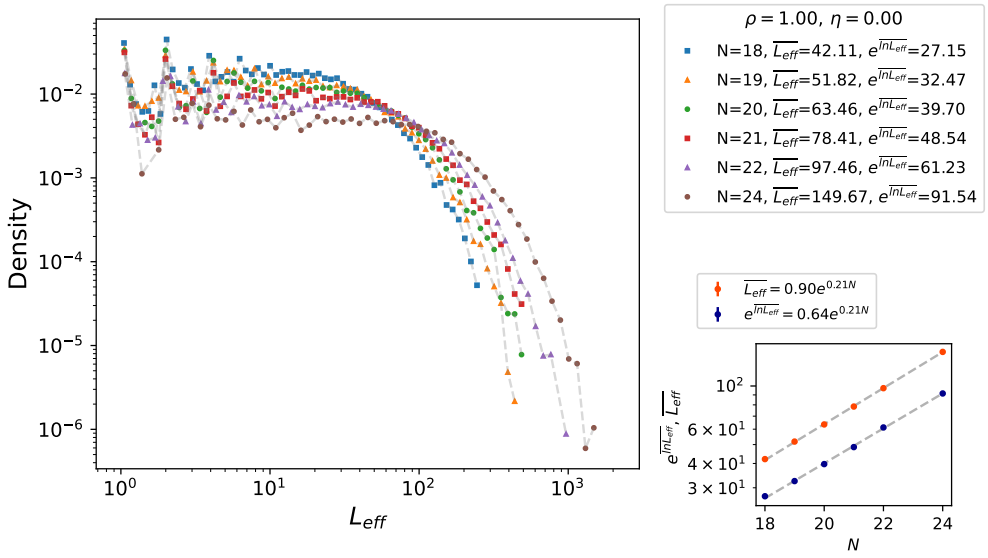
Transient Dynamics and Cycle Properties

The transient dynamics, characterized by the mean distance to attractors $\langle MD \rangle$, follows a Log-Normal distribution with parameters μ and σ, yielding a mean value $\mathbb{E}[\langle MD \rangle] = e^{\mu+\sigma^2/2}$. As shown in Fig. 6.2a, the average transient length $\overline{\langle MD \rangle}$ grows exponentially with system size, consistent with the complex, multi-step relaxation processes characteristic of disordered systems.



(a) Distribution of mean distances $\langle MD \rangle$ with Log-Normal fits. Inset: exponential growth of $\overline{\langle MD \rangle}$ with N.

In this regime, the mean cycle length $\overline{\langle L \rangle}$ grows exponentially with system size as $\overline{\langle L \rangle} \sim A_L \exp(\alpha_L N)$, with $\alpha_L = 0.21 \pm 0.01$ obtained from the fit. This exponent is compatible with previous observations [20, 106, 76]. The distribution of the cycle length statistics and the correlated fit is in Fig. 6.3a. The typical cycle length $e^{\overline{\ln L_{\text{eff}}}}$ shows similar exponential growth, indicating that the scaling behavior is robust across different averaging procedures.



(a) Distribution of effective cycle lengths L_{eff} with exponential fits for $\overline{L_{\text{eff}}}$ and $e^{\overline{\ln L_{\text{eff}}}}$.

Figure 6.3 Transient and cycle statistics for fully connected, fully asymmetric networks. (a) Mean distances follow Log-Normal distributions and grow exponentially with system size, while (b) cycle lengths exhibit exponential growth with exponent $\alpha_L = 0.21$.

6.1.2 Fully Connected Networks with Varying Coupling Correlation

We now investigate how synaptic correlation η modifies the properties of the system in fully connected networks ($\rho = 1$). As η increases from the fully asymmetric limit ($\eta = 0$), we observe systematic changes in both basin structure and cycle statistics that signal a passage from complex to ordered dynamical regimes.

Basin Structure Evolution and Chaoticity

The distribution $\Pi(Y)$ undergoes significant transformation with increasing correlation. For $\eta \lesssim 0.33$, $\Pi(Y)$ exhibits the characteristic singularities at $Y = 1/n$ previously observed in the fully asymmetric case. However, as shown in Figure C.5, these singularities disappear for $\eta \gtrsim 0.33$, indicating a fundamental change in basin organization. For $\eta \gtrsim 0.5$, the distribution becomes sharply peaked around $Y \to 0$, reflecting a landscape dominated by numerous small basins. We also report the probability $P(\text{same}|\text{NN})$ that neighboring initial configurations converge to the same attractor that remains finite across all correlation values. Regarding the chaoticity p_{chaos} of the system, we show in Fig. C.6 the transformation in the distribution

increasing the correlation η. The global behavior mirrors to the one of the distribution $\Pi(Y)$. In this figure we can see that for $\eta \leq \eta_d \approx 0.33$, distributions are size-independent and exhibit singularities at $p = 1 - 1/n$ for $n \geq 1$ (panels a-c). These singularities disappear at $\eta \simeq 0.33$, and for $\eta \gtrsim 0.5$, distributions concentrate around increasing values of p_{chaos}. The comparison with $P(\text{diff}|\text{random})$ provides additional insight into sensitivity to initial conditions.

Cycle Length Statistics

The scaling of cycle lengths with system size reveals a clear correlation-induced change. For $\eta \gtrsim 0.5$, both the simple average $\overline{\langle L \rangle}$ and basin-weighted average $\overline{L_{\text{eff}}}$ become essentially size-independent, with $\overline{\langle L \rangle} \sim 2$ and $\alpha_L \to 0$, indicating the dominance of short cycles (primarily fixed points and period-2 oscillations). The distributions do not present long tails and are sharply peaked around their mean values. In contrast, for low correlation regimes ($\eta \lesssim 0.33$), cycle length distributions are dominated by rare but extremely long cycles, bringing to substantial differences between mean and typical values. Quantitative analysis reveals distinct scaling behaviors. We performed both a power law $\overline{\langle L \rangle} = A_L N^{\beta_L}$ and an exponential fit $\overline{\langle L \rangle} = A_L \mathrm{e}^{\alpha_L N}$ obtaining $\beta_L = 0.11$ at $\eta = 0.33$, $\alpha_L = 0.09$ and $\beta_L = 1.91$ at $\eta = 0.20$ (compatible with both power-law and exponential fits), and $\alpha_L = 0.16$ in the fully asymmetric case $\eta = 0$. These exponents are consistent with previous studies [106] and reflect the persistence of complex cycling behavior in weakly correlated networks. The systematic suppression of long cycles with increasing η demonstrates how synaptic correlation serves as a control parameter. The crossover occurs gradually rather than at a sharp critical point, with the most significant changes occurring in the range $0.3 \lesssim \eta \lesssim 0.5$ as will be described in Sec. 6.2.

6.1.3 Fully Asymmetric Networks with Varying Dilution

We now examine how connection density ρ modulates the dynamical landscape along the fully asymmetric axis ($\eta = 0$). Dilution introduces structural sparsity that fundamentally alters both the attractor statistics and consequently the cycling behavior, revealing a density-dependent crossover in network dynamics.

Evolution of Basin Statistics and Chaoticity with Dilution

The distribution $\Pi(Y)$ exhibits systematic changes as the network becomes increasingly sparse. For moderate dilution ($\rho \gtrsim 0.5$), the basin structure remains qualitatively similar to the fully connected case (compare Fig. 6.1b), preserving the characteristic features of dense asymmetric networks. However, a significant change occurs around $\rho \sim 0.2$, where $\Pi(Y)$ becomes sharply peaked near $Y \to 0$ (Fig. C.3b). This concentration of probability mass at small Y values indicates the emergence of an extensive number of attractors with really small basin sizes. The behavior of $P(\text{same}|\text{NN})$, the probability that neighboring initial configurations converge to the same attractor, reveals that local similarity in initial conditions poorly predicts asymptotic dynamics in highly diluted regimes. This loss of local predictability, shown by the dotted curves in all figures reporting p_{chaos}, suggests that dilution promotes state space fragmentation. In this regime, we observe no

significant variation in the network's chaoticity distribution compared to the dense case until substantial dilution ($\rho \sim 0.5$). In Fig. C.4 we show the evolution of chaoticity distributions $p_{\mathbf{chaos}}$ with increasing dilution. A significant distinction emerges between the two sensitivity measures: while p_{chaos} reaches a maximum value of approximately 0.8 across all dilution levels, the probability $\overline{P(\text{diff}|\text{random})}$ approaches unity as ρ decreases.

Cycle Length Distribution and Scaling

Dilution profoundly affects cycle length statistics and their scaling with system size. At high densities ($\rho \sim 0.5$), the distribution is dominated by heavy tails that produce exponential growth of mean cycle lengths with N, reminiscent of the fully connected asymmetric case. However, as dilution increases beyond $\rho \lesssim 0.3$, the distribution undergoes a qualitative transformation. In highly diluted regimes ($\rho \sim 0.1$), cycle length distributions become concentrated around $L \sim 4$, with substantially reduced tail contributions. A low connectivity suppresses the emergence of very long cycles, characteristic of dense asymmetric networks, and concentrates the distribution around characteristic short periods. The observed dilution-induced crossover from exponentially growing cycles with heavy-tailed distributions to concentrated distributions around characteristic short periods, demonstrates how structural sparsity can regulate dynamical complexity. This density-dependent control mechanism operates independently of synaptic correlation effects as will be visible in the entire map of the plane $a_1 = 0$ in Fig. 6.5.

6.2 Results for $\mathbf{a_1 = 0}$: Absence of Directed Links

Now we concentrate on the results obtained when directed (unidirectional) links are absent ($a_1 = 0$). In this constrained regime, the probability of double (bidirectional) links equals the connection density ($a_2 = \rho$), and consequently, the probability of no connection is $a_0 = 1 - \rho$. This specific topological constraint, where the network connectivity is symmetrically diluted when $\rho < 1$, allows a direct comparison with the formulation employed in prior works [54, 87]. A detailed mapping between our generalized parametrization (η, ρ, a_1) and the notation used in those studies is provided in Appendix C.1 where we show that their approach implicitly assumed $a_1 = 0$ always: if a link $i \to j$ is absent, necessarily also the reciprocal link $j \to i$ is not present.

By systematically varying both synaptic correlation η and connection density ρ, we construct a complete phase diagram, the one schematically highlighted in the green plane of Fig. 5.2, characterizing the fundamental properties of attractor landscapes, including cycle length statistics, chaoticity measures, and basin size distributions.

Our analysis start with a general exploration of the trends in the average values of the observables in Tab. 5.1 as functions of the parameters (ρ, η), as reported in Figs. (C.9-C.13). To quantify scaling behavior, we performed exponential fits of key observables[1] as functions of system size N, extracting growth exponents α_i (with $i = C, L$ for attractor count and cycle length respectively). We established

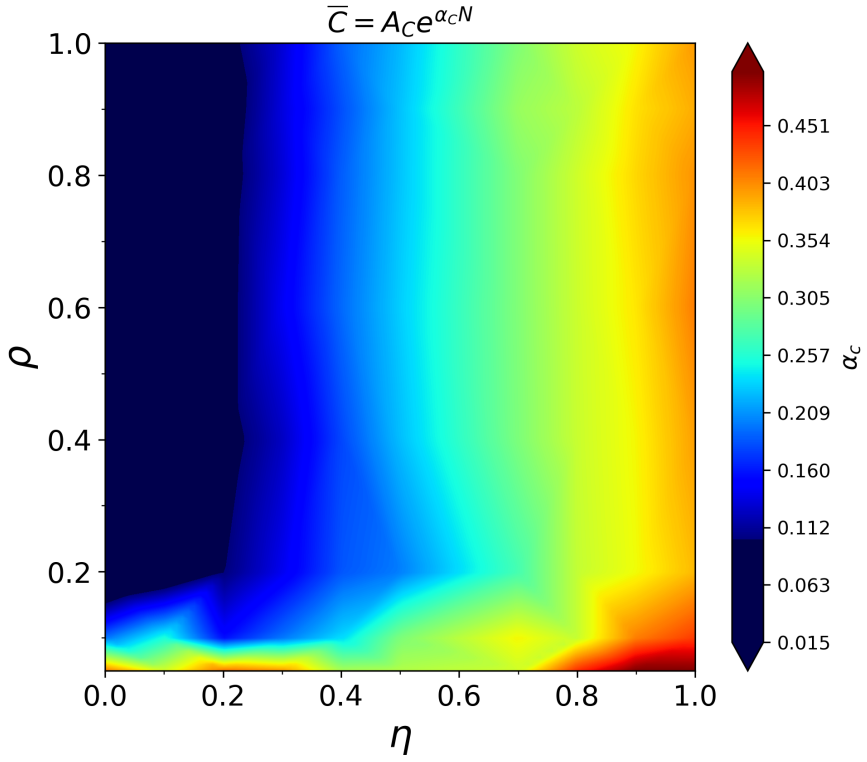
[1] We have also analyzed the behavior of typical values.

thresholds of $\alpha_C = 0.10$ and $\alpha_L = 0.05$ to distinguish significant exponential growth, complemented by power-law fits with threshold $\beta_i = 2$. These boundaries are based on fit validity criteria: power-law scaling becomes nonphysical for $\beta > 2$, while exponential growth is negligible for $\alpha < 0.05, 0.1$. This approach allows us to distinguish between chaotic and ordered regimes.

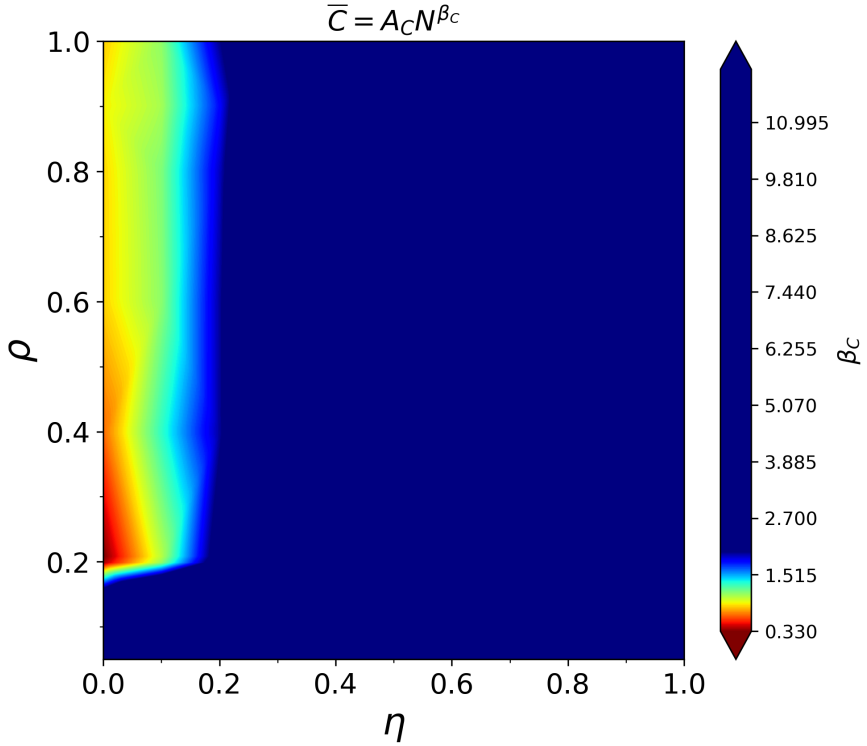
Scaling of the Mean Number of Attractors

The phase diagrams in Figure 6.4 reveal how attractor proliferation depends on the structural parameters (ρ, η) and that the region of validity of an exponential growth of the mean number of attractors $\overline{C}$ with the system size N, is complementary to the one governed by a polynomial or power law scaling. The exponential growth rate α_C (top panel) reaches maximum values in two distinct regimes: (1) highly correlated systems ($\eta \to 1$) with increasing density and, more surprisingly, (2) highly diluted networks ($\rho \ll 1$) across the entire range of correlation values η. While the first aligns with classical symmetric systems behavior, the second represents an interesting characteristic which mimic well biologically plausible sparse networks.
In the fully correlated, fully connected limit ($\eta = 1, \rho = 1$), we find $\alpha_C \simeq 2\Sigma_{SK}$, where $\Sigma_{SK} = 0.1992$ is the complexity of metastable states in the Sherrington-Kirkpatrick model [28, 136]. Indeed, with these values of the parameters, the network is equivalent to a Sherrington-Kirkpatrick system up to a rescaling of the J that are by definition of $O(1)$ in our case while have variance of $O(1/N)$ in the original model. However, this factor is not relevant in the evaluation of the configurational entropy since the dynamics considers just the sign of the local field $h_i(t)$ applied to the spin variables. The factor 2 is a direct consequence of the parallel upgrade that causes the presence of only cycles of length $L = 1, 2$ in the fully symmetric case $\eta = 1$ with the dominant contribution coming from 2-cycles. From both simulations and analytic computation we have verified that the complexity of fixed point Σ_1 is the same of the SK model $\Sigma_1(\eta = 1, \rho = 1) \equiv \Sigma_{SK}$. The complexity of 2-cycles Σ_2 is instead related to $\alpha_C(\eta = 1, \rho = 1) \simeq \Sigma_2(\eta = 1, \rho = 1) \simeq 2\Sigma_{SK}$. The two quantities $\Sigma_{1,2}$ are numerically calculated from exponential fit of the growth with the size N of the mean number of fixed points or 2-cycles.

The power-law exponent β_C (bottom panel) reveals complementary structure of the phase diagram. Most of the parameter space is dominated by large values of β_C indicating that a power law behavior is not representative and valid to describe the attractors scaling. Except for a narrow region at low η and high ρ where β_C decreases sharply, marking a clear crossover in scaling behavior. Together, these two complementary phase maps reveal distinct regions in parameter space where either exponential or power-law scaling dominates, and they confirm the existence of critical boundaries that separate qualitatively different growth regimes for $\overline{C}$.
We collect in Figs. C.17 the corresponding colormaps of Figs. 6.4-6.5 obtained using the nomenclature adopted in [54, 87, 76] which represents a special case of our parametrization as discussed in Sec. C.1.



(a) Exponential growth rate α_C across parameter space

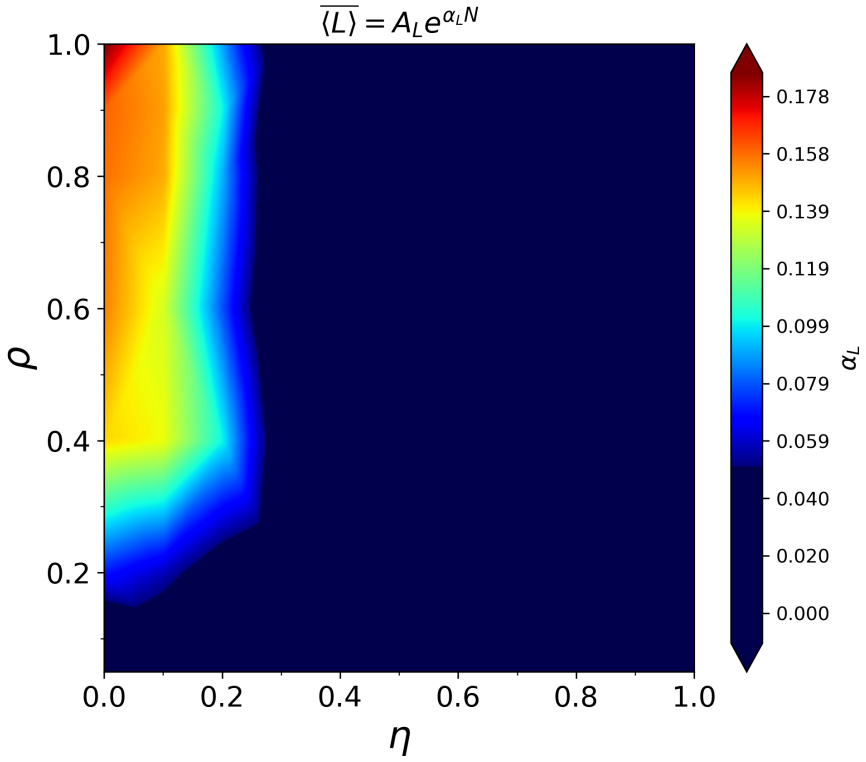


(b) Power-law exponent β_C across parameter space.

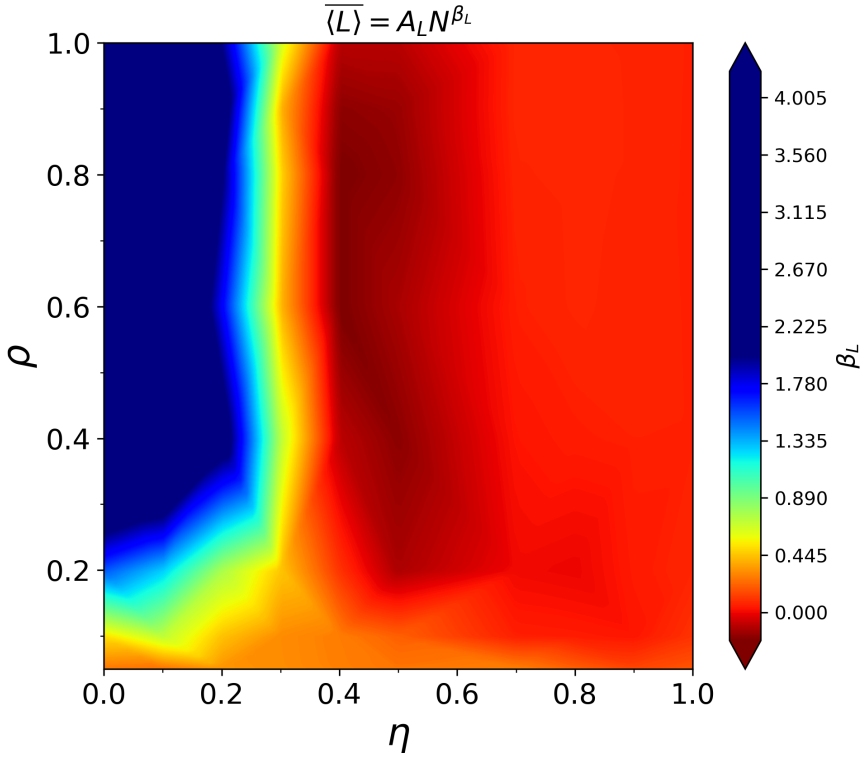
Figure 6.4 Scaling behavior of mean number of attractors $\overline{C}$. *(a)* Exponential growth rate α_C across parameter space. Dark blue indicates regions where exponential scaling becomes insignificant. *(b)* Power-law exponent β_C, with blue marking regions where $\beta_C > 2$ where the power law is no more acceptable.

Scaling of the Mean Cycle Lengths

The cycle length statistics in Fig. 6.5 exhibit complementary structure to the attractor count $\overline{C}$ analysis shown in Fig. 6.4 . In the top panel, the exponential growth rate α_L exhibits increasing values in a narrow region at low η and high ρ, indicating that the exponential scaling of $\overline{\langle L \rangle}$ with system size N occurs only in this specific region corresponding to the so called *chaotic regime*. The power-law exponent β_L (bottom panel) shows reasonable values across most parameter space except for the same chaotic region where it increases dramatically. Looking at the coefficient A_L and the power-law exponent β_L, we have noticed a strong negative correlation. In Fig. 6.5b is represented in light blue the region with $\beta_L \sim 1$ and in yellow the one with $\beta_L = 0.5$; for $\beta_L \sim 0.5$ the coefficient A_L concentrates around $A_L \sim 1$ corresponding to $\overline{\langle L \rangle} \sim \sqrt{N}$. This scaling regime represents an intermediate behavior between the short cycles of ordered phases and exponentially long cycles of the chaotic regime. Our results confirm and extend previous findings of a dynamical transition [21, 76] at $\eta_d \in (0.3, 0.4)$ for fully connected networks $\rho = 1$. For high correlation values $\eta > \eta_d$, typical cycle lengths are short $\overline{\langle L \rangle} \sim 2$. For $\eta_d < \eta \lesssim 0.5$, the typical cycles have still length 2, but their basins of attraction carry vanishing weights, causing the average cycle lengths to be dominated by the tails of the distribution [21, 106] (we have found in this region values of $A_L \geq 2$ due to these deviations from the typical values). In the supplementary Fig. C.1 we shows the distribution of the effective cycle length L_{eff} and its scaling of mean and typical values with system size N that is influenced by tail of the distribution. In any case, is already visible the difference with respect to the exponential scaling observed in the fully asymmetric case $\eta = 0$ (Fig. 6.3a). Indeed, for $0 < \eta \lesssim \eta_d$, the typical cycles are exponentially long with respect to the system size, and the basin weights are distributed as in a Random Map C.2 with reversal symmetry [21].
We have shown that this transition persists, at the same correlation, under substantial dilution, as evidenced by the sharp vertical boundary in the phase diagrams of Fig. 6.5. The value η_d divides the region in which $\beta_L \neq 0$ from the red area where $\beta_L \to 0$ and the mean cycle length values $\overline{\langle L \rangle}$ is fully influenced by the coefficient A_L. In general, even in the other colormaps (Fig. 6.4) is visible a clear changing in the behavior of the mean number of attractors $\overline{C}$, remarking the probability of a dynamical transition not only in the fully connected scenario $\rho = 1$ but even when dilution is added to the synaptic matrix. These results shows a clear crossover between the exponential and power-law scaling regimes at a critical value η_d that seems to be the same of the dense case. This changing in the shape of the cycle length distributions with increasing dilution and the effects on the mean values are reported in Fig. 6.6 where we show (for $\eta = 0$ but it's the same regardless to the correlation) the behavior of the weighed cycle lengths L_{eff}. In the top panel the density $\rho \sim 0.5$ therefore the distribution is still comparable to the fully connected scenario that is chaotic for low correlation η. Instead, in the bottom panel we can see the effect of dilution since the distribution is picked around $L_{\text{eff}} \sim 4$ and the mean and typical values are less influenced by distribution tails. Increasing the dilution we can drive the system toward ordered regimes dominated by short cycles and for $\rho \lesssim 0.2$ this behavior becomes η-independent. This demonstrates that extreme dilution effectively suppresses chaotic dynamics regardless of coupling symmetry.

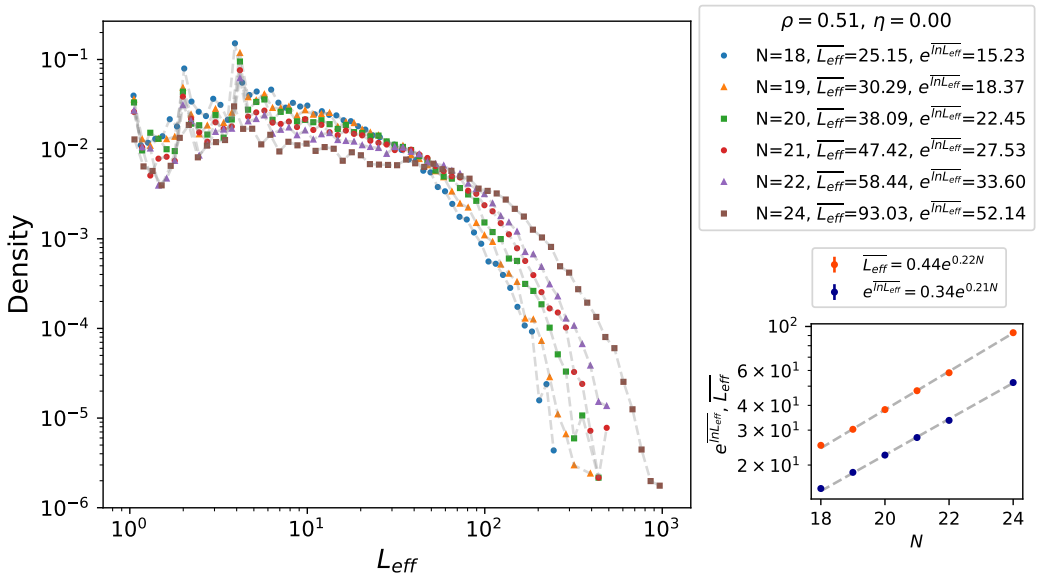


(a) Exponential growth rate α_L across parameter space

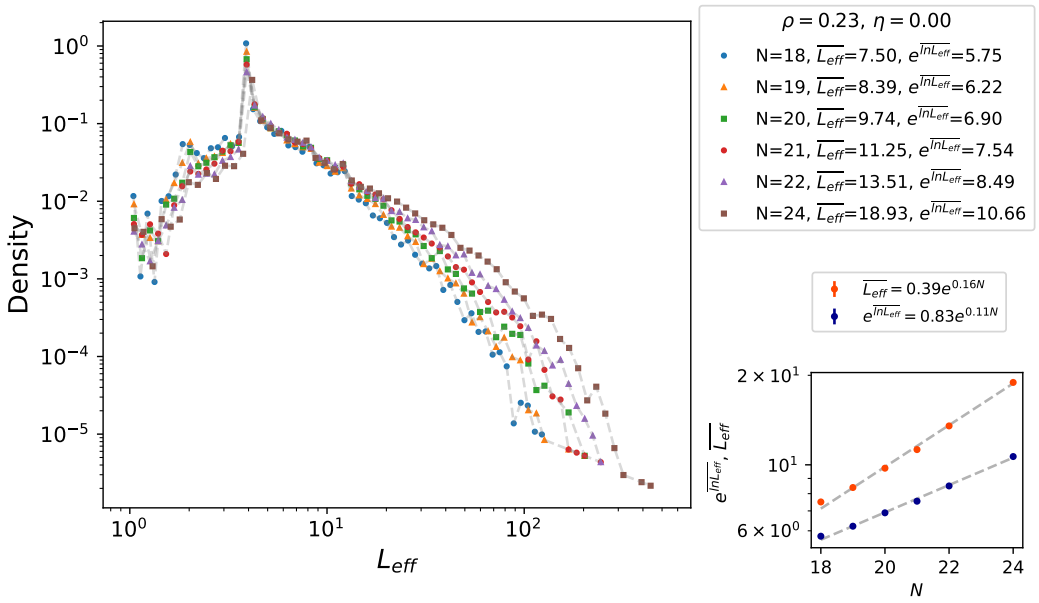


(b) Power-law exponent β_L across parameter space.

Figure 6.5 Scaling behavior of mean cycle lengths $\overline{\langle L \rangle}$. *(a)* Exponential growth rate α_L across parameter space. Dark blue indicates regions where exponential scaling becomes insignificant. *(b)* Power-law exponent β_L, with blue marking regions where $\beta_L > 2$ where the power law is no more acceptable.

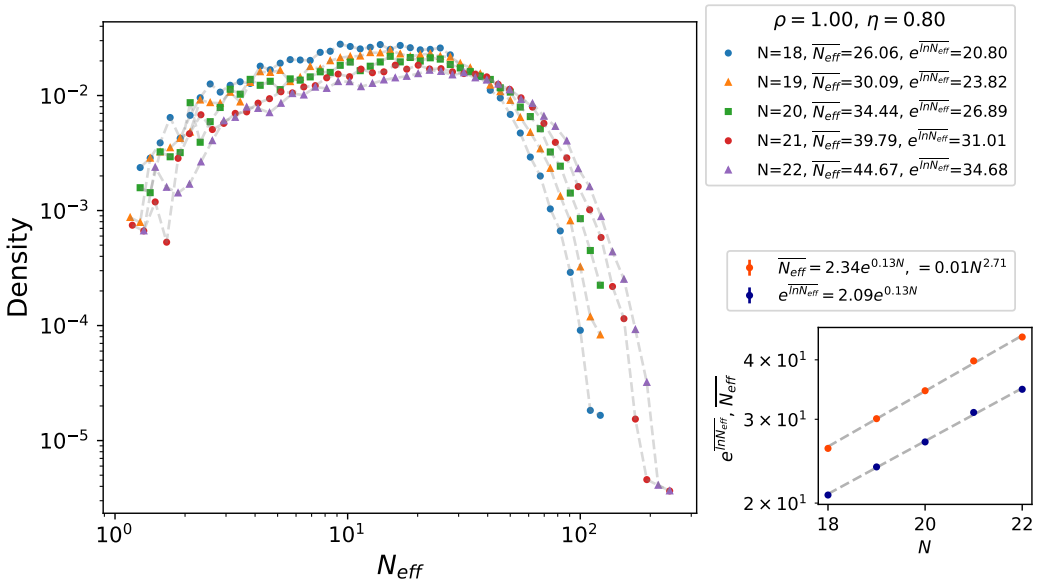


(a) $\rho = 0.51$: The overall behavior remains comparable to the exponential chaotic regime of the fully connected case ($\rho = 1$).

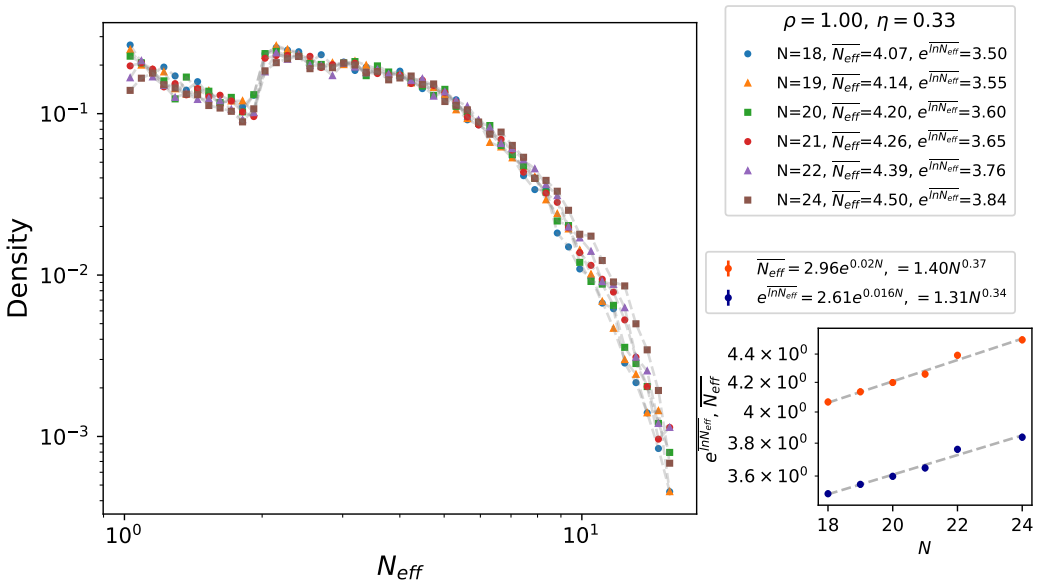


(b) $\rho = 0.23$: The distribution becomes peaked around $L_{\text{eff}} \sim 4$, with substantially reduced tails.

Figure 6.6 Effective cycle lengths in fully asymmetric diluted networks. Distribution of the weighted cycle lengths L_{eff} for $\eta = 0$ at different levels of density ρ. Insets show scaling of average and typical values with system size N. The shift from exponential to localized behavior with increasing dilution is clearly visible.



(a) $\eta = 0.80$: Oscillatory regime characterized by exponential growth in the mean number of attractors.



(b) $\eta = 0.33$: Chaotic regime dominated by few effective attractors.

Figure 6.7 Effective attractor distributions across correlation regimes. (**a**) Well-defined distribution of N_{eff} at high correlation ($\eta = 0.8$) with exponential scaling with system size (inset). (**b**) Crossover to a regime dominated by 1–2 effective attractors at lower correlation ($\eta = 0.33$), characteristic of the chaotic regime.

Our analysis reveals distinct scaling regimes for attractor statistics across the parameter space. In fully connected networks ($\rho = 1$) and moderately diluted systems ($\rho \gtrsim 0.2$) with high correlation ($\eta \to 1$), the mean number of attractors follows

exponential scaling $\overline{C} \simeq Ae^{\alpha_C N}$, with similar behavior observed for $\overline{N_{\text{eff}}}$. The complexity parameter α_C decreases monotonically with increasing asymmetry.However, when coupling asymmetry exceeds a critical value $\eta_d(\rho)$ or when dilution becomes substantial ($\rho \lesssim 0.2$), the scaling evolves into a power-law regime. In the dense, fully asymmetric limit ($\eta = 0$, $\rho = 1$), we observe linear dependence $\overline{C} \propto N$, as demonstrated in Fig. 6.1a. Effective cycle lengths $\overline{L_{\text{eff}}}$ exhibit significant size dependence for $\eta \lesssim 0.33$ coinciding with distinct behavior in the chaoticity parameter p_{chaos}.
The effective number of attractors reveals complementary structure (Fig. 6.7). For $\eta \lesssim 0.33$, the distribution of N_{eff} typically features 1–2 dominant effective attractors (panel 6.7b), while at higher correlations ($\eta = 0.8$, panel 6.7a), the distribution becomes well-defined with exponential scaling, approaching the characteristics of more symmetric scenarios.
In order to fully characterize the phase diagram to better understand the general behavior of the observables as a function of the due parameters, we show the colormaps even for the other quantities reported in Table 5.1. In particular, we represent the behavior of the transient time or mean distance from the attractors $\overline{\langle MD \rangle}$ in Fig. C.14 and in Fig. C.15-C.16 the effective quantities $\overline{N_{\text{eff}}}$ and $\overline{L_{\text{eff}}}$, where averages are weighted by basin sizes. The phase diagrams for effective quantities show scaling behaviors consistent with those observed for their unweighted counterparts in Fig. 6.4 and 6.5.
In contrast to the case of $\overline{\langle L \rangle}$, where the behavior exhibits a drastic change at $\eta \simeq 0.5$, the N-dependence of the timescale $\overline{\langle MD \rangle}$ evolves smoothly with η, as illustrated in Fig. C.2. In the symmetric limit, convergence to an attractor typically requires just 1–2 steps. For $\eta \lesssim 0.5$, the distribution reflects the behavior of the fully asymmetric case $\eta = 0$ shown in Fig. 6.2a.

6.2.1 High Dilution Regime $\rho \lesssim 0.2$

We focus on the high dilution part of the phase diagram that represents a particularly interesting case study since it shows biological plausibility represented by sparse connectivity limiting the chaotic behavior of the system. In this region, networks exhibit ordered behavior characterized by exponential attractor counts but short cycle lengths, $\overline{\langle L \rangle} = 4$, with reduced distribution tails that less dominate the average values. This behavior emerges independently of synaptic correlation η. The exception is the fully correlated case $\eta = 1$, where $\overline{\langle L \rangle} = 2$ due to the dominance 2-cycles characteristic of symmetric systems with parallel dynamics. As shown in Fig. C.19, this stability in the mean cycle length is largely independent of the total number of attractors $C(\alpha)$ in each sample. The exception occurs in the fully symmetric case ($\eta = 1$), where $\langle L \rangle$ converges exactly to 2, a characteristic feature that persists independently of system size N. In this regime where we have found an exponential growth in the number of attractors. Our goal is to identify scaling relations that distinguish this regime from the fully connected and symmetric one which, despite having numerous attractors (see C.7), is not biologically plausible. Regarding the distribution and the mean number of attractors, we have seen a drastic change in behavior for a high value of dilution $\rho \lesssim 0.2$ as visible from Fig. 6.4. We report in Fig. C.8 the distribution of the attractors and in the insets the fits for the average values $\overline{C}$ and the typical values $\mathrm{e}^{\overline{\ln C}}$ as a function of the system size N. For

$\rho \lesssim 0.2$ we observe an exponential scaling of the mean number of attractors, with the exponential coefficient α_C increasing as density ρ decreases. In particular, at $\rho = 0.08$ and $N = 24$, we find $\overline{C} = 1676.10$, demonstrating the substantial growth of attractors in extremely diluted regimes. We emphasize that this exponential growth behavior is η-independent, persisting even in the fully asymmetric case, the scenario most prone to chaotic dynamics. In Fig. C.8 we show the distribution in the asymmetric case $\eta = 0$ in order to show that, even in this worst correlated scenario, we still have an exponential growth of the mean number of attractors. We would like to underline that the shape of the distribution is much more different from that observed in relatively dense, highly correlated networks (compare with Fig. C.7), highlighting the different properties of the high dilution regime. Let $C(\alpha)$ denote the number of attractors in sample α (corresponding to a specific realization of the coupling matrix $\mathbf{J}$)[2], and let Ω_i represent the number of initial conditions that converge to the i-th attractor within sample α. For each sample α, we define the average log-basin size as:

$$\langle \log B \rangle = \frac{1}{C(\alpha)} \sum_{i=1}^{C(\alpha)} \log \Omega_i \tag{6.1}$$

with the normalization condition $\sum_{i=1}^{C(\alpha)} \Omega_i = 2^N$ ensuring conservation of state space volume. In the following Fig. 6.8 we report the behavior of $\langle \log B \rangle$ as a function of $\log C(\alpha)$.

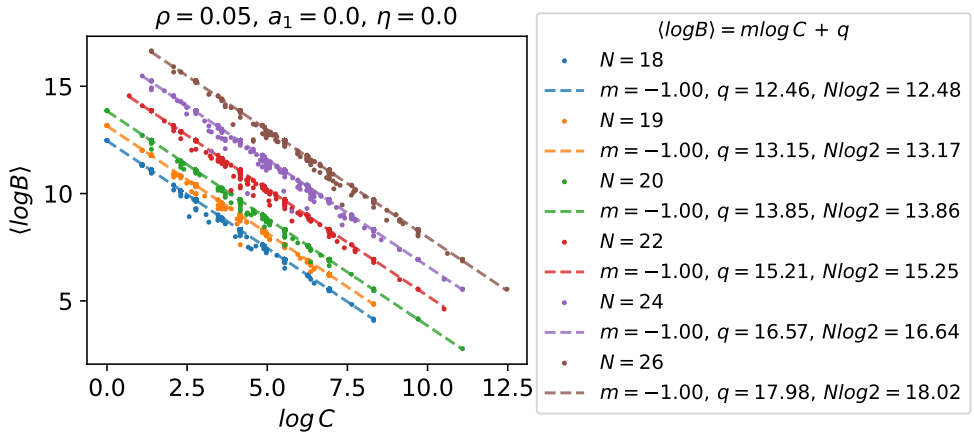
Our analysis reveals that in the high dilution regime with symmetric topology ($a_1 = 0$), the configuration space partitions into basins of approximately equal size across all correlation values η. This homogeneous division of state space persists throughout the parameter range, indicating a fundamental property of highly diluted networks.

The mathematical foundation for this observation follows directly from uniform basin sizing. If $\Omega_i = \Omega$ for all attractors i, then from the normalization condition $\sum_{i=1}^{C(\alpha)} \Omega_i = 2^N$, we obtain $\Omega = 2^N / C(\alpha)$. Consequently, the average log-basin size becomes:

$$\langle \log B \rangle = -\log C(\alpha) + N \log 2. \tag{6.2}$$

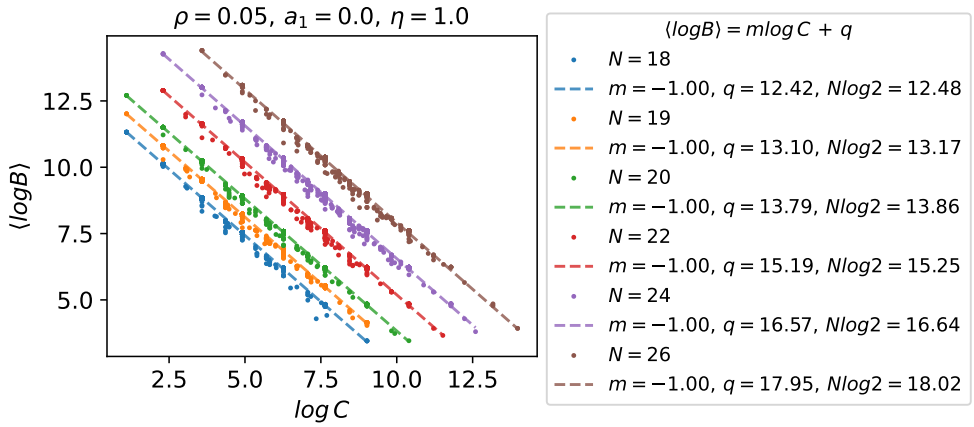
The linear relationship described by Eq. (6.2) remains robust across all correlation values η for $\rho \leq 0.1$, although increased fluctuations become apparent as shown in Fig. C.18.

[2]Note: In previous sections and figures, we have used the simplified notation $C(\alpha) = C$.



(a) Fully asymmetric case $\boldsymbol{\eta = 0}$

Figure 6.8 Basin size uniformity in highly diluted networks. Relationship between $\langle \log B \rangle$ and $\log C(\alpha)$ for networks with high dilution ($\rho = 0.05$) and symmetric topology ($a_1 = 0$).



(b) Fully symmetric case $\boldsymbol{\eta = 1}$

Figure 6.8 Basin size uniformity in highly diluted networks (continued). Both correlation regimes demonstrate uniform partitioning of state space, with attractor basins approaching equal sizes regardless of synaptic correlation η.

Generalized Entropy Measures

To quantify the distribution of basin sizes and characterize attractor dominance, we observe Rényi entropy measures. The Rényi entropy of order γ, where $0 < \gamma < \infty$ and $\gamma \neq 1$, is defined as:

$$S_\gamma = \frac{1}{1-\gamma} \log \left(\sum_{i=1}^{n} p_i^\gamma \right). \tag{6.3}$$

The Shannon entropy emerges as the limiting case $\gamma \to 1$:

$$S_1 = -\sum_{i=1}^{n} p_i \log p_i \tag{6.4}$$

while the case $\gamma = 2$ is known as collision entropy:

$$S_2 = -\log \sum_{i=1}^{n} p_i^2 = -\log P(\mathrm{X} = \mathrm{Y}) \tag{6.5}$$

where X and Y are independent and identically distributed random variables. In our framework, Ω_i represents the number of initial conditions converging to attractor i in sample α, and $W_i = \Omega_i / 2^N$ denotes the corresponding normalized weight (satisfying $\sum_{i=1}^{C(\alpha)} W_i = 1$). We can define the Shannon entropy in our situation as:

$$S_1 = -\sum_{i=1}^{C(\alpha)} W_i \log W_i = N\log 2 - \frac{1}{2^N} \sum_{i=1}^{C(\alpha)} \Omega_i \log \Omega_i \tag{6.6}$$

and the collision entropy as:

$$S_2 = -\log \sum_{i=1}^{C(\alpha)} W_i^2 = -\log Y = N\log 4 - \log \left(\sum_{i=1}^{C(\alpha)} \Omega_i^2 \right). \tag{6.7}$$

These entropy measures capture different aspects of basin structure: the collision entropy S_2 emphasizes dominant attractors with large basins, as squaring probabilities reduces the influence of smaller basins, while the Shannon entropy S_1 is more sensitive to small probabilities represented by the presence of attractors with few initial conditions. In the ideal case of uniform basin distribution ($\Omega_i = 2^N / C(\alpha)$ for all i), both entropy measures coincide:

$$S_1 = S_2 = \log C(\alpha). \tag{6.8}$$

We analyze the fluctuations of both entropy measures from this uniform baseline $\log C(\alpha)$, obtained in the uniform case where the initial conditions are uniformly distributed across attractors. In the high dilution regime we observe that the histogram of the fluctuations from the uniform situation is sharply peaked around 0 suggesting that the distribution of initial conditions among attractors is nearly uniform across the samples α. In particular for $\rho = 0.05$ it is independent both from the system size N and the correlation η as visible from Fig. C.20. We started to see a changing in the distribution in the correlation η: Fig. 6.9 presents the fluctuations of Shannon entropy S_1 (left panels) and collision entropy S_2 (right panels) from the uniform distribution baseline $\log C(\alpha)$ for $N = 26$, systematically varying both dilution ρ and correlation η showing how the convergence to uniform state space partitioning depends on both dilution and synaptic correlation. To complete the analysis on deviations from uniform basin distributions across the parameter space, we extract the mode values of $S_\gamma - \log C(\alpha)$ for $\gamma = 1, 2$ from the entropy fluctuation distributions. These modal values are used to construct the phase diagrams in Fig. 6.10 that give a complete visualization of how basin entropy structure evolves across the full parameter space (ρ, η).

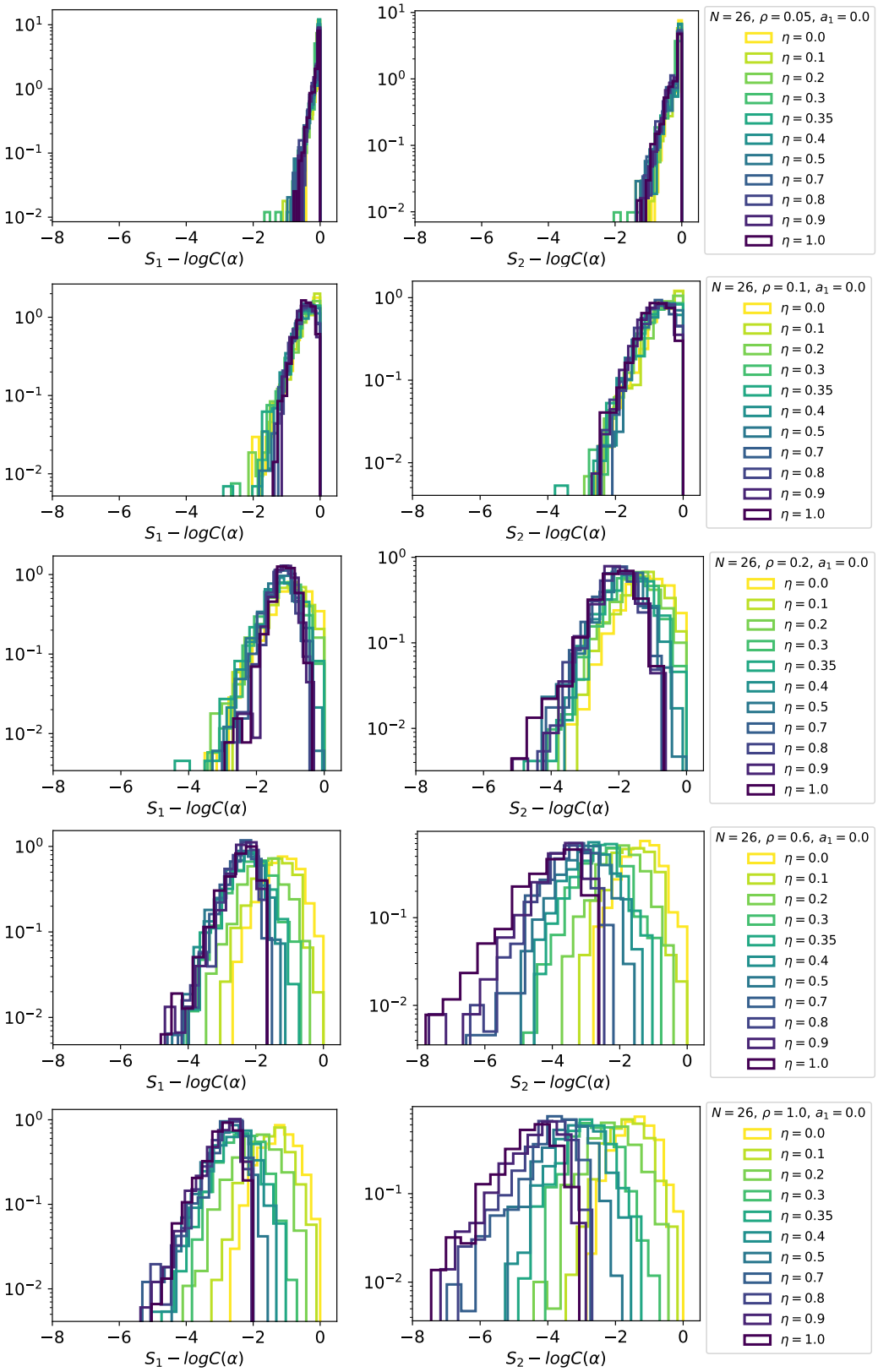


Figure 6.9 Evolution of entropy fluctuations with dilution and correlation ($N = 26$). Histograms of fluctuations from the uniform baseline $\log C(\alpha)$ for (**left**) Shannon entropy S_1 and (**right**) collision entropy S_2. Each row corresponds to a different dilution level ρ, illustrating the transition from left-skewed distributions at lower densities to more symmetric distributions at high density.

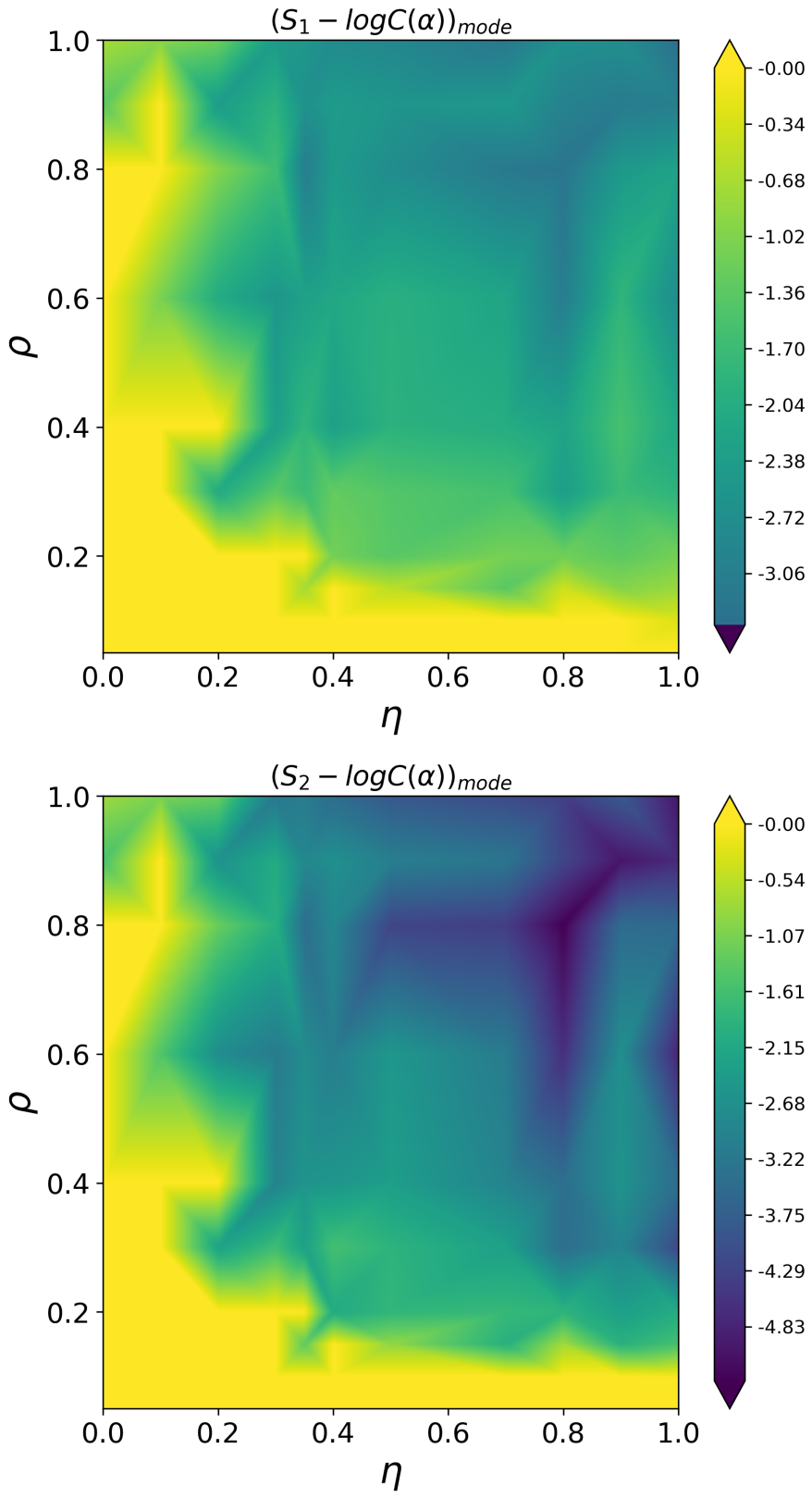


Figure 6.10 Phase diagrams of entropy deviations from uniform basin distribution. (**Top**) Mode values of $S_1 - \log C(\alpha)$ across the (ρ, η) parameter space. (**Bottom**) Mode values of $S_2 - \log C(\alpha)$. Both measures reveal systematic patterns in basin size heterogeneity, with the high dilution regime showing minimal deviations (yellow regions) indicating nearly uniform partitioning, while higher density correlated regions exhibit significant departures from uniformity.

6.3 Dependence on the Number of Direct Links a_1

In this last section we present the analysis conducted introducing also the number of direct links a_1 as a free parameter. The resulting phase diagram is a prism with a triangular base in the (ρ, a_1) plane, as explained in Fig. 5.2, explored at different levels of coupling correlation η. In order to create the parameters diagrams shown in Figs. 6.11 and 6.12, we have conducted the same analysis described in Sec 6.2. Due to inherent constraints, the lower-left vertex of the phase diagrams at $\rho = 0$ and $a_1 = 0$ remains experimentally inaccessible. The smallest density simulated was $\rho = 0.01$ with $a_1 = 0$[3].

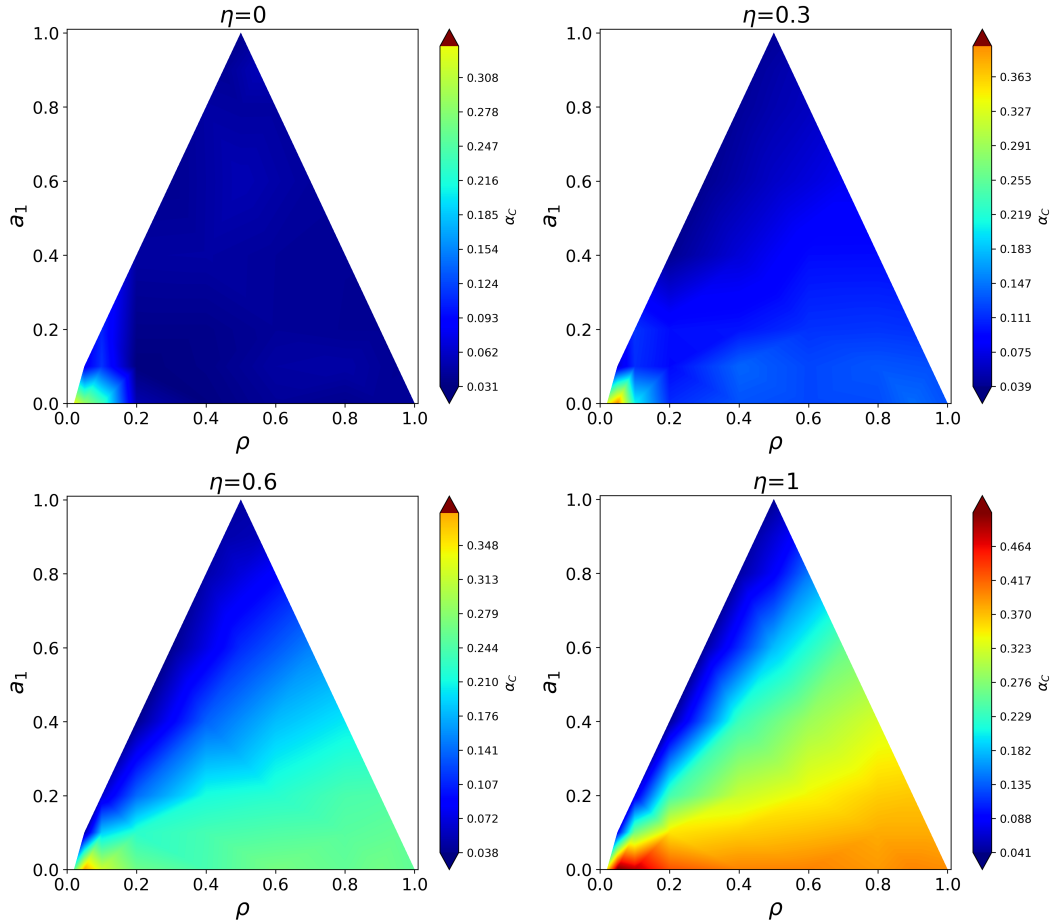


Figure 6.11 Attractor Proliferation Across the Full Parameter Space Exponential scaling coefficient α_C for the mean number of attractors $\overline{C} \sim e^{\alpha_C N}$ across the parameter space (ρ, a_1) for varying synaptic correlation η. The triangular geometry reflects the constraint $a_1 \leq a_1^{\max}(\rho)$ of Eq. (5.10). Each panel corresponds to a fixed η value, as labeled. Note the exponential growth (higher α_C, warmer colors) in the high-dilution regime ($\rho < 0.2$) across all correlation levels and at low probability of direct link a_1.

[3]Unlike the square colormaps in Figures 6.4–6.5, the triangular geometry of these phase diagrams precludes the use of interpolation to complete the vertex region.

In Fig. 6.11 we show colormaps for the mean number of attractors $\overline{C}$, while in Fig. 6.12 we report the corresponding maps for the mean cycle length $\overline{\langle L \rangle}$. From a biological perspective, the most relevant and plausible conditions correspond to $a_2 = 0$ (represented along the left edge of the triangular diagrams) and lower values of correlation η. In this regime, we observe the least favorable conditions for attractor proliferation and, consequently the dynamics is highly chaotic as evidenced by the data in Figs. 6.11 and 6.12. Again as in the case of symmetric topology, dilution seems to have a positive effect in increasing significantly the number of attractors, as indicated by the variation in color within the bottom-left regions of the diagrams. This shift corresponds indeed to higher values of the exponential coefficient α_C.

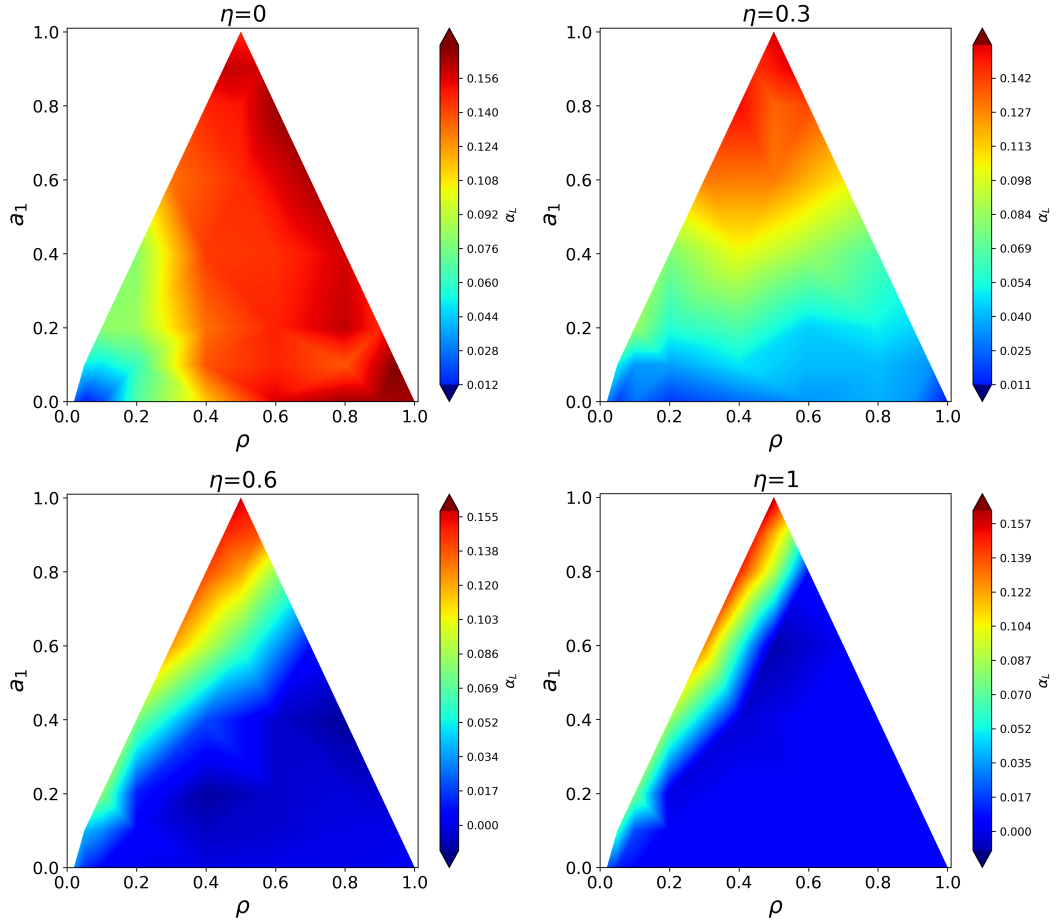


Figure 6.12 Cycle Length Scaling Across the Full Parameter Space
Exponential scaling coefficient α_L for the mean cycle length $\overline{\langle L \rangle} \sim e^{\alpha_L N}$ in the (ρ, a_1) plane for different correlation values η. Comparison with Fig. 6.11 reveals the dynamical dichotomy: regions exhibiting high attractor density ($\alpha_C > 0.10$, Fig. 6.11) typically show short cycles ($\alpha_L \approx 0$, cooler colors here). This complementarity underscores the distinct roles of topological sparsity and synaptic symmetry in shaping network dynamics. The variation with η illustrates how synaptic symmetry increases attractor proliferation and short cycles lengths even if not in a biological plausible way.

6.4 Take home message of Chapters 5 and 6

Overview: A Comprehensive Framework for Asymmetric Diluted Neural Networks In Chapters 5 and 6, we have presented a systematic exploration of how structural parameters shape the dynamical landscape of finite neural networks. Our generalized framework, parameterized by the triplet (ρ, η, a_1) gives the first complete characterization that explicitly distinguishes between:

- **Connection density** ρ: The fraction of existing synaptic connections
- **Synaptic correlation** η: The degree of symmetry in bidirectional couplings
- **Topological asymmetry** a_1: The probability of unidirectional connections

This distinction, often overlooked in previous literature, reveals that topological and synaptic asymmetries operate through fundamentally different mechanisms, each influences differently the network dynamics. In this section we collect the main messages and results discovered through the numerical analysis conducted in the asymmetric diluted neural network scenario. The comprehensive exploration of the (ρ, η, a_1) parameter space presented in Fig. 5.2, reveals fundamental principles that characterize neural network dynamics across different architectural regimes. Our analysis has shown that dilution ρ, synaptic correlation η, and topological asymmetry a_1 represent complementary control parameters that shape distinct systems with different main characteristics.

6.4.1 Synthesis and Interpretation of Results

In Sec. 6.2 we presented the results for the topological symmetric networks, so without direct links $a_1 = 0$, understanding the effects of dilution ρ and symmetry in the coupling, the correlation η, on the main observable reported in Table 5.1 revealing three fundamental dynamical regimes emerging from the interplay between these two parameters:

1. **Chaotic Regime** (High Density, Low Correlation):
    - **Location in the parameter space**: $\rho \gtrsim 0.2$, $\eta \lesssim 0.33$
    - **Key characteristics:**
        - **Exponential growth of mean cycle lengths** $\overline{\langle L \rangle} \sim e^{\alpha_L N}$ with $\alpha_L \sim 0.16$ in the dense asymmetric case ($\rho = 1, \eta = 0$)
        - **Power law scaling of attractor counts** $\overline{C} \sim N^{\beta_C}$ and a linear scaling in the dense asymmetric case ($\rho = 1, \eta = 0$): $\overline{C} \sim 0.35N$
        - **Heterogeneous basin structure**: Few dominant attractors with $\overline{N_{\text{eff}}} \sim 2.5$ when $\rho = 1$, $\eta = 0$
        - **Singular distribution** $\Pi(Y)$ with peaks $Y = 1/n$, indicating hierarchical basin organization
        - **Chaotic exploration of configuration space**: The mean transient time $\overline{\langle MD \rangle} \sim e^{\alpha_{MD} N}$ grows exponentially with N.
    - **Physical interpretation**: In this regime, the network shows random deterministic chaos—trajectories and explore extensive regions of configuration space through long, complex cycles.

2. **Oscillatory/Ordered Regime** (High Correlation):
   - **Location in the parameter space**: $\eta \gtrsim 0.5$, any ρ
   - **Key characteristics:**
     - **Short, stable cycles** $\overline{\langle L \rangle} \sim 2$ (predominantly period-2 oscillations and fixed points with no increase with system size N)
     - **Exponential attractor proliferation** $\overline{C} \sim e^{\alpha_C N}$ with α_C maximal at $(\rho = 1, \eta = 1)$
     - **Homogeneous basin distribution**: $\Pi(Y)$ sharply peaked near $Y \to 0$, indicating many attractors with comparable basin sizes
     - **Rapid convergence**: $\overline{\langle MD \rangle} \sim 1 - 2$ steps, independent of system size N.
   - **Physical interpretation**: High synaptic correlation induces a regular, predictable landscape dominated by simple periodic orbits. In particular:
     - **Dense Correlated Networks ($\rho = 1, \eta = 1$)**: Maximal synaptic correlation produces 2-cycles characteristic of symmetric systems. The exponential growth of attractor counts with $\alpha_C \simeq 2\Sigma_{SK}$ demonstrates maximal memory storage within dense architectures, more aligned with the traditional spin glass scenario or stable memory-like behavior typical for example of the Hopfield model.
3. **Intermediate Regime** (Correlation Threshold $\eta_d \approx 0.33$, Dilution Threshold $\rho \approx 0.2$):
   - This regime is demarcated by critical boundaries in the parameter space, at $0.33 \lesssim \eta \lesssim 0.5$ and the dilution threshold $\rho \approx 0.2$ and separates exponentially growing cycles (chaotic) from short, stable cycles (ordered). This region is governed by power-law scaling with mixed properties and:
     - **Persists under dilution** The vertical boundary at η_d extends through the entire density range $\rho \in (0.2, 1)$
     - **Is continuous, not abrupt** Properties evolve smoothly across η_d, with the most significant changes, especially for the cycle lengths, occurring in $0.33 \lesssim \eta \lesssim 0.5$
     - **Affects multiple observables simultaneously** basin structure ($\Pi(Y)$), and chaoticity (p_{chaos}) all exhibit qualitative changes at η_d.
4. **Highly Diluted Networks**:
   - **Location in the parameter space**: $\rho \lesssim 0.2$, any η
   - **Key characteristics:**
     - **Short, stable cycles** $\overline{\langle L \rangle} = 2 - 4$, size independent for large N
     - **Exponential attractor growth** $\overline{C} \sim e^{\alpha_C N}$ with α_C increasing as ρ decreases
     - **Uniform phase space partitioning**: Basins sizes satisfy $\Omega_i \approx 2^N/C$ across all attractors

  - **Rapid convergence**: $\overline{\langle MD \rangle} \sim 1 - 2$ steps, independent of system size N
  - $\boldsymbol{\eta}$**-independence**: All statistical properties become insensitive to synaptic correlation for $\rho \lesssim 0.2$.
- **Physical interpretation**: This regime represents an extension of the ordered one whose characteristics for density $\rho \lesssim 0.2$ becomes independent on the correlation η. Sparse connectivity fragments the configuration space into numerous, compact attractor basins of approximately equal size; it is also biologically plausible and can be used as a basis for understanding real neural structures.

The systematic evolution from exponential to power-law scaling in both attractor counts and cycle lengths delineates a fundamental trade-off in network design. High correlation $\eta \to 1$ produces numerous but simple (short-cycle) attractors, ideal for pattern storage, whereas low correlation $\eta \to 0$ allows complex, long-period dynamics suitable for temporal processing. The critical region $0.3 \lesssim \eta \lesssim 0.5$ represents a transition zone where networks balance these competing capabilities. This additional dynamical transition within the chaotic regime was first identified in the fully connected case for spin variables in [21] and we have shown that it is present also adding dilution in the coupling matrix $\mathbf{J}$. Although the average cycle length undergoes an abrupt change at $\eta_C \approx 0.5$, the stability region for 2-cycles extends down to $\eta_d \approx 0.33$. In this regime, almost all trajectories converge to 2-cycles, despite their basins of attraction having vanishingly small weight. For $\eta < \eta_d$, the probability of finding a 2-cycle drops abruptly to zero, and the typical attractors—now increasing only linearly with N—become exponentially long with system size. The phase maps in Figs. 6.4 and 6.5 reveal a complementary structure that defines these distinct dynamical regimes. In Fig. 6.4a, the exponential growth rate α_C reaches maximum values in two distinct regions: (1) highly correlated systems $\eta \to 1$ with increasing density, and (2) strongly diluted networks $\rho \ll 1$ across the entire correlation spectrum. This second case is particularly significant from a biological perspective, suggesting that sparse connectivity can promote the proliferation of attractors independently of the synaptic symmetry. Panel 6.4b shows the complementary structure: most of the parameter space is dominated by high values of β_C, indicating that power-law scaling is not representative. The exception is a narrow region at low η and high ρ, where β_C decreases drastically, precisely where exponential scaling fails in panel 6.4a. Fig. 6.5 confirms this dichotomy from the perspective of cycle lengths. The exponential scaling of $\overline{\langle L \rangle}$ (6.5a) emerges only in the same chaotic region identified for $\overline{C}$, while the rest of the parameter space shows power-law behavior (6.5b). The sharp vertical boundary at $\eta_d \approx 0.33$ persists even under substantial dilution, extending the known dynamical transition from the fully-connected case to more general topologies. The relationship between attractor proliferation and cycle complexity is non-monotonic. As highlighted by comparing Figs. 6.4 and 6.5, the chaotic regime (low η, high ρ) shows simultaneous linear growth of $\overline{C}$ with N and exponentially long cycles. In contrast, the high dilution regime $\rho \lesssim 0.2$ produces exponential growth of $\overline{C}$ but with short, stable cycles ($\overline{\langle L \rangle} \sim 4$, independent of N for sufficiently large N). This dichotomy suggests distinct underlying mechanisms: in the first case, an expansion of the phase space

through long cycles; in the second, a fragmentation into compact attractors.

High Diluted Regime

A key result of the analysis in this region, is the non-monotonic relationship between dilution and storage capacity in asymmetric networks. The average number of attractors $\overline{C}$ shows a pronounced peak at very high levels of dilution. This indicates the existence of an optimal, sparse connectivity regime that maximizes memory storage. Furthermore, the intensity of this peak, corresponding to the maximum achievable storage capacity, scales exponentially with the network size, where the exponential scaling exponent α_C is found to be comparable to that of fully-connected, symmetric networks. This is a remarkable finding: it demonstrates that a highly diluted, asymmetric topology can achieve a storage capacity on par with the optimal, but biologically unrealistic, case of a fully-connected symmetric network. A second characteristic of this region is the uniform partitioning, whose evidence comes from the linear relationship (Eq. 6.2) $\langle \log B \rangle = -\log C(\alpha) + N \log 2$ and is shown in Fig. 6.8-C.18 where is visible that holds robustly across all η values, confirming that each attractor captures an equal fraction $\sim 1/C$ of the state space. To better quantify the degree of heterogeneity in the attractor landscape, we analyzed the differences $S_\gamma - \log C(\alpha)$ for $\gamma = 1, 2$, where S_1 and S_2 denote respectively the Shannon and collision entropies and where by construction, a value close to zero corresponds to a uniform partitioning of the state space. In Fig. 6.9 we show the transition from uniform to heterogeneous partitioning of the phase space as dilution and synaptic correlation vary. For $\rho = 0.05$ (top row), the distributions of S_1 and S_2 are concentrated around zero for all values of η, confirming that initial conditions distribute uniformly among attractors. This interesting property, the uniform partitioning independent of synaptic correlation, characterizes the high dilution regime. Increasing the density to $\rho = 0.1$, more pronounced fluctuations emerge, especially at low correlation η. For $\rho \gtrsim 0.2$, the distributions develop left-skewed tails, indicating the emergence of dominant basins. This asymmetry is more marked for the collision entropy S_2 (right column), which emphasizes attractors with large basins by squaring the probabilities, compared to the Shannon entropy S_1, which is more sensitive to small basins. The dependence on correlation becomes evident at high densities: the light curves (low η) show greater deviations from uniformity compared to the dark curves (high η), suggesting that synaptic symmetry promotes a more homogeneous attractor landscape. The phase diagrams in Fig. 6.10 give a complete visualization of the entropic structure of the basins across the (ρ, η) parameter space. The modal values of $S_1 - \log C(\alpha)$ (top panel) and $S_2 - \log C(\alpha)$ (bottom panel) quantify the deviations from the uniform baseline $\log C(\alpha)$. The extensive yellow region for $\rho \lesssim 0.2$ confirms the nearly uniform partitioning in the high dilution regime, consistent with the analysis in Figs. 6.8 and C.18. This minimal entropic deviation independent of η suggests that topological sparsity alone is sufficient to promote homogeneity in the dynamical landscape, independent of synaptic properties and an absence of bias towards specific attractors. The transition towards significant deviations occurs gradually for $\rho \gtrsim 0.2$, with steeper gradients in the case of collision entropy S_2 that more strongly emphasizes the presence of dominant attractors through the quadratic weighting of probabilities, explaining

the sharper gradients compared to S_1. The dependence on synaptic correlation η becomes pronounced only for $\rho \gtrsim 0.5$, where high symmetry $\eta \to 1$ tends to produce more homogeneous landscapes. This confirms that in the densely connected regime, both topology and synaptic properties contribute to the structure of the attractor landscape.

Biological Motivation for Diluted Connectivity

The assumption of full connectivity in classical models is biologically untenable. Empirical evidence indicates that neural connectivity is inherently sparse, or diluted. For instance, the connection probability between any two neurons in the hippocampal CA3 region is approximately 4% [137], while for nearby neocortical pyramidal cells, it is on the order of 10% [122]. This prevalence of sparse wiring is not merely a structural constraint but may confer functional advantages. As proposed by [123], increased connectivity can paradoxically reduce memory capacity. Their numerical studies suggest that higher dilution promotes network stability and retrieval accuracy by mitigating spiking-related noise and the stochastic fluctuations it induces. The optimal dilution level identified in our simulations of asymmetric networks aligns remarkably well with the connection densities observed anatomically in the neocortex and hippocampus [146]. This correspondence suggests that the brain's connectivity may be tuned to a sparse regime that optimally balances the trade-off between high storage capacity and stable, rapid retrieval dynamics. Indeed, our results suggest that such architectures naturally produce:

- Rich exponential attractor count
- Rapid, stable convergence (short cycles, low transients)
- Unbiased state space coverage (uniform basin distribution)
- Robustness to synaptic variability (η-independence)

6.4.2 Extension to General Topologies: The Role of Directed Links

In Sec. 6.3 we completed the exploration of the full three-dimensional parameter space (Fig. 5.2) by introducing topological asymmetry through directed links $a_1 \neq 0$. The triangular phase diagrams reported in Figs. 6.11-6.12 reveal what in general the presence of direct links, as expected, suppress attractor counts especially in the low correlated cases. However, again in the high diluted regime $\rho \lesssim 0.2$, and intermediate values of as_1, there is an increasing in the mean growth of the mean number of attractors that is η-independent as in the previous topological symmetric case $a_1 = 0$.

6.4.3 Limitations and Future Exploration

Our results want to represent a complete framework for understanding how structural constraints shape dynamical systems. Future work should address open emerging questions:

- What are the precise local mechanisms through which dilution generates this exponential growth in the number of attractors and how is it different from the dense correlated case?
- How do these architectural principles scale to larger systems approaching biological scales?
- Can these findings influence the design of artificial neural systems biologically inspired?
- How do learning and plasticity rules interact with these innate architectural constraints?

While correlation creates one pathway to regulate dynamics, dilution represents an alternative route to attractors richness within biologically plausible architectures characterized by the emergence of exponential attractor counts with stable, short cycles and no biased dominance of specific configurations. The η-independence of key properties in high dilution demonstrates that this kind of "biological networks" are robust to variation in the symmetry of the coupling links. We have characterized the statistical properties of high-dilution regimes, but for now, the precise mechanisms by which a diluted ensable of synaptic matrix enables both exponential attractor counts and rapid convergence remain to be fully elucidated.
Our simulations are limited to $N \leq 26$ due to the $O(2^N)$ computational complexity of exhaustive enumeration of the configuration space, a constraint that prevents direct observation of thermodynamic limit but that should be contextualized within methodological choices for this specific dynamics study. Previous studies have used stochastic sampling approaches to access larger system sizes for analyzing average cycle lengths $\langle L \rangle$, which can be estimated from trajectory sampling. Instead, a complete analysis of the number and distribution of attractors C fundamentally requires complete knowledge of the configuration space, as some disjoint interesting component may be missed by stochastic methods. The trade-off between system size and completeness of attractor enumeration thus represents a fundamental future challenge.

Conclusions

The work presented in this thesis explored the mechanisms that allow neural systems to store, retrieve, and reorganize information efficiently, combining tools from statistical mechanics and dynamical systems. Starting from the classical Hopfield model, we have progressively extended theoretical framework of associative memory in two complementary directions: the development of a new learning rule able to overcome the limitations of Hebbian prescription, and the investigation of structural heterogeneity in non-Hamiltonian networks that more closely resemble biological architectures. Understanding how the brain operates, its collective functioning, emergent properties, and cognitive processes, remains one of the most fascinating challenges in modern science. The main goal of this thesis has been to investigate why neural networks are so effective in solving optimization problems and processing high-dimensional data, and how insights from biological systems can guide the design of more efficient and adaptive learning architectures. In this work we move beyond the classical Hebbian paradigm to explore two fundamental pathways for improving associative memory in neural networks: one through a new learning algorithm, and the other through a systematic investigation of biologically realistic network structures. The work bridges two complementary perspectives on associative memory:

- In the **first part** we focus on an *algorithmic innovation* within the framework of symmetric, fully connected networks, introducing and analyzing the *Daydreaming algorithm*
- In the **second part** we explore how asymmetry and dilution in synaptic couplings reshape the attractor landscape of non-Hamiltonian neural systems.

Our findings demonstrate that the limitations of the standard Hopfield model can be overcome by either refining the learning rule within its framework or by embracing the dynamical richness of asymmetric, non-Hamiltonian systems.

Part I: The Daydreaming Algorithm—Algorithmic Enhancement of Symmetric Networks

In the first part we address the critical storage limit of the Hopfield model ($\alpha_c \approx 0.138$) by introducing the *Daydreaming algorithm*. This iterative, local learning rule continuously reshapes the synaptic coupling matrix $\mathbf{J}$ through an iterative process

combining Hebbian reinforcement and anti-Hebbian unlearning. This dual mechanism simultaneously stabilizes stored patterns, expanding their basins of attraction, while destabilizing spurious attractors sampled by the network's own dynamics. This learning rule demonstrates that local, unsupervised plasticity can dynamically organize synaptic weights to favor genuine memories over spurious states, extending the capacity of symmetric networks.
Crucially, the algorithm achieves these effects without fine-tuning, relying only on a single timescale parameter τ, whose specific value has minimal impact on performance for sufficiently large values. This robustness makes Daydreaming effectively parameter-free and fully local in its updates. In particular this algorithm:

- **Achieves theoretical maximum capacity:** For uncorrelated patterns, it reaches theoretical limit $\alpha_c = 1$, nearly an order of magnitude improvement over the standard Hebbian limit ($\alpha_c \approx 0.138$), with all stored patterns becoming stable fixed points, though with vanishing basins of attraction.

- **Operates robustly without catastrophic forgetting:** Unlike previous unlearning procedures that require careful early stopping, Daydreaming can iterate indefinitely without performance degradation, eliminating the delicate fine-tuning requirements that plague most learning algorithms.

- **Autonomously exploits data structure:** For correlated data generated by the random-features model, it not only stabilizes the stored patterns but spontaneously discovers and stabilizes the underlying hidden features without any supervision. Correlated patterns develop significantly larger basins of attraction than uncorrelated ones, enabling stable retrieval even at loads $\alpha > 1$, thereby surpassing theoretical bound that constrains random patterns.

- **Bridges associative memory with unsupervised learning:** On the MNIST dataset, it exhibits a striking phase transition as a function of load. At low loads ($\alpha < 0.8$), it functions as a high-capacity episodic memory, storing individual training images as stable attractors. At high loads ($\alpha \gtrsim 50$), individual examples lose stability, but the network develops emergent attractors corresponding to class prototypes. Dynamics initialized from either training or test images converge to these generalized representations, demonstrating unsupervised abstraction and achieving approximately 67.5% test classification accuracy—surpassing previous Hopfield-based approaches despite using a minimalist architecture.

These results highlight the algorithm's ability to uncover latent structure without supervision, effectively bridging the gap between associative memory and feature learning. In summary, the Daydreaming algorithm unlocks the latent potential of the Hopfield architecture, transforming it into a powerful, self-organizing system for content-addressable memory and feature extraction.

Part II: Structural Analysis of Asymmetric and Diluted Networks

The second part departed from both the Hebbian prescription and the assumption of symmetric connectivity to investigate how network topology fundamentally governs dynamics in biologically realistic architectures. Biological networks exhibit sparse connectivity (typical connection probabilities ∼4–10% in hippocampus and cortex) and predominantly asymmetric, often unidirectional synaptic pathways. These features fundamentally alter network dynamics: the energy function ceases to exist, detailed balance is violated, and the system becomes non-Hamiltonian, admitting limit cycles in addition to fixed points.
We developed a generalized framework to construct synaptic matrices $\mathbf{J}$ controlled by three independent parameters: connection density ρ, synaptic correlation η, and topological asymmetry a_1. This three-dimensional parameter space (ρ, η, a_1) covers all possible network architectures, from the classical symmetric Hopfield model to fully asymmetric sparse networks resembling biological connectivity.
Using exhaustive enumeration of all 2^N configurations for system sizes up to $N \simeq 30$, we systematically studied how these structural parameters affect the number, length, and degeneracy of attractors, as well as measures of configurational entropy and chaoticity. Through this comprehensive mapping, we revealed:

- **Richness in dynamical phases:** The analysis revealed distinct regimes ranging from *ordered phases* dominated by short cycles ($L \sim 2$) and a consistent number of stable attractors to *chaotic regimes* characterized by few exponentially long cycles and extended transient times. A dynamical transition occurs at a critical correlation $\eta_d \approx 0.33$, separating these regimes, and this transition persists under substantial dilution.

- **Distinct roles of synaptic and topological asymmetry:** By independently varying η and a_1, we demonstrated that directed links (topological asymmetry) and coupling decorrelation (synaptic asymmetry) have fundamentally different effects on attractor structure. Increasing synaptic symmetry ($\eta \to 1$) generally suppresses chaos and promotes ordered dynamics, while topological asymmetry ($a_1 > 0$) and dilution interact in complex, non-monotonic ways.

- **Discovery of the optimal high-dilution regime $\rho \lesssim 0.2$:** An optimal sparsity regime exhibiting new properties:

  - **Exponential attractor proliferation:** $C \sim e^{\alpha_C N}$ with growth rates comparable to or exceeding those of dense symmetric networks, achieved through structural sparsity alone
  - **Short, stable cycles:** Mean cycle lengths remain $\mathcal{O}(1)$ (typically $L \sim 2$–4) independent of system size, ensuring rapid, predictable convergence
  - **Uniform state-space partitioning:** Basin sizes satisfy $\Omega_i \approx 2^N/C$ across all attractors, quantified through Rényi entropy analysis. This "democratic" organization, where no specific memories are privileged, indicates unbiased accessibility across the configuration space.

  - **Robustness to synaptic variability:** All properties are independent of correlation η
  - **Biological plausibility:** The optimal dilution level $\rho \sim 0.05$–0.1 matches connectivity densities in hippocampal CA3 ($\sim$4%) and neocortex ($\sim$10%)

This analysis shows that dilution represents an alternative pathway to high capacity that does not require an energy function or synaptic optimization. In highly diluted networks, the phase space naturally fragments into numerous compact basins through topological constraints alone, achieving attractor proliferation comparable to optimally trained dense networks. This provides a deep structural insight offering an explanation for the specific connectivity patterns found in neural circuits.
Extending the analysis to include topological asymmetry $a_1 \neq 0$ revealed that while directed links generally reduce attractor counts (especially at low η), the high dilution regime remains more or less robust even with moderate values of a_1. This systematic exploration provides a crucial bridge between abstract neural models and biological reality, arguing that structural sparsity, rather than symmetry, is the dominant enabler of rich attractor landscapes in biologically plausible architectures. The sparse, asymmetric connectivity observed in biological brains is not a constraint to be overcome but a finely-tuned design principle that suppresses pathological transients, enables rapid computation, and maximizes memory capacity through topological fragmentation of phase space.

Synthesis: Complementary Mechanisms for Efficient Memory

The two proposed points of view are complementary:

- **Daydreaming optimizes within a Hebbian framework**, demonstrating that considerable capacity improvements are achievable through learning alone
- **Structural analysis on general synaptic matrix** reveals the presence of a rich attractor landscapes even with random couplings, suggesting that couplings structure is a fundamental component.
- **Combined effects remain to be explored**, though preliminary analysis suggests that applying adaptive algorithms to sparse asymmetric networks could potentially achieve both the capacity benefits of structural sparsity and the retrieval robustness of adaptive plasticity

This synthesis points toward a unified framework: rather than viewing optimization and architecture as separate concerns, they should be understood as interacting degrees of freedom that jointly determine networks capabilities.

Outlook and Future Directions

Our findings suggest specific hypotheses and design principles in different directions that need further investigation:

- **For theoretical neuroscience:** If biological networks have evolved to maximize memory capacity while minimizing metabolic costs, they should operate near the high dilution regime $\rho \sim 0.1$. Empirical measurements in hippocampal and cortical circuits are consistent with this prediction. The uniform basin partitioning implies that all memories are equally accessible, a property that could underlie flexible cognition and transfer learning.

- **Analytical theory of the high dilution regime:** Develop mean-field or cavity equations that predict the exponential scaling of attractor counts and the uniformity of basin sizes in sparse networks, in order to have a more complete analytical understanding enabling an optimization of network parameters.

- **Learning in asymmetric networks:** Adapt Daydreaming for networks without energy functions. This would test whether structural and learning-based algorithm can combine in order to obtain systems the benefits of both approaches.

- **Scaling to larger systems:** Extending the structural analysis of systems with $N \gg 30$, bridging to realistic cortical scales and allowing direct observation of asymptotic scaling laws.

- **Biological validation:** Comparing the model's predictions with experimentally measured connectivity matrices (through connectomics) and attractor dynamics.

This thesis has demonstrated that the storage capacity limitation of the classical Hopfield model is a consequence of specific assumptions about its synaptic formulation and network architecture. By relaxing these assumptions through adaptive plasticity (Daydreaming) and structural optimization (sparse asymmetric networks), we have shown that both near-optimal capacity and biological plausibility are simultaneously achievable. The Daydreaming algorithm achieves theoretical maximum capacity for uncorrelated patterns, exploits structure in correlated data to exceed this limit, and spontaneously extracts latent features—all through simple, local, biologically plausible updates. The high dilution regime supports exponential attractor proliferation, uniform basin partitioning, and efficient convergence independently of synaptic correlation, matching biological connectivity statistics.
By uniting these perspectives, this work contributes to a deeper theoretical understanding of memory formation and retrieval, and gives the basis for new bio-inspired computational paradigms that integrate learning dynamics and structural realistic design.
We are at the end of this journey. I hope that you enjoyed the story or at least the music!

Appendices

Appendix A

Appendix for Chapter 1

A.1 Calculation of the Free Energy in the Low Load Regime

In this Appendix we present in details the calculation of the constrained free energy $f_N(\boldsymbol{m})$ reported in Eq. (1.39). We begin with the expression for the partition function, rewritten using the integral representation of the δ-function:

$$\delta(m_\mu - m_\mu(\boldsymbol{\sigma})) = \int \frac{N dx_\mu}{2\pi}\, e^{iNx_\mu(m_\mu - m_\mu(\boldsymbol{\sigma}))}. \tag{A.1}$$

Inserting this into the definition of $Z(\boldsymbol{m})$ from Eq. (1.35) we obtain:

$$Z = \int \left(\prod_{\mu=1}^{P} dm_\mu\right) Z(\boldsymbol{m}) = \int \left(\prod_{\mu=1}^{P} dm_\mu\right) \sum_{\boldsymbol{\sigma}} \left(\prod_{\mu=1}^{P} \delta(m_\mu - m_\mu(\boldsymbol{\sigma}))\right) e^{\frac{\beta N}{2}\sum_\mu m_\mu^2}. \tag{A.2}$$

$$= \left(\frac{N}{2\pi}\right)^P \int \left(\prod_{\mu=1}^{P} dm_\mu dx_\mu\right) e^{iN\sum_\mu x_\mu m_\mu + \frac{\beta N}{2}\sum_\mu m_\mu^2} \sum_{\boldsymbol{\sigma}} e^{-i\sum_\mu x_\mu \sum_i \sigma_i \xi_i^\mu}. \tag{A.3}$$

The sum over spins $\boldsymbol{\sigma}$ now factorizes:

$$\sum_{\boldsymbol{\sigma}} e^{-i\sum_i \sigma_i \sum_\mu x_\mu \xi_i^\mu} = \prod_{i=1}^{N} \sum_{\sigma_i = \pm 1} e^{-i\sigma_i(\boldsymbol{x}\cdot\boldsymbol{\xi}_i)} = \prod_{i=1}^{N} 2\cos\left(\boldsymbol{x}\cdot\boldsymbol{\xi}_i\right). \tag{A.4}$$

Thus, the partition function becomes:

$$Z = \left(\frac{N}{2\pi}\right)^P \int \left(\prod_{\mu=1}^{P} dm_\mu dx_\mu\right) \exp\left(N\left[i\boldsymbol{x}\cdot\boldsymbol{m} + \frac{\beta}{2}\boldsymbol{m}^2 + \frac{1}{N}\sum_{i=1}^{N} \ln\left(2\cos\left(\boldsymbol{x}\cdot\boldsymbol{\xi}_i\right)\right)\right]\right). \tag{A.5}$$

In the exponent of the Eq. (A.5) there are both the energetic

$$e_N(\boldsymbol{m}) = -\frac{1}{2}\sum_{\mu=1}^{P} m_\mu^2(\boldsymbol{\sigma}) = -\frac{1}{2}\boldsymbol{m}^2 \tag{A.6}$$

and entropic contributions

$$s_N(\boldsymbol{m}) = i\sum_{\mu=1}^{P} x_\mu m_\mu + \frac{1}{N}\sum_{i=1}^{N} \ln\left(2\cos\left(\sum_{\mu=1}^{P} x_\mu \xi_i^\mu\right)\right) = i\boldsymbol{x}\cdot\boldsymbol{m} + \mathbb{E}[\ln 2\cos(\boldsymbol{\xi}\cdot\boldsymbol{x})] \tag{A.7}$$

with:

$$\mathbb{E}[O(\boldsymbol{\xi})] = \frac{1}{N}\sum_{i=1}^{N} O(\boldsymbol{\xi}_i). \tag{A.8}$$

In thermodynamic limit $N \to \infty$, the integral is dominated by the saddle point of the exponent. We extremize with respect to both $\boldsymbol{m}$ and $\boldsymbol{x}$.

$$\frac{\partial}{\partial m_\mu}: \quad ix_\mu + \beta m_\mu = 0 \quad \Rightarrow \quad x_\mu = i\beta m_\mu, \tag{A.9}$$

$$\frac{\partial}{\partial x_\mu}: \quad im_\mu + \frac{1}{N}\sum_i \xi_i^\mu \tan\left(\boldsymbol{x}\cdot\boldsymbol{\xi}_i\right) = 0. \tag{A.10}$$

Substituting the result from Eq. (A.9), $\boldsymbol{x} = i\beta\boldsymbol{m}$, into the exponent of Eq. (A.5) makes the argument real. Noting that $\cos(iy) = \cosh(y)$, the exponent becomes:

$$N\left[-\beta\boldsymbol{m}^2 + \frac{\beta}{2}\boldsymbol{m}^2 + \frac{1}{N}\sum_{i=1}^{N}\ln\left(2\cosh\left(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi}_i\right)\right)\right] = \tag{A.11}$$

$$= N\beta\left[-\frac{1}{2}\boldsymbol{m}^2 + \frac{1}{\beta N}\sum_i \ln\left(2\cosh\left(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi}_i\right)\right)\right]. \tag{A.12}$$

The expression in brackets is identified as $-\beta f_N(\boldsymbol{m})$. Therefore, the intensive constrained free energy is:

$$f_N(\boldsymbol{m}) = \frac{1}{2}\boldsymbol{m}^2 - \frac{1}{\beta N}\sum_{i=1}^{N}\ln\left(2\cosh\left(\beta\,\boldsymbol{m}\cdot\boldsymbol{\xi}_i\right)\right). \tag{A.13}$$

The extremization condition $\partial f/\partial\boldsymbol{m} = 0$ of the free energy (A.13) gives the equation of state:

$$\boldsymbol{m} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\xi}_i\tanh(\beta\boldsymbol{m}\cdot\boldsymbol{\xi}_i). \tag{A.14}$$

A.1.1 Properties of the Solution and Self-Averaging

The sum $\frac{1}{N}\sum_{i=1}^{N}$ in the microscopic expression for the magnetization is a random variable dependent on the specific realization of the patterns $\{\xi_i^\mu\}$. However, in thermodynamic limit ($N \to \infty$), the system exhibits the crucial property of *self-averaging*. This means that the intensive free energy $f(\boldsymbol{m})$ and the order parameters $\boldsymbol{m}$ converge to their expectation values over the disorder, effectively becoming non-random. The average $\mathbb{E}[\cdot]$ in Eqs. (1.39) and (1.40) can therefore be interpreted as an average over the pattern distribution [11]:

$$\mathbb{E}\left[O(\boldsymbol{\xi})\right] = \frac{1}{2^P}\sum_{\xi^1=\pm1}\cdots\sum_{\xi^P=\pm1} O(\boldsymbol{\xi}) = \sum_{\boldsymbol{\xi}\in\{-1,+1\}^P} P(\boldsymbol{\xi})O(\boldsymbol{\xi}), \tag{A.15}$$

where $P(\boldsymbol{\xi}) = 1/2^P$ for patterns drawn from a uniform distribution. The equation of state (A.14) becomes:

$$\boldsymbol{m} = \mathbb{E}[\tanh(\beta \boldsymbol{m} \cdot \boldsymbol{\xi})] \tag{A.16}$$

that is final expression reported in Eq. (1.40). In the same way, for the free energy in thermodynamic limit $f(\boldsymbol{m})$ we can invoke self-averaging and the sum over sites i is replaced by an average over the disorder, giving the final result in Eq. (1.39):

$$f(\boldsymbol{m}) = \frac{1}{2}\boldsymbol{m}^2 - \frac{1}{\beta}\,\mathbb{E}[\ln(2\cosh(\beta \boldsymbol{\xi} \cdot \boldsymbol{m}))]. \tag{A.17}$$

For patterns with independent and identically distributed (i.i.d.) components $\xi^\mu = \pm 1$, the following moments hold:

$$\begin{aligned}
\mathbb{E}[\xi^\mu] &= 0, \\
\mathbb{E}[\xi^\mu \xi^\nu] &= \delta_{\mu\nu}, \\
\mathbb{E}[\xi^\mu \xi^\nu \xi^\rho \xi^\lambda] &= \delta_{\mu\nu}\delta_{\rho\lambda} + \delta_{\mu\rho}\delta_{\nu\lambda} + \delta_{\mu\lambda}\delta_{\nu\rho} - 2\delta_{\mu\nu}\delta_{\nu\rho}\delta_{\rho\lambda}.
\end{aligned}$$

A simple solution to the mean-field equations (1.40) is the *retrieval* state, where the network configuration is aligned with a single pattern, say $\boldsymbol{\xi}^1$. This corresponds to an ansatz:

$$m_1 \neq 0, \quad \text{and} \quad m_\mu = 0 \quad \text{for all} \quad \mu > 1. \tag{A.18}$$

Substituting this ansatz into Eq. (1.40) and using the properties of the pattern distribution, the equation for m_1 decouples and reduces exactly to the Curie-Weiss self-consistency equation:

$$m = \tanh(\beta m), \tag{A.19}$$

where we have set $m \equiv m_1$. This indicates a second-order phase transition at the critical temperature $T_c = 1$ ($\beta_c = 1$). For $T < 1$, the system can retrieve stored patterns, with the magnetization m serving as the order parameter for the transition.

A.2 Calculation of the Hessian and its Eigenvalues

To determine the stability of the symmetric mixture states, we analyze the Hessian matrix of the free energy $f(\boldsymbol{m})$, defined in Eq. (1.39). The elements of the Hessian are given by the second derivatives:

$$A_{\mu\nu} = \frac{\partial^2 f}{\partial m_\mu \partial m_\nu}. \tag{A.1}$$

A solution $\boldsymbol{m}$ is a local minimum of the free energy (and thus a stable state) if the Hessian matrix evaluated at that point is positive definite, meaning all its eigenvalues are positive.
For the Hopfield model free energy, the Hessian matrix elements are:

$$A_{\mu\nu} = \delta_{\mu\nu} - \beta\left(\delta_{\mu\nu} - Q_{\mu\nu}\right), \tag{A.2}$$

where the matrix $\boldsymbol{Q}$ is defined as:

$$Q_{\mu\nu} = \mathbb{E}\left[\xi^{\mu}\xi^{\nu}\tanh^{2}(\beta, \boldsymbol{\xi}\cdot\boldsymbol{m})\right]. \tag{A.3}$$

We now evaluate this Hessian for a symmetric n-mixture state of the form reported in Eq. (1.53). Due to the symmetry among the first n patterns and among the remaining $P-n$ patterns, the matrix $\boldsymbol{Q}$, and consequently the Hessian $\boldsymbol{A}$, takes a highly structured form. We can define the diagonal terms of the Hessian $A_{\mu\nu}$ and the out of diagonal elements $A_{\mu\nu}$ from the definitions of the overlap. We have:

- The self-overlap q. For any pattern μ in the mixture ($\mu \leq n$), the diagonal element is:
$$q \equiv Q_{\mu\mu} = \mathbb{E}\left[\tanh^{2}(\beta, \boldsymbol{\xi}\cdot\boldsymbol{m})\right]. \tag{A.4}$$
This is the same for all μ due to the symmetry of the mixture.
- The cross-overlap Q. For any two distinct patterns μ, ν within the mixture ($\mu, \nu \leq n; \mu \neq \nu$), the off-diagonal element is:
$$Q \equiv Q_{\mu\nu} = \mathbb{E}\left[\xi^{\mu}\xi^{\nu}\tanh^{2}(\beta\boldsymbol{\xi}\cdot\boldsymbol{m})\right]. \tag{A.5}$$

Using these definitions in Eq. (A.2), the Hessian matrix $\boldsymbol{A}$ has the following structure:

- Diagonal elements: $A_{\mu\mu} \equiv a = 1 - \beta(1-q)$ for all μ.
- Off-diagonal elements within the mixture ($\mu, \nu \leq n; \mu \neq \nu$): $A_{\mu\nu} = \beta Q$.
- Off-diagonal elements involving patterns outside the mixture ($\mu, \nu > n$, $\mu \neq \nu$): $A_{\mu\nu} = 0$.

This structure implies that $\boldsymbol{A}$ is a block matrix:

$$A_{\mu\nu} = \left(\begin{array}{ccccc|ccccc} a & \beta Q & \cdots & \cdots & \beta Q & & & & & \\ \beta Q & a & \beta Q & \cdots & \vdots & & & & & \\ \vdots & \beta Q & a & \ddots & \vdots & & & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \vdots & & & & & \\ \beta Q & \cdots & \cdots & \cdots & a & & & & & \\ \hline & & & & & a & 0 & \cdots & \cdots & 0 \\ & & & & & 0 & a & 0 & \cdots & 0 \\ & & 0 & & & \vdots & \vdots & \ddots & \ddots & \vdots \\ & & & & & \vdots & \vdots & \ddots & \ddots & \vdots \\ & & & & & 0 & 0 & \cdots & \cdots & a \end{array}\right)$$

The Hessian $A_{\mu\nu}$ is a diagonal block matrix: this means that the eigenvalues of the matrix are the eigenvalues of its individual blocks. The first block is relative to a symmetric matrix $n \times n$ while the second is the determinant of a diagonal matrix $P - n \times P - n$. In order to find the eigenvalues we have to calculate the $\det(\underline{A} - \lambda\mathbb{1}) = 0$.

The second block is already diagonal. Therefore, it contributes one distinct eigenvalue [13]:

$$(a-\lambda)^{P-n}=0 \rightarrow \quad \lambda_1 = a = 1-\beta(1-q) \quad \text{with degeneracy} \;\; \nu_1 = P-n. \tag{A.6}$$

The first block has one eigenvector of the form $(1,1,\ldots,1)$, with the eigenvalue:

$$\lambda_3 = a+(n-1)\beta Q = 1-\beta\left[1-q-(n-1)Q\right]. \tag{A.7}$$

This eigenvalue has a degeneracy of $\nu_3 = 1$. The remaining $n-1$ eigenvectors are any vectors orthogonal to $(1,1,\ldots,1)$ (e.g., $(1,-1,0,\ldots,0)$). These all share the same eigenvalue:

$$\lambda_2 = a-\beta Q = 1-\beta(1-q+Q), \tag{A.8}$$

with a degeneracy of $\nu_2 = n-1$. In summary, the Hessian for an n-mixture state has three distinct eigenvalues:

- $\lambda_1 = a = 1-\beta(1-q)$, with $\nu_1 = P-n$.
- $\lambda_2 = a-\beta Q = 1-\beta(1-q+Q)$, with $\nu_2 = n-1$.
- $\lambda_3 = a+(n-1)\beta Q = 1-\beta(1-q-(n-1)Q)$, with $\nu_3 = 1$.

These are the results reported in Sec. 1.3.1.
Now we can give a physical interpretation of these eigenvalues. Each of them corresponds to the curvature of the free energy landscape along a specific type of fluctuation mode:

- λ_3 (Amplitude Mode). This mode corresponds to a fluctuation where the magnitudes of all n non-zero overlaps change uniformly ($\delta m_1 = \delta m_2 = \ldots = \delta m_n$). A positive λ_3 indicates stability against a collective increase or decrease in the strength of the retrieved mixture.
- λ_2 (Symmetry-Breaking Mode): This mode corresponds to fluctuations that break the symmetry between the n patterns. For example, the overlap with one pattern increases while the overlap with another decreases ($\delta m_1 = -\delta m_2$). A positive λ_2 is required for the symmetric mixture to be stable against breaking apart into an asymmetric state.
- λ_1 (Activation Mode): This mode corresponds to fluctuations that activate an overlap with a pattern not currently in the mixture ($\delta m_\mu > 0$ for some $\mu > n$). A positive λ_1 ensures that the system does not spontaneously begin recalling an additional pattern.

A symmetric mixture state is a stable attractor if and only if all three eigenvalues are positive:

$$\lambda_1 > 0, \quad \lambda_2 > 0, \quad \lambda_3 > 0. \tag{A.9}$$

Analyzing these conditions as a function of temperature T and mixture size n we obtain the stability phase diagram described in Sec. 1.7, where odd-n mixtures can become stable at low temperatures, while even-n mixtures are always unstable. In the trivial saddle point ($(\boldsymbol{m}=0)$ $A_{\mu\mu} = \delta_{\mu\nu}(1-\beta)$ so that at $T = T_c = 1$ this state destabilizes.

A.3 Calculation of the Free Energy in the High Load Regime

The calculation of the free energy $f(\boldsymbol{m}, q, r)$ in the RS ansatz in Eq. (1.64) and consequently of the order parameters $\boldsymbol{m}$, q, r in Eqs. (1.65-1.67) start from the expression of the replicated partition function Z^n. We decide to not include in this manuscript the derivation of the Replica Method and to apply it directly to the particular case of the Hopfield Model. For a comprehensive and detailed exposition of the replica method and theory of spin glasses, we refer the reader of course to [97]. The analysis begins with the average replicated partition function, $\mathbb{E}[Z^n]$. The replicated partition function Z^n is defined as:

$$Z^n = \sum_{\boldsymbol{\sigma^1}} .... \sum_{\boldsymbol{\sigma^n}} e^{-\beta \sum_{a=1}^{n} H(\boldsymbol{\sigma^a})} = \sum_{\boldsymbol{\sigma^a}} \prod_{\mu=1}^{P} \exp\left(\frac{\beta}{2N} \sum_{a=1}^{n} \sum_{i,j} \xi_i^\mu \xi_j^\mu \sigma_i^{(a)} \sigma_j^{(a)} \right). \quad \text{(A.1)}$$

We insert now the Hamiltonian in Eq. (1.62) that separates the condensed ($\mu \leq l$) and non-condensed ($\mu > l$) patterns in the expression of the expectation value $\mathbb{E}[Z^n]$:

$$\mathbb{E}[Z^n] = \sum_{\boldsymbol{\sigma}^a} e^{-\frac{\beta n P}{2}} \prod_{\mu=1}^{l} \exp\left(\frac{\beta}{2N} \sum_{a=1}^{n} \left(\sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu \right)^2 \right) \mathbb{E}\left[\prod_{\mu>l}^{P} \exp\left(\frac{\beta}{2N} \sum_{a=1}^{n} \left(\sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu \right)^2 \right) \right]. \quad \text{(A.2)}$$

In the previous expression, the $\mathbb{E}[\cdot]$ is done only on the $P - l$ patterns in which the disorder is present.

A.3.1 Handling the Non-Condensed Patterns (The Noise Term)

We focus on the expectation term for the non-condensed patterns, that we denote by A:

$$A = \mathbb{E}\left[\prod_{\mu>l}^{P} \exp\left(\frac{\beta}{2N} \sum_{a=1}^{n} \left(\sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu \right)^2 \right) \right]. \quad \text{(A.3)}$$

A standard Gaussian linearization (Hubbard-Stratonovich transformation) is applied to each $\mu > l$:

$$\exp\left(\frac{\beta}{2N} \left(\sum_i \sigma_i^{(a)} \xi_i^\mu \right)^2 \right) = \int \frac{dz_a^{(\mu)}}{\sqrt{2\pi}} \exp\left(-\frac{(z_a^{(\mu)})^2}{2} + \sqrt{\frac{\beta}{N}} z_a^{(\mu)} \sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu \right). \quad \text{(A.4)}$$

Introducing n auxiliary Gaussian variables $\boldsymbol{z}^{(\mu)} = (z_1^{(\mu)}, ..., z_n^{(\mu)})$ for each pattern $\mu > l$, the term becomes:

$$A = \mathbb{E}\left[\prod_{\mu>l}^{P} \int D\boldsymbol{z}^{(\mu)} \exp\left(\sqrt{\frac{\beta}{N}} \sum_{a=1}^{n} z_a^{(\mu)} \sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu \right) \right], \quad \text{(A.5)}$$

where

$$\int D\boldsymbol{z}^{(\mu)} = \prod_{a=1}^{n} \int \left(\frac{1}{\sqrt{2\pi}} dz_a^{(\mu)} \exp\left(-\frac{(z_a^{(\mu)})^2}{2} \right) \right). \quad \text{(A.6)}$$

Now we evaluate the $\mathbb{E}[\cdot]$ over the independent, unbiased patterns $\xi_i^\mu = \pm 1$:

$$A = \prod_{\mu>l}^{P} \int D\boldsymbol{z}^{(\mu)} \prod_{i=1}^{N} \mathbb{E}_\xi \left[\exp\left(\sqrt{\frac{\beta}{N}} \xi_i^\mu \sum_{a=1}^{n} z_a^{(\mu)} \sigma_i^{(a)} \right) \right] = \prod_{\mu>l}^{P} \int D\boldsymbol{z}^{(\mu)} \prod_{i=1}^{N} \cosh\left(\sqrt{\frac{\beta}{N}} \sum_{a=1}^{n} z_a^{(\mu)} \sigma_i^{(a)} \right). \tag{A.7}$$

For large N, the logarithm of the product can be expanded:

$$\prod_{i=1}^{N} \cosh(\cdot) = \exp\left(\sum_{i=1}^{N} \ln\cosh(\cdot) \right) \approx \exp\left(\frac{\beta}{2N} \sum_{a,b}^{n} z_a^{(\mu)} z_b^{(\mu)} \sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)} \right), \tag{A.8}$$

where we used $\ln\cosh(x) \approx x^2/2$ for small x. The whole noise factor is therefore:

$$A = \prod_{\mu>l}^{P} \int Dz \exp\left(\frac{\beta}{2N} \sum_{a,b}^{n} z_a^{(\mu)} z_b^{(\mu)} \sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)} \right). \tag{A.9}$$

Defining the spin overlap as

$$Q_{ab} = \sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)} \tag{A.10}$$

the noise term A in Eq. (A.9) simplifies to:

$$A = \prod_{\mu>l}^{P} \int D\boldsymbol{z}^{(\mu)} \exp\left(\frac{\beta}{2N} \sum_{a,b}^{n} z_a^{(\mu)} z_b^{(\mu)} Q_{ab} \right). \tag{A.11}$$

The integral over $\boldsymbol{z}^{(\mu)}$ is an n-dimensional Gaussian integral:

$$\int \exp\left(-\frac{1}{2} \sum_{i,j}^{n} M_{ij} x_i x_j \right) d^n x = \sqrt{\frac{(2\pi)^n}{\det M}}. \tag{A.12}$$

We can define the matrix $\boldsymbol{\mathcal{M}}$ with elements $\mathcal{M}_{ab} = \delta_{ab} - \frac{\beta}{N} Q_{ab}$ and find that the integral to evaluate is:

$$\int D\boldsymbol{z} \exp\left(\frac{\beta}{2N} \sum_{a,b} z_a z_b Q_{ab} \right) = \frac{1}{\sqrt{\det \boldsymbol{\mathcal{M}}}} = \exp\left(-\frac{1}{2} \ln\det\left(\mathbb{1} - \beta \boldsymbol{q} \right) \right), \tag{A.13}$$

with $\boldsymbol{q}$ the intensive overlap matrix:

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)}. \tag{A.14}$$

The product in Eq. (A.9) is over $P - l$ identical terms. However, in the high storage limit, the difference between P and $P - l$, when l is finite, is negligible in thermodynamic limit and we can consider $P - l \approx \alpha N$. Therefore, the full noise contribution A is:

$$A = \exp\left(-\frac{\alpha N}{2} \ln\det\left(\mathbb{1} - \beta \boldsymbol{q} \right) \right). \tag{A.15}$$

Taking up the expression of $\mathbb{E}[Z^n]$ in Eq. (A.2) by entering the expression for A found, we can calculate:

$$\mathbb{E}[Z^n] = \sum_{\boldsymbol{\sigma}^a} \exp\left(-\frac{\beta n \alpha N}{2} - \frac{\alpha N}{2} \ln\det\left(\mathbb{1} - \beta \boldsymbol{q}\right) + \frac{\beta N}{2} \sum_{\mu=1}^{l} \sum_{a=1}^{n} (m_\mu^{(a)})^2\right) \tag{A.16}$$

where $m_\mu^{(a)} = \frac{1}{N}\sum_i \sigma_i^{(a)} \xi_i^\mu$ represents the l nominated pattern overlaps $m_\mu^{(a)}(\boldsymbol{\sigma})$ for each replica a.
The sum over spin configurations is constrained by the definitions of the order parameters $\boldsymbol{q}$ and $\boldsymbol{m}$. This is enforced by inserting integral representations of the Dirac delta functions:

$$1 = \int d\boldsymbol{q}, \prod_{a<b} \delta\left(N q_{ab} - \sum_i \sigma_i^{(a)} \sigma_i^{(b)}\right) \propto \int d\boldsymbol{q} d\boldsymbol{\lambda}, e^{iN \sum_{a<b} \lambda_{ab} \left(q_{ab} - \frac{1}{N}\sum_i \sigma_i^{(a)} \sigma_i^{(b)}\right)}, \tag{A.17}$$

$$1 = \int d\boldsymbol{m}, \prod_{\mu,a} \delta\left(N m_\mu^{(a)} - \sum_i \sigma_i^{(a)} \xi_i^\mu\right) \propto \int d\boldsymbol{m} d\boldsymbol{x}, e^{iN \sum_{\mu,a} x_\mu^{(a)} \left(m_\mu^{(a)} - \frac{1}{N}\sum_i \sigma_i^{(a)} \xi_i^\mu\right)} \tag{A.18}$$

The $\mathbb{E}[Z^n]$ becomes:

$$\mathbb{E}[Z^n] = \int d\boldsymbol{q} d\boldsymbol{m}\, D(\boldsymbol{q}, \boldsymbol{m}) \exp\left(-\frac{\beta n \alpha N}{2} - \frac{\alpha N}{2} \ln\left[\det\left(\mathbb{1} - \beta \boldsymbol{q}\right)\right] + \frac{\beta N}{2} \sum_{\mu=1}^{l} \sum_{a=1}^{n} (m_\mu^{(a)}(\boldsymbol{\sigma}))^2\right) \tag{A.19}$$

with:

$$D(\boldsymbol{q}, \boldsymbol{m}) = \sum_{\boldsymbol{\sigma^a}} \prod_{a<b}^{n} \delta\left(q_{ab} - \frac{1}{N}\sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)}\right) \prod_{\mu=1}^{l} \prod_{a=1}^{n} \delta\left(m_\mu^{(a)}(\boldsymbol{\sigma}) - \frac{1}{N}\sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu\right) \tag{A.20}$$

(A.21)

where

$$d\boldsymbol{q} = \prod_{a<b}^{n} dq_{ab} \quad \text{and} \quad d\boldsymbol{m} = \prod_{\mu=1}^{l} \prod_{a=1}^{n} dm_\mu^{(a)}. \tag{A.22}$$

Using the Fourier integral representation of the Dirac delta functions in Eq. (A.21) we obtain:

$$D(\boldsymbol{q}, \boldsymbol{m}) = \sum_{\boldsymbol{\sigma}^a} \int d\boldsymbol{\lambda} d\boldsymbol{x}, \left(\frac{N}{2\pi}\right)^{n^2+nl} \exp\left[iN\left(\sum_{a<b} \lambda_{ab} q_{ab} + \sum_{a=1}^{n}\sum_{\mu=1}^{l} x_\mu^{(a)} m_\mu^{(a)}\right)\right] \times \tag{A.23}$$

$$\times \exp\left[-i\left(\sum_{a<b} \lambda_{ab} \sum_{i=1}^{N} \sigma_i^{(a)} \sigma_i^{(b)} + \sum_{a=1}^{n}\sum_{\mu=1}^{l} x_\mu^{(a)} \sum_{i=1}^{N} \sigma_i^{(a)} \xi_i^\mu\right)\right], \tag{A.24}$$

with the integration measures:

$$d\boldsymbol{\lambda} = \prod_{a,b}^{n} d\lambda_{ab} \quad \text{and} \quad d\boldsymbol{x} = \prod_{\mu=1}^{l} \prod_{a=1}^{n} dx_\mu^{(a)}. \tag{A.25}$$

The last term of the previous expression can be written considering that it factors completely over the site index i. Consequently, the sum over all replicated spin configurations $\boldsymbol{\sigma}^a$ also factors over sites:

$$\sum_{\boldsymbol{\sigma}^a} \prod_{i=1}^{N} (\cdots) = \prod_{i=1}^{N} \left(\sum_{\sigma_i^{(a)}} (\cdots) \right). \tag{A.26}$$

The quantity $D(\boldsymbol{q}, \boldsymbol{m})$ can therefore be rewritten as:

$$D(\boldsymbol{q}, \boldsymbol{m}) = \left(\frac{N}{2\pi} \right)^{n^2+nl} \int D\boldsymbol{\lambda} D\boldsymbol{x}, e^{iN\Phi[\boldsymbol{\lambda}, \boldsymbol{x}]} \left(\sum_{\sigma^a} e^{-i\Psi[\sigma^a, \boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{\xi}_i]} \right)^N, \tag{A.27}$$

where we define:

$$\Phi[\boldsymbol{\lambda}, \boldsymbol{x}] = \sum_{a<b} \lambda_{ab} q_{ab} + \sum_{a=1}^{n} \sum_{\mu=1}^{l} x_\mu^{(a)} m_\mu^{(a)}, \tag{A.28}$$

$$\Psi[\sigma^a, \boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{\xi}] = \sum_{a<b} \lambda_{ab} \sigma^{(a)} \sigma^{(b)} + \sum_{a=1}^{n} \sigma^{(a)} \left(\sum_{\mu=1}^{l} x_\mu^{(a)} \xi^\mu \right). \tag{A.29}$$

The sum in the last term, $\sum_{\sigma^a} e^{-i\Psi[\cdots]}$, depends on the site i *only* through the specific realization of the condensed patterns $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^l)$ stored at that site. In thermodynamic limit ($N \to \infty$), the average over these quenched disorder variables $\boldsymbol{\xi}_i$ is performed. Following the same logic leading to Eq. (A.15), the average over sites is equivalent to an average over the distribution of the quenched patterns:

$$\frac{1}{N} \sum_{i=1}^{N} O(\boldsymbol{\xi}_i) \longrightarrow \mathbb{E}_{\boldsymbol{\xi}}[O(\boldsymbol{\xi})] = \frac{1}{2^l} \sum_{\boldsymbol{\xi} \in \{-1,+1\}^l} O(\boldsymbol{\xi}) \quad \text{as } N \to \infty. \tag{A.30}$$

Thus, the sum over sites is replaced by an expectation over the pattern distribution, leading to the final form:

$$\left(\sum_{\sigma^a} e^{-i\Psi[\cdots]} \right)^N = \exp\left(N \ln \left(\sum_{\sigma^a} e^{-i\Psi[\cdots]} \right) \right) \longrightarrow \exp\left(N \mathbb{E}_{\boldsymbol{\xi}} \left[\ln \left(\sum_{\sigma^a} e^{-i\Psi[\sigma^a, \boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{\xi}]} \right) \right] \right). \tag{A.31}$$

The resulting expression of the mean replicated partition function $\mathbb{E}[Z^n]$ can be written in the form:

$$\mathbb{E}[Z^n] = \int d\boldsymbol{q} d\boldsymbol{m} \exp[-\beta n N f_N(\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\lambda})] \tag{A.32}$$

where

$$f_N(\boldsymbol{q}, \boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\lambda}) = \frac{\alpha}{2} + \frac{\alpha}{2\beta n} \ln \det(\mathbb{1} - \beta \boldsymbol{q}) - \frac{1}{2n} \sum_{\mu,a} (m_\mu^{(a)})^2 - \frac{1}{\beta n} \left[i \sum_{a<b} \lambda_{ab} q_{ab} + i \sum_{\mu,a} x_\mu^{(a)} m_\mu^{(a)} \right] + \tag{A.33}$$

$$- \frac{1}{\beta n} \mathbb{E} \left[\ln \left(\sum_{\sigma^a} \exp\left(-i \sum_{a<b} \lambda_{ab} \sigma^a \sigma^b - i \sum_{\mu,a} x_\mu^{(a)} \sigma^a \xi^\mu \right) \right) \right]. \tag{A.34}$$

In thermodynamic limit ($N \to \infty$), the integrals are evaluated by the saddle-point method. The intensive free energy is then:

$$f = -\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{\beta n N}\ln \mathbb{E}[Z^n] = \lim_{n\to 0}\operatorname{extr}_{\boldsymbol{q},\boldsymbol{m},\boldsymbol{\lambda},\boldsymbol{x}} f_N(\boldsymbol{q},\boldsymbol{m},\boldsymbol{\lambda},\boldsymbol{x}). \tag{A.35}$$

A.3.2 Replica Symmetric Ansatz

Requiring the stationarity from Eq. (A.34) we obtain:

$$\frac{\partial f_N}{\partial m_\mu^{(a)}} = 0 \to\ x_\mu^{(a)} = i\beta m_\mu^{(a)} \tag{A.36}$$

and:

$$\frac{\partial f_N}{\partial q_{ab}} = 0 \to\ \lambda_{ab} = i\frac{\beta\alpha}{2}(\mathbb{1} - \beta\boldsymbol{q})^{-1}_{ab}. \tag{A.37}$$

where we have used the following relation in the calculation of the last derivative:

$$\frac{\partial \ln\det(\boldsymbol{M})}{\partial M_{il}} = (\boldsymbol{M})^{-1}_{ij}. \tag{A.38}$$

Eq. (A.36) can be used to eliminate the $x_\mu^{(a)}$ from the free energy Eq. (A.34); this gives a pseudo-free energy which coincides with the true free energy at its saddle points:

$$f_N(\boldsymbol{q},\boldsymbol{m},\boldsymbol{\lambda}) = \frac{\alpha}{2} + \frac{\alpha}{2\beta n}\ln\det(\mathbb{1}-\beta\boldsymbol{q}) + \frac{1}{2n}\sum_{\mu=1}^{l}\sum_{a=1}^{n}(m_\mu^{(a)}(\boldsymbol{\sigma}))^2 + \tag{A.39}$$

$$-\frac{1}{\beta n}\left[i\sum_{a<b}\lambda_{ab}q_{ab} + \mathbb{E}\left[\ln\left(\sum_{\boldsymbol{\sigma^a}}\exp\left(-i\sum_{a<b}\lambda_{ab}\sigma^{(a)}\sigma^{(b)} + \beta\sum_{a,\mu}m_\mu^{(a)}\sigma^{(a)}\xi^\mu\right)\right)\right]\right]. \tag{A.40}$$

(A.41)

At this point, it is necessary to assume a precise form for the overlap order parameters. The extremization is performed under the Replica Symmetric (RS) Ansatz [33, 13]:

$$m_\mu^{(a)} = m_\mu\ \forall a,\ q_{ab} \quad = \delta_{ab} + q(1-\delta_{ab}),\ \lambda_{ab} = \frac{i\alpha\beta^2 r}{2}(1-\delta_{ab}),\ x_\mu^{(a)} \quad = i\beta m_\mu.$$

This ansatz has a clear interpretation:

- $m_\mu^{(a)}(\boldsymbol{\sigma})$ is the Mattis magnetization between the state of the system and the μth pattern that is assumed to be replica-independent.
- q_{ab} is the spin glass order parameter defined even as Edward-Anderson order parameter [49] that indicates the overlap of different spin replicas. The diagonal is fixed to 1 ($q_{aa} = 1$), since each replica has maximal overlap with itself, and all off-diagonal elements are assumed equal to q, representing the typical overlap between different replicas.

- $r_{ab} = \frac{1}{\alpha} \sum_{\mu \geq 2} m_a^\mu \, m_b^\mu$ is the overlaps corresponding to non-nominated patterns. The factor $1/\alpha$ effectively makes it an average over the P overlaps, and cancels the dependence on $1/\sqrt{N}$ of the m_μ. From this parameter we define λ_{ab} as the conjugate to q_{ab}.

Under the RS ansatz, the matrix $\mathbf{\Lambda} = \mathbb{1} - \beta \boldsymbol{q}$ in the free energy expression in Eq. (A.41), has a specific structure: the diagonal elements are $\Lambda_{aa} = 1 - \beta$ and the off-diagonal elements are $\Lambda_{ab} = -\beta q$ for $a \neq b$. Its determinant can be computed by finding its eigenvalues. The structure is identical to that of the Hessian matrix in the low load analysis in Sec. A.2. The eigenvalues and their multiplicities are:

$$\begin{aligned} \lambda_1 &= 1 - \beta(1-q) - \beta q n && \text{multiplicity } \nu_1 = 1 \\ \lambda_2 &= 1 - \beta(1-q) && \text{multiplicity } \nu_2 = n - 1. \end{aligned}$$

Thus, the log-determinant is:

$$\ln \det(\mathbf{\Lambda}) = \ln \left(\prod_{i=1}^{2} \lambda_i^{\nu_i} \right) = \ln(\lambda_1) + (n-1) \ln(\lambda_2) = \tag{A.42}$$

$$= \ln\left(1 - \beta(1-q) - \beta q n\right) + (n-1) \ln\left(1 - \beta(1-q)\right). \tag{A.43}$$

Taking the limit $n \to 0$ and expanding to first order in n we have:

$$\frac{1}{n} \ln \det(\mathbf{\Lambda}) = \ln\left(1 - \beta(1-q)\right) - \frac{\beta q}{1 - \beta(1-q)} + \mathcal{O}(n). \tag{A.44}$$

The final step is to evaluate the term containing the sum over spin configurations in Eq. (A.41) under the RS ansatz. Substituting the form for λ_{ab} and $m_\mu^{(a)}$, this term becomes:

$$\mathbb{E}_{\boldsymbol{\xi}} \left[\ln \left(\sum_{\sigma^a} \exp \left(\frac{\alpha \beta^2 r}{2} \sum_{a \neq b} \sigma^{(a)} \sigma^{(b)} + \beta \sum_{a=1}^{n} \sigma^{(a)} \left(\sum_{\mu=1}^{l} m_\mu \xi^\mu \right) \right) \right) \right]. \tag{A.45}$$

The sum over $a \neq b$ can be rewritten using the identity $\sum_{a \neq b} \sigma^{(a)} \sigma^{(b)} = \left(\sum_a \sigma^{(a)} \right)^2 - n$. The term $-n$ contributes a constant which can be absorbed. The crucial step is to linearize the square, $\left(\sum_a \sigma^{(a)} \right)^2$, using a Hubbard-Stratonovich transformation. This introduces a Gaussian auxiliary variable z:

$$\exp\left(\frac{\alpha \beta^2 r}{2} \left(\sum_{a=1}^{n} \sigma^{(a)} \right)^2 \right) = \int Dz \exp\left(\beta \sqrt{\alpha r} z \sum_{a=1}^{n} \sigma^{(a)} \right), \tag{A.46}$$

where $\int Dz = \int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}$. After this transformation, the sum over replicas factorizes completely:

$$\sum_{\sigma^a} \exp\left(\beta \sum_{a=1}^{n} \sigma^{(a)} \left(\sqrt{\alpha r} z + \boldsymbol{m} \cdot \boldsymbol{\xi} \right) \right) = \prod_{a=1}^{n} \sum_{\sigma^{(a)} = \pm 1} \exp\left(\beta \sigma^{(a)} \left(\sqrt{\alpha r} z + \boldsymbol{m} \cdot \boldsymbol{\xi} \right) \right) = \tag{A.47}$$

$$= \prod_{a=1}^{n} 2\cosh\left(\beta\left(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi}\right)\right) = \left[2\cosh\left(\beta\left(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi}\right)\right)\right]^n . \tag{A.48}$$

Taking the logarithm and the limit $n \to 0$, using the identity $\lim_{n\to 0}(X^n - 1)/n = \ln X$, we end up with:

$$\frac{1}{n}\ln\left(\sum_{\sigma^a}(\cdots)\right) \longrightarrow \int Dz, \ln\left[2\cosh\left(\beta\left(\sqrt{\alpha r}, z + \boldsymbol{m}\cdot\boldsymbol{\xi}\right)\right)\right] + \text{const.} \tag{A.49}$$

Assembling all these results, we arrive at the final expression for the intensive replica-symmetric free energy per spin:

$$f_{\text{RS}}(q, r, \boldsymbol{m}) = \frac{\alpha}{2} + \frac{\alpha}{2\beta}\left[\ln(1-\beta(1-q)) - \frac{\beta q}{1-\beta(1-q)}\right] + \frac{1}{2}|\boldsymbol{m}|^2 \tag{A.50}$$

$$+\frac{1}{2}\alpha\beta r(1-q) - \frac{1}{\beta}\int Dz\, \mathbb{E}_{\boldsymbol{\xi}}\left[\ln\left(2\cosh\left(\beta\left(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi}\right)\right)\right)\right] . \tag{A.51}$$

This is the expression reported in Eq. (1.64).
This equation gives a complete description of the stable states of the system, in the limit $N \to \infty$. The mean-field theory, is exact because the network is fully connected, as in the SK model. The equations of state are derived by extremization with respect to the order parameters. They correspond to the minima of the free energy f and represent the expectation values in thermodynamic limit:

$$\frac{\partial f}{\partial \boldsymbol{m}} = 0 \quad \rightarrow \quad \boldsymbol{m} = \int Dz\ \mathbb{E}\Big[\boldsymbol{\xi}\tanh\Big(\beta\Big(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi}\Big)\Big)\Big] \tag{A.52}$$

Equation (A.52) generalizes the mean-field equation of the $\alpha \to 0$ limit: in this case, the Gaussian integral becomes trivial and we recover the low load regime formulation of the Eq. (A.16).
The derivation of the free energy f wit respect to the overlap q gives a relation between the spin glass order parameter q and its conjugate r whose expression is in Eq. (1.67):

$$\frac{\partial f}{\partial q} = 0 \quad \rightarrow \quad r = \frac{q}{\left(1-\beta(1-q)\right)^2} \tag{A.53}$$

The derivation of the q equation reported in the main text in Eq. (1.66) involves an application of Stein's lemma (or integration by parts) during the extremization with respect to r. Suppose x is a normally distributed random variable with expectation μ and variance σ^2 and F a differentiable function for which $\mathbb{E}[F'(x)] < \infty$. Then:

$$\mathbb{E}[(x-\mu)F(x)] = \sigma^2\,\mathbb{E}[F'(x)]. \tag{A.54}$$

In the calculation of q, the normally distributed random variable x is Gaussian integration variable z and the differentiable function $F(x)$ is $\tanh(\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi}))$. In particular, we have:

$$\frac{\partial f}{\partial r} = 0 \ \rightarrow \alpha\beta(1-q) - \sqrt{\frac{\alpha}{r}}\int Dz\ \mathbb{E}\left[\tanh\Big(\Big(\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi})\Big)z\right] = 0 \tag{A.55}$$

$$\alpha\beta(1-q) - \sqrt{\frac{\alpha}{r}} \int Dz \; \mathbb{E}\left[\Big(1 - \tanh^2\big((\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi})\big)\Big)\beta\sqrt{\alpha r}\right] = 0 \tag{A.56}$$

$$\alpha\beta(1-q) - \alpha\beta + \alpha\beta \int Dz \; \mathbb{E}[\tanh^2((\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi})))] = 0 \tag{A.57}$$

$$q = \int Dz \; \mathbb{E}[\tanh^2((\beta(\sqrt{\alpha r}z + \boldsymbol{m}\cdot\boldsymbol{\xi})))]. \tag{A.58}$$

where $q > 0$ indicates a frozen disorder within the structure of metastable states. This is exactly the expression reported in Eq. (1.66) in the main text.

Appendix B

Appendix for Chapter 5

B.1 Algorithmic Implementation

In this section we present the algorithmic framework employed in our numerical simulations. To characterize the dynamical properties of the networks, we generate ensembles of connectivity matrices $\mathbf{J}$ with specified correlation η, dilution ρ, and asymmetry parameter a_1 across various system sizes N. For each parameter set, we iterate through all 2^N initial configurations under the deterministic dynamics defined in Eq.(5.1). After transient behavior, the system converges to an attractor (fixed point or cycle), and key statistical properties are computed. Multiple realizations (S samples) of the coupling matrix are simulated for each parameter combination to ensure statistical reliability.
The core simulation structure is summarized in Algorithm 1:

Algorithm 1 Core Simulation Loop

```
Input: N, ρ, η, a_1, S                          ▷ Network parameters
for s = 1 to S do
    J ← MakeJ(N, ρ, η, a_1)                     ▷ Generate synaptic matrix J
    FillNextConf(J)                             ▷ Compute state transitions
    ComputeProperties()                         ▷ Extract dynamical features
end for
```

This algorithm delineates the iterative process of generating coupling matrices, evolving spin configurations and computing relevant properties across multiple samples to comprehensively assess the behavior of the neural networks under consideration. In the following sections, we briefly describe all the function used.

B.1.1 State Transition Mapping

The function FillNextConf computes the subsequent configuration $\boldsymbol{\sigma}(t+1)$ from each initial configuration $\boldsymbol{\sigma}(t)$ using the deterministic parallel dynamics. Given the finite phase space (2^N configurations) and deterministic evolution, each initial configuration maps uniquely to its successor. The next configuration is chosen randomly but is unique. Once this correspondence is established, the dynamics can

be, therefore, executed only once per sample. Specifically, the process is outlined as follows:

Algorithm 2 FillNextConf - State Transition Computation

Input: J (synaptic matrix), initial 2^N configurations
for $\mathcal{C} = 0$ **to** $2^N - 1$ **do**
  $\boldsymbol{\sigma} \leftarrow$ DecodeConfiguration($\mathcal{C}$) ▷ Convert integer to spin vector
  for $i = 0$ **to** $N - 1$ **do**
    $h_i \leftarrow \sum_{j=0}^{N-1} J_{ij}\sigma_j$ ▷ Compute local field
    $\sigma_i' \leftarrow \mathrm{sgn}(h_i)$ ▷ Parallel update
  end for
  $\mathcal{C}' \leftarrow$ EncodeConfiguration($\boldsymbol{\sigma}'$) ▷ Convert back to integer
  nextConf[$\mathcal{C}$] $\leftarrow \mathcal{C}'$ ▷ Store transition
end for
Output: next 2^N configurations

Configurations are encoded as integers where each bit represents a spin value. Bitwise operations enable efficient encoding/decoding between integer representations and spin vectors. The configuration $\mathcal{C}$ is interpreted as an integer binary number, where each bit b corresponds to one spin $\sigma_i = 1 - 2(b \,\&\, 1)$. A right shift $b >>= 1$ moves to the next bit. During the execution of parallel dynamics, the obtained spin vector needs to be re-associated with a configuration $\mathcal{C}$ among the 2^N possible. A loop iterates over the elements of $\boldsymbol{\sigma}$ from the most significant bit to the least significant bit, evaluating a binary representation of σ_i': $b+ = (1 - \sigma_i')/2$ followed by a left shift by one. Finally, a right shift by one bit is performed to discard the extra bit introduced by the last left shift. This process results in a binary number derived from $\boldsymbol{\sigma'}$, associated with a configuration $\mathcal{C}'$.

B.1.2 Attractor Analysis

The function ComputeProperties identifies attractors and characterizes their properties by tracing trajectories from all initial conditions. Within a loop over the configurations $\mathcal{C}$, we save the trajectories and create a position array pos of 2^N entries, all initialized to 0 except for the first, which is set to 1. Until a configuration already present is encountered, the loop proceeds to the next one, updating the corresponding entry in pos to 1. When pos[$\mathcal{C}$]$\neq 0$, indicating that a configuration has been encountered before, either a fixed point or a cycle has been found, and we save the label of the configuration (for a cycle, we save the minimum label). A synthesis of the function is provided in the following pseudocode:

Algorithm 3 ComputeProperties - Attractor Analysis

```
numAttractors ← 0
Initialize arrays: basinSize[], cycleLen[], meanDist[]
for C = 0 to 2^N − 1 do
    traj ← [C]                                   ▷ Initialize trajectory
    pos[C] ← 1                                   ▷ Mark initial position
    curr ← C, step ← 1
    while pos[nextConf[curr]] = 0 do
        curr ← nextConf[curr]
        step ← step + 1
        traj.push(curr)
        pos[curr] ← step
    end while
    first ← pos[nextConf[curr]]
    label ← min(traj[first..step])               ▷ Minimal label in cycle
    if cycleLen[label] = 0 then                  ▷ New attractor found
        numAttractors ← numAttractors + 1
        cycleLen[label] ← step + 1 − first
    end if
    basinSize[label] ← basinSize[label] + 1
    meanDist[label] ← meanDist[label] + (first − 1)
end for
```

B.1.3 Boundary Analysis

The chaos probability p_{chaos} (discussed in Sec. 5.3) quantifies basin boundary complexity and is calculated using the following function:

Algorithm 4 ComputeBoundaries - Chaos Probability

```
count ← 0
for C_1 = 0 to 2^N − 1 do
    for i = 0 to N − 1 do
        C_2 ← C_1 ⊕ (1 ≪ i)                      ▷ Flip spin i
        if label[C_1] ≠ label[C_2] then
            count ← count + 1
        end if
    end for
end for
p_chaos ← count/(N · 2^N)
```

From each of the 2^N initial configurations, we create a new configuration by flipping only one spin at position i ($\mathcal{C}_2 \leftarrow \mathcal{C}_1 \oplus (1 \ll i)$) and then verify if these two configurations are in different basins.

B.1.4 Synaptic Matrix Generation

The function MakeJ generates the coupling matrix $\mathbb{J}$ with controlled topology. This component of the code is the only aspect that undergoes modification to account different levels or symmetry, dilution and topology.

Algorithm 5 MakeJ - Synaptic Matrix Generation

procedure MakeJ(N, ρ, a_1, η)
  $a_2 \leftarrow \rho - a_1/2$, $a_0 \leftarrow 1 - a_1 - a_2$
  Initialize $\mathbf{A}$ as $N \times N$ matrix of ones
  Set $\mathbf{A}_{ii} \leftarrow 0$ for all i ▷ No self-connections
  for $i = 0$ **to** $N - 1$ **do**
    for $j = 0$ **to** $i - 1$ **do**
      $r \sim \mathcal{U}[0, 1]$
      if $r < a_0$ **then**
        $A_{ij}, A_{ji} \leftarrow 0$ ▷ No connection
      else if $r < a_0 + a_2$ **then**
        $A_{ij}, A_{ji} \leftarrow 1$ ▷ Double connection
      else
        $A_{ij} \leftarrow 1$, $A_{ji} \leftarrow 0$ **or** vice versa ▷ Single connection
      end if
    end for
  end for
  for $i = 0$ **to** $N - 1$ **do**
    for $j = 0$ **to** $N - 1$ **do**
      if $A_{ij} = 1$ **then**
        $J_{ij} \sim \mathcal{N}(0, 1)$ with correlation η for symmetric pairs
      else
        $J_{ij} \leftarrow 0$
      end if
    end for
  end for
  return J
end procedure

This method is different from the one proposed in [54] where the synaptic matrix $\mathbb{J}$ is created as a combination of a symmetric and anti-symmetric matrix as discussed in Appendix C.1.

Appendix C

Appendix for Chapter 6

C.1 Comparison with Previous Formulations

In order to make this work more readable, we report here the comparison with our method of generation of the synaptic matrix J described in Sec. 5.2 with the one followed in [54, 87] which is related to an implicit symmetry in the network's topology. The coupling matrix elements are there realized through the relation:

$$J_{ij} = \left(1 - \frac{\epsilon}{2}\right) S_{ij} + \frac{\epsilon}{2} A_{ij} \tag{C.1}$$

with $S_{ij} = S_{ji}$ symmetric and $A_{ij} = -A_{ji}$ antisymmetric. In this case, the elements of S_{ij} and A_{ij} are extracted from the distribution

$$P(x) = (1 - \bar{\rho})P_0(x) + \bar{\rho}\delta(x) \tag{C.2}$$

with $P_0(x) = \mathcal{N}(0, 1)$ or $P_0(x) = \frac{1}{2}\theta(1 - x^2)$. The parameters are:

- ϵ indicating the degree of symmetry of the couplings,
- $\bar{\rho}$, which counts the density of zeros in the matrix[1].

Following the previous approach, the elements of the lower triangular matrix are first randomly generated from a chosen distribution ($\mathcal{N}(0, 1)$, $\mathcal{U}[-1, 1]$, etc.), with a probability $\bar{\rho}$ of being set to zero. The upper triangular matrix is then constructed for both symmetric and antisymmetric cases, with the diagonals set to zero in both matrices (for A it is automatic). Therefore, the two matrices S and A are created by filling the lower triangular matrix so that

$$\mathbb{P}[S_{ij} = 0] = \mathbb{P}[A_{ij} = 0] = \bar{\rho}. \tag{C.3}$$

From Eq. (C.1) we obtain[2]

$$\mathbb{P}[J_{ij} = 0] = \bar{\rho}^2(1 - \delta_{\epsilon,0}) + \bar{\rho}\delta_{\epsilon,0}. \tag{C.4}$$

[1] Note: While [54, 87] used ρ for this parameter, we employ $\bar{\rho}$ to maintain consistency with our notation and facilitate comparison.

[2] If we also consider values of $\epsilon \in [0, 2]$ (so including even the completely antisymmetric case $\epsilon = 2$), the previous expression must be modified in this way: $\mathbb{P}[J_{ij} = 0] = \bar{\rho}^2\big(1-\delta_{\epsilon,0}\big)\big(1-\delta_{\epsilon,2}\big)+\bar{\rho}(\delta_{\epsilon,0}+\delta_{\epsilon,2})$.

Fully asymmetric case $\epsilon = 1$ In the fully asymmetric case, the synaptic coupling matrix J_{ij} is determined by the sum of symmetric S_{ij} and antisymmetric A_{ij} components, each generated for example from Gaussian distributions. Specifically, the elements of J_{ij} exhibit the following probabilistic behavior:

$$J_{ij} = \begin{cases} 0 & \text{with probability } p = \bar{\rho}^2 \text{ if } S_{ij} = 0 \text{ and } A_{ij} = 0 \\ \frac{z}{2} & \text{with probability } p = 2\bar{\rho}(1-\bar{\rho}) \text{ if } S_{ij} = z \text{ and } A_{ij} = 0, \text{ or vice versa} \\ \frac{z}{\sqrt{2}} & \text{with probability } p = (1-\bar{\rho})^2 \text{ if } S_{ij} \text{ and } A_{ij} \sim \mathcal{N}(0,1) \end{cases} \tag{C.5}$$

This formulation leverages the fact that the sum of two independent Gaussian random variables X and Y follows a normal distribution with mean $\mu_Z = \mu_X + \mu_Y$ and variance $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$. Consequently, J_{ij}, being half of the sum of such variables is distributed as $J_{ij} \sim \mathcal{N}(0, 1/2)$[3]. By standardization, J_{ij} can be expressed in terms of the standard normal variable $z \sim \mathcal{N}(0,1)$ returning the expressions shown. Moreover, examining pairs of elements (J_{ij}, J_{ji}) in the J matrix reveals:

$$(J_{ij}, J_{ji}) = \begin{cases} (0,0) & \text{with } p = \bar{\rho}^2 \\ \left(\frac{z}{2}, \frac{z}{2}\right) & \text{with } p = \bar{\rho}(1-\bar{\rho}) \\ \left(\frac{z}{2}, -\frac{z}{2}\right) & \text{with } p = \bar{\rho}(1-\bar{\rho}) \\ \left(\frac{z_1}{\sqrt{2}}, \frac{z_2}{\sqrt{2}}\right) & \text{with } p = (1-\bar{\rho})^2 \end{cases} \tag{C.6}$$

This analysis demonstrates that only in the dense situation, where $\bar{\rho} = 0$, J_{ij} and J_{ji} are mutually independent, since $S + A$ differs from $S - A$. However, in diluted scenarios, this independence is not maintained.
Due to the creation of J as a linear combination of symmetric and antisymmetric matrices, automatically $\mathbb{P}[J_{ij} = 0 | J_{ji} = 0] = 1$ so both elements of the couple (J_{ij}, J_{ji}) are simultaneously set to zero. This means that:

$$a_0 = \mathbb{P}[J_{ij} = 0 \;\&\; J_{ji} = 0] = \mathbb{P}[J_{ij} = 0 \mid J_{ji} = 0] \cdot \mathbb{P}[J_{ij} = 0] \implies \tag{C.7}$$

$$a_0 = \mathbb{P}[J_{ij} = 0] = \bar{\rho}^2(1 - \delta_{\epsilon,0})(1 - \delta_{\epsilon,2}) + \bar{\rho}(\delta_{\epsilon,0} + \delta_{\epsilon,2}). \tag{C.8}$$

For complementarity, can be obtained even a_2 that indicates the probability of double links in J.
From the point of view of couplings symmetry we have that:

- The fully symmetric case $\epsilon = 0$ and the fully asymmetric case $\epsilon = 2$ are different from all others with the same dilution.
- If $\epsilon \in (0, 1]$ symmetry does not change the nature of the links (it only depends on $\bar{\rho}$).

So at different levels of asymmetry, changes in observables (average number and effective number of attractors, cycle lengths, etc.) are due to the intensity of the links (and their correlation) and not to the matrix structure (the number of direct, double links, and single nodes does not change with asymmetry at the same dilution).

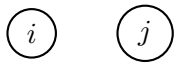
[3] $Var(aZ) = a^2 Var(Z)$.

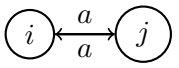
The critical value $\epsilon_c = 0.84$ in the dense case [76], where there is a transition from the oscillatory regime (many attractors with cycles of $L = 1, 2$) to a chaotic regime with a linear average number of attractors in N and an exponential cycles length, is not due to the matrix structure, which does not undergo variations with symmetry.

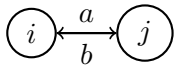
Topological Constraints and Implications This construction imposes specific topological constraints:

- **Absence of directed links**: $a_1 = 0$ for all parameter values
- **Topological symmetry**: $J_{ij} = 0 \iff J_{ji} = 0$ (zeros occur in pairs)
- **Fully connected case ($\bar{\rho} = 0$)**: Only double links exist, which are:
  - Equal if $\epsilon = 0$ (symmetric)
  - Different if $\epsilon \in (0, 2]$ (asymmetric to antisymmetric)

Graphically:

- with $a_0 = \bar{\rho}^2(1 - \delta_{\epsilon,0})(1 - \delta_{\epsilon,2}) + \bar{\rho}(\delta_{\epsilon,0} + \delta_{\epsilon,2})$ 

- with $p = (1 - \delta_{\epsilon,2})(1 - \bar{\rho})\Big((1 - \delta_{\epsilon,0})\bar{\rho} + \delta_{\epsilon,0}\Big)$ 

- with $p = (1 - \bar{\rho})(1 - \delta_{\epsilon,0})$ 

The last two relations sum up to a_2.

Calculating $\mathbb{E}[J_{ij}J_{ji}]$ yields:

$$\mathbb{E}[J_{ij}J_{ji}] = \mathbb{E}\left[\left(\left(1 - \frac{\epsilon}{2}\right)S_{ij} + \frac{\epsilon}{2}A_{ij}\right)\left(\left(1 - \frac{\epsilon}{2}\right)S_{ij} - \frac{\epsilon}{2}A_{ij}\right)\right] \tag{C.9}$$

$$= \left(1 - \frac{\epsilon}{2}\right)^2 \mathbb{E}[S_{ij}^2] - \frac{\epsilon^2}{4}\mathbb{E}[A_{ij}^2] = (1 - \bar{\rho})(1 - \epsilon) \tag{C.10}$$

where the anti-symmetry of A, the independence of S_{ij} and A_{ij}, and the fact that if not null (with $p = 1 - \bar{\rho}$) they are distributed as normal variables with mean $\mathbb{E}[z] = 0$ and variance $\mathbb{E}[z^2] = 1$ are used. Similarly, it is found that $\mathbb{E}[J_{ij}^2]$ is:

$$\mathbb{E}[J_{ij}^2] = \mathbb{E}\left[\left(\left(1 - \frac{\epsilon}{2}\right)S_{ij} + \frac{\epsilon}{2}A_{ij}\right)^2\right] = (1 - \bar{\rho})\left(1 - \epsilon + \frac{\epsilon^2}{2}\right). \tag{C.11}$$

This formulation reveals that changes in dynamical observable across different asymmetry levels (ϵ) are primarily driven by variations in coupling strengths and their correlations, rather than structural changes in the network topology, which remains constant for fixed $\bar{\rho}$.

We report the formulation for the two methods in order to make visible the differences. For clarity, we define the following quantities for the probabilities associated with double links:

$\tilde{a}_2$: the probability to have a double link with $J_{ij} = J_{ji}$;
$\hat{a}_2$: the probability to have a double link with $J_{ij} \neq J_{ji}$.

with $\tilde{a}_2 + \hat{a}_2 = a_2$ defined in Eq. (5.9).

Correspondence in the Absence of Direct Links ($a_1 = 0$) A direct comparison between the two methods is possible when direct links are absent, $a_1 = 0$ in our method. Indeed, the previous formulation inherently assumes $a_1 = 0$.

Our Method (a_1, η, ρ)

$$a_1 : \text{given} \tag{C.12}$$

$$a_2 = \rho - \frac{a_1}{2} \tag{C.13}$$

$$a_0 = 1 - a_2 - a_1 \tag{C.14}$$

$$\tilde{a}_2 = \delta_{\eta,1} a_2 \tag{C.15}$$

$$\hat{a}_2 = (1 - \delta_{\eta,1}) a_2 \tag{C.16}$$

$$\mathbb{P}[J_{ij} = J_{ji}] = \delta_{\eta,1} a_2 + a_0 \tag{C.17}$$

$$\frac{\mathbb{E}[J_{ij} J_{ji}]}{\mathbb{E}[J_{ij}^2]} = \eta \tag{C.18}$$

Previous Formulation $(\epsilon, \bar{\rho})$

$$a_1 = 0 \quad \forall \epsilon, \bar{\rho} \tag{C.19}$$

$$a_2 = (1-\bar{\rho}^2)(1-\delta_{\epsilon,0})(1-\delta_{\epsilon,2}) + (1-\bar{\rho})(\delta_{\epsilon,0}+\delta_{\epsilon,2}) \tag{C.20}$$

$$a_0 = 1 - a_2 \tag{C.21}$$

$$\tilde{a}_2 = (1 - \delta_{\epsilon,2})(1 - \bar{\rho})\Big((1 - \delta_{\epsilon,0})\bar{\rho} + \delta_{\epsilon,0}\Big) \tag{C.22}$$

$$\hat{a}_2 = (1 - \bar{\rho})(1 - \delta_{\epsilon,0}) \tag{C.23}$$

$$\mathbb{P}[J_{ij} = J_{ji}] = \delta_{\epsilon,0} + \bar{\rho}(1 - \delta_{\epsilon,0}) \tag{C.24}$$

$$\frac{\mathbb{E}[J_{ij} J_{ji}]}{\mathbb{E}[J_{ij}^2]} = \frac{1 - \epsilon}{1 - \epsilon + \epsilon^2/2} \tag{C.25}$$

In particular, the fully correlated case $\eta = 1$ can be mapped easily to the previous formulation's $\epsilon = 0$ by setting $\rho = 1 - \bar{\rho}$; the same for completely the antisymmetric case $\epsilon = 2$ ($\eta = -1$).

Our Method $(a_1 = 0, \eta = 1, \rho)$

$$a_2 = \rho, \quad a_0 = 1 - \rho \tag{C.26}$$

$$\tilde{a}_2 = a_2 = \rho, \quad \hat{a}_2 = 0 \tag{C.27}$$

$$\mathbb{P}[J_{ij} = J_{ji}] = a_2 + a_0 = 1 \tag{C.28}$$

Previous Formulation $(\epsilon = 0, \bar{\rho})$

$$a_2 = 1 - \bar{\rho}, \quad a_0 = \bar{\rho} \tag{C.29}$$

$$\tilde{a}_2 = 1 - \bar{\rho}, \quad \hat{a}_2 = 0 \tag{C.30}$$

$$\mathbb{P}[J_{ij} = J_{ji}] = 1 \tag{C.31}$$

Our Method $(a_1 = 0, \eta = -1, \rho)$

$$a_2 = \rho, \quad a_0 = 1 - \rho \tag{C.32}$$

$$\tilde{a}_2 = 0, \quad \hat{a}_2 = a_2 = \rho \tag{C.33}$$

$$\mathbb{P}[J_{ij} = J_{ji}] = a_0 = 1 - \rho \tag{C.34}$$

Previous Formulation $(\epsilon = 2, \bar{\rho})$

$$a_2 = 1 - \bar{\rho}, \quad a_0 = \bar{\rho} \tag{C.35}$$

$$\tilde{a}_2 = 0, \quad \hat{a}_2 = 1 - \bar{\rho} \tag{C.36}$$

$$\mathbb{P}[J_{ij} = J_{ji}] = \bar{\rho} \tag{C.37}$$

In the fully connected situation ($\rho = 1$, $\bar{\rho} = 0$), both formulations describe a

network with only double links, which are all equal if $\eta = 1$ ($\epsilon = 0$).

General Asymmetric Case ($\eta \in (-1, +1)$):
For the general case of asymmetric couplings, a correspondence can be found by matching the correlation parameter η from our method with the expression from the previous formulation and setting the density:

$$\rho = 1 - \bar{\rho}^2 \tag{C.38}$$

In this way we get:

Our Method ($a_1 = 0, \eta \in (-1, +1), \rho$)		**Previous Formulation** ($\epsilon \in (0, 2), \bar{\rho}$)	
$a_2 = \rho, \quad a_0 = 1 - \rho$	(C.39)	$a_2 = 1 - \bar{\rho}^2, \quad a_0 = \bar{\rho}^2$	(C.43)
$\tilde{a}_2 = 0, \quad \hat{a}_2 = a_2 = \rho$	(C.40)	$\tilde{a}_2 = (1 - \bar{\rho})\bar{\rho}, \quad \hat{a}_2 = 1 - \bar{\rho}$	(C.44)
$\mathbb{P}[J_{ij} = J_{ji}] = a_0 = 1 - \rho$	(C.41)	$\mathbb{P}[J_{ij} = J_{ji}] = \bar{\rho}$	(C.45)
$\frac{\mathbb{E}[J_{ij}J_{ji}]}{\mathbb{E}[J_{ij}^2]} = \eta$	(C.42)	$\frac{\mathbb{E}[J_{ij}J_{ji}]}{\mathbb{E}[J_{ij}^2]} = \frac{1 - \epsilon}{1 - \epsilon + \epsilon^2/2}$	(C.46)

A full analogy between the two methods can not be achieved (except for the fully connected or the trivial fully diluted case). However, even if the nature of double links is different ($\tilde{a}_2$, $\hat{a}_2$) for the two methods, the general behavior of the network is the same, indicating that the nature of double links does not affect the main observables.

C.2 Random Map

The random map model represents a disordered system with deterministic dynamics, where the phase space consists of M points each representing a configuration $1 \leq \mathcal{C} \leq M$. If the system is in configuration $\mathcal{C}$ at time t, it will evolve to configuration $T(\mathcal{C})$ at time $t + 1$: the dynamical flow in configuration space is determined by a random mapping which uniquely assigns to each configuration its immediate successor. The randomness stems from the map T itself, and for finite M, the system's properties depend on the specific map chosen, thus lacking self-averaging. Since the map remains constant over time, the dynamics are deterministic, ensuring that each initial configuration eventually converges to a cycle of length $L \geq 1$ after a certain time. In our context, we consider $M = 2^N$. The average number of attractors $\overline{C}$ in a Random Map is computed in [44, 45] as:

$$\overline{C} = \frac{1}{2} \ln M + \mathcal{O}(1) \sim \frac{\ln 2}{2} N \sim 0.35N. \tag{C.1}$$

This result is in agreement with the one obtained in [21, 68, 76] for fully connected and fully asymmetric networks and recovered by our simulations in 6.1.1.

C.3 Results for $a_1 = 0$: Absence of Directed Links

In this section we collect supplementary results obtained in the case of absence of direct links $a_1 = 0$. In Sec. C.3.1 we collect additional distributions that can be useful for a complete understanding of the analysis performed. In Sec. C.3.2 we report phase diagram maps for other observable of Table 5.1.

C.3.1 Additional Distributions

In this section, we present additional distributions of the key quantities defined in Table 5.1 to fully characterize the behavior of the network realized in the constraint of symmetric topology $a_1 = 0$.

Cycle Lengths

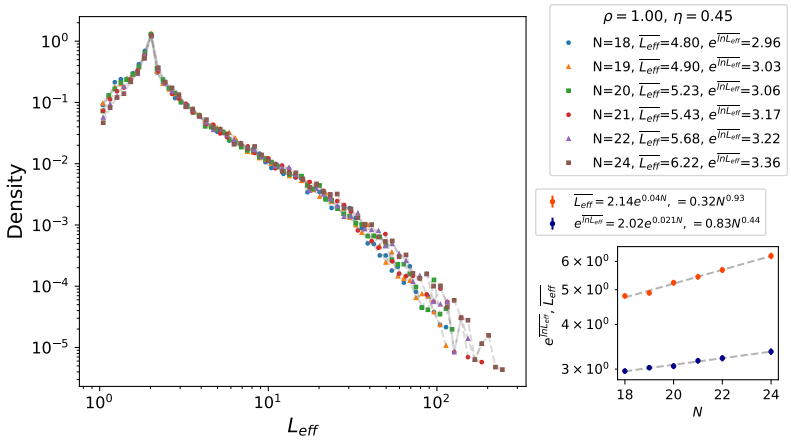
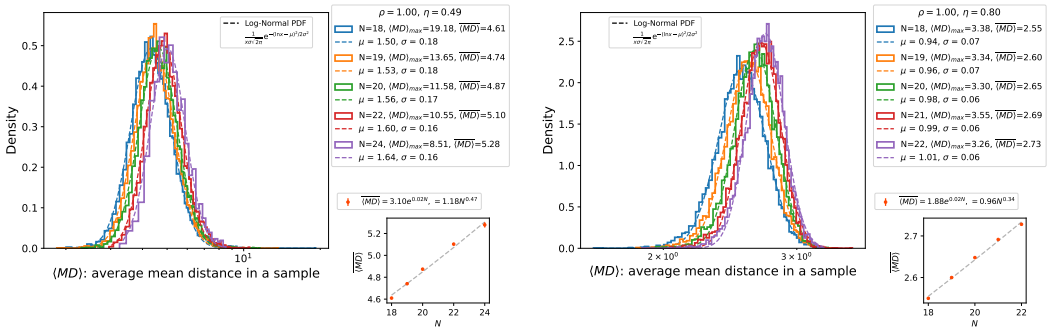


Figure C.1 Weighted cycle length L_{eff} in fully connected asymmetric network. Distribution of the weighted cycle length L_{eff} and relative fit of the mean and typical value as a function of the system size N in the inset. The correlation is $\eta = 0.45$.

Back to the Chapter

Transient Time

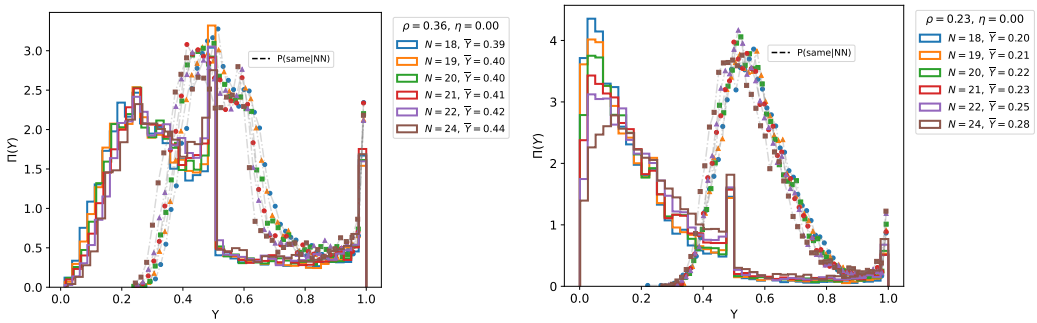


(a) $\eta = 0.49$ **(b) $\eta = 0.80$**

Figure C.2 Transient time distributions across correlation regimes. Distributions of mean distance to attractors $\overline{\langle MD \rangle}$ in fully connected networks ($\rho = 1$) for different correlation values η. The number of steps required to reach attractors decreases systematically with increasing correlation. See Fig. 6.2a for comparison to the asymmetric case $\eta = 0$.

Back to the Chapter

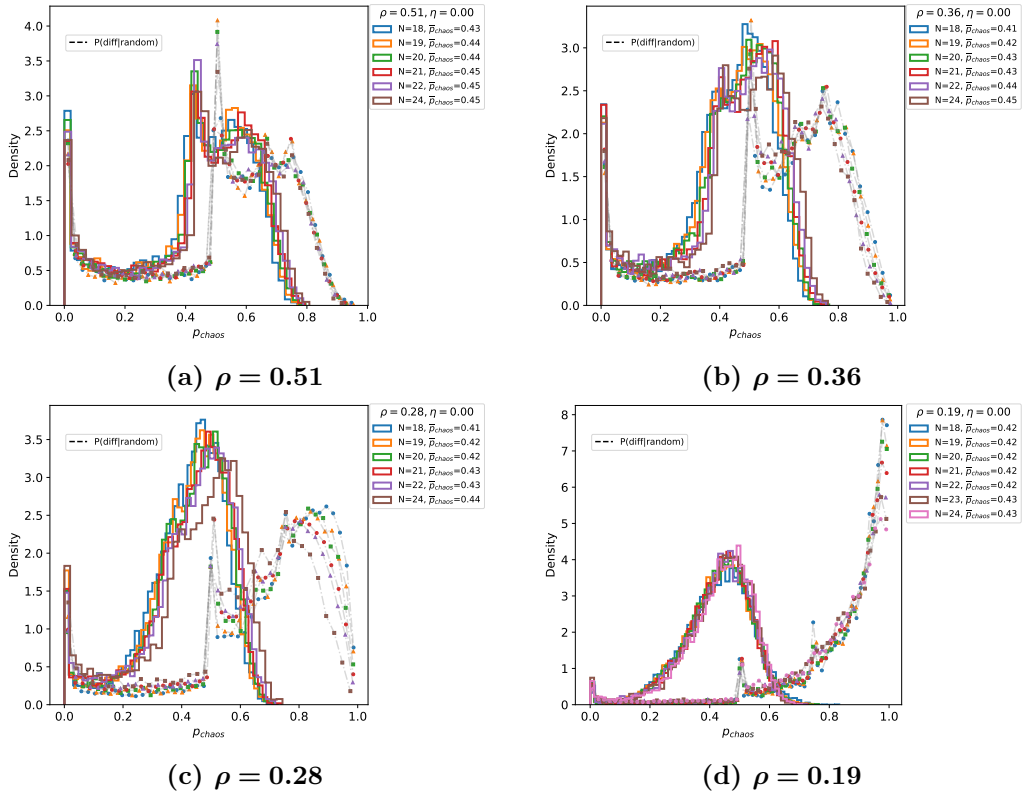
Basin Statistics and Chaoticity p_{chaos} in Fully Asymmetric Networks at Different Dilution



(a) $\rho = 0.36$ **(b) $\rho = 0.23$**

Figure C.3 Basin statistics evolution with dilution in fully asymmetric networks. Distribution $\Pi(Y)$ (solid lines) and probability $P(\text{same}|\text{NN})$ (dashed lines) for varying connection density ρ at $\eta = 0$. The distribution shifts toward $Y \to 0$ with increasing dilution, indicating proliferation of small basins and enhanced dynamical sensitivity.

Back to the Chapter



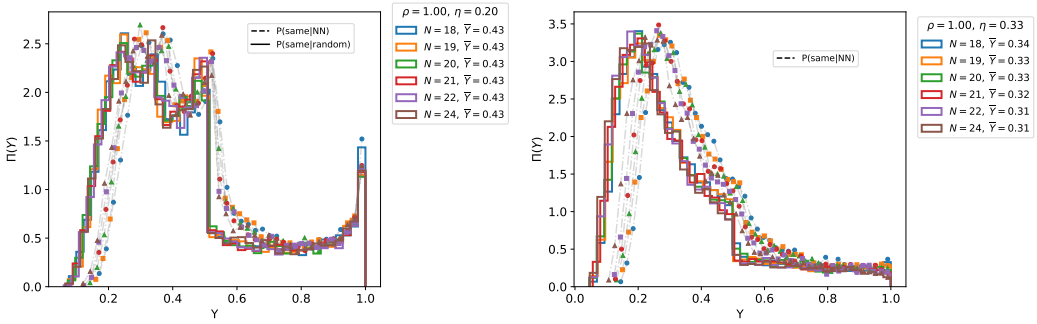
(a) $\rho = 0.51$ (b) $\rho = 0.36$

(c) $\rho = 0.28$ (d) $\rho = 0.19$

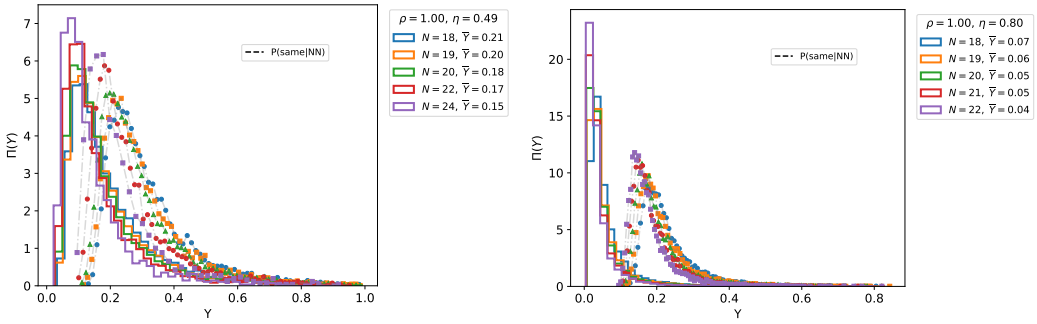
Figure C.4 Chaoticity distributions under increasing dilution. Evolution of p_{chaos} distributions for fully asymmetric networks ($\eta = 0$) across dilution levels ρ. The characteristic singularities at $1 - 1/n$ disappear with increasing dilution. Dashed lines represent $\overline{P(\text{diff}|\text{random})}$ for comparison.

Back to the Chapter

Basin Statistics and Chaoticity p_{chaos} in Fully Connected Networks at Different Correlation

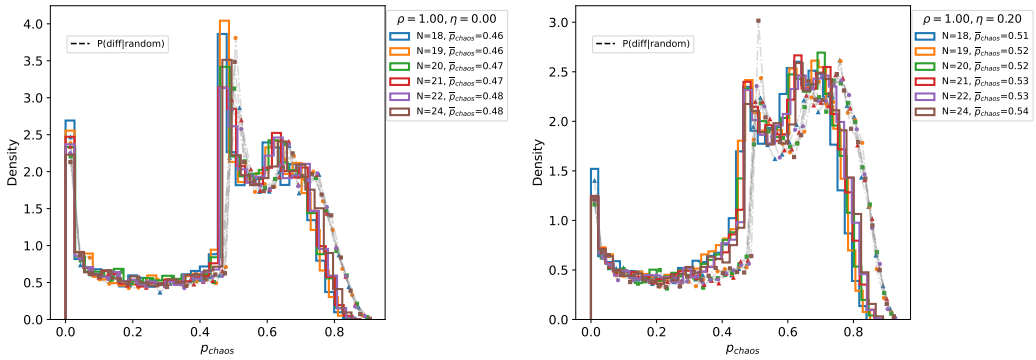


(a) $\eta = 0.20$ (b) $\eta = 0.33$



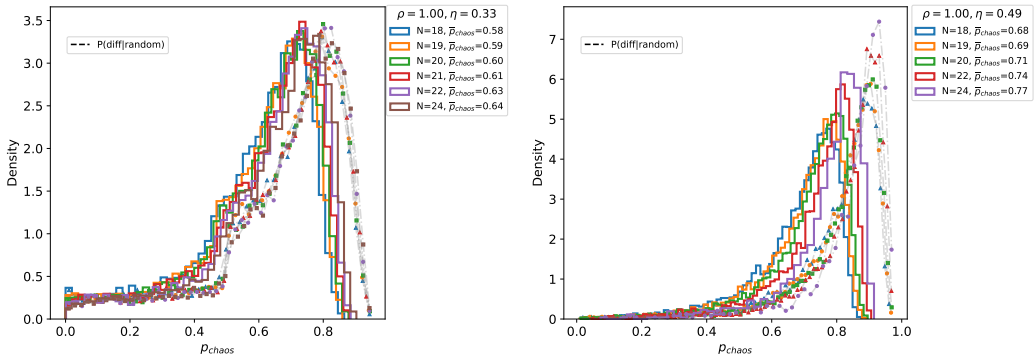
(c) $\eta = 0.49$ **(d) $\eta = 0.80$**

Figure C.5 Evolution of basin statistics with synaptic correlation. Distribution $\Pi(Y)$ (solid lines) and probability $P(\text{same}|\text{NN})_{\text{min}}$ (dotted lines) for varying correlation η in fully connected networks. The singular structure at $Y = 1/n$ disappears for $\eta \gtrsim 0.33$, with distributions becoming concentrated near $Y \to 0$ for $\eta \gtrsim 0.5$.



(a) $\eta = 0$ **(b) $\eta = 0.20$**

Back to the Chapter



(c) $\eta = 0.33$ **(d) $\eta = 0.49$**

Figure C.6 Chaoticity distributions in fully connected networks. Evolution of p_{chaos} distributions across correlation values η. Singularities at $p = 1 - 1/n$ ($n \geq 1$) disappear by $\eta \approx 0.33$ (c), with distributions concentrating around increasing p_{chaos} values for $\eta \gtrsim 0.5$ (d). The overall behavior is complementary to the one in Fig. C.5.

Attractors

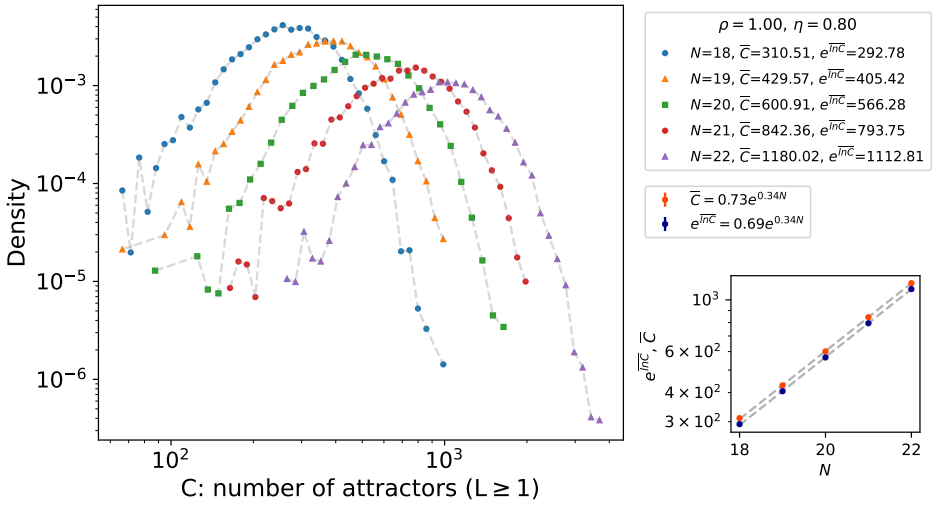
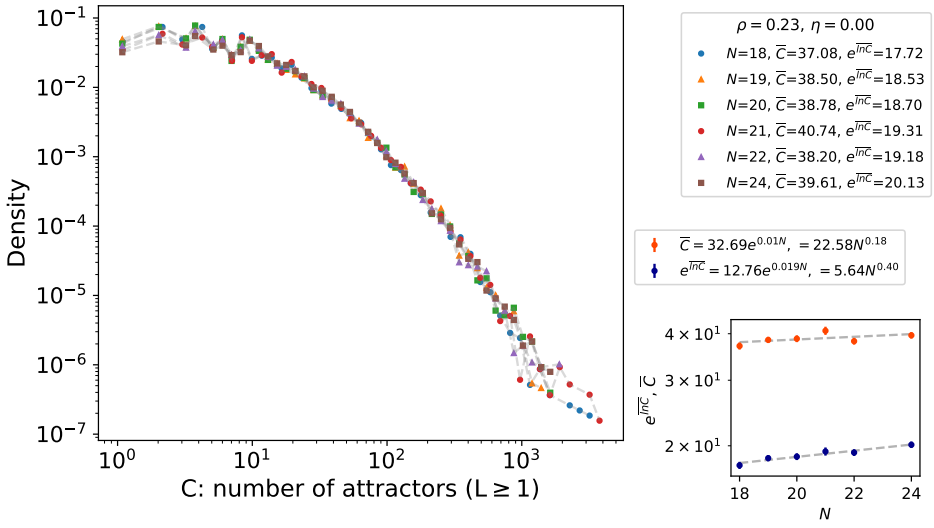
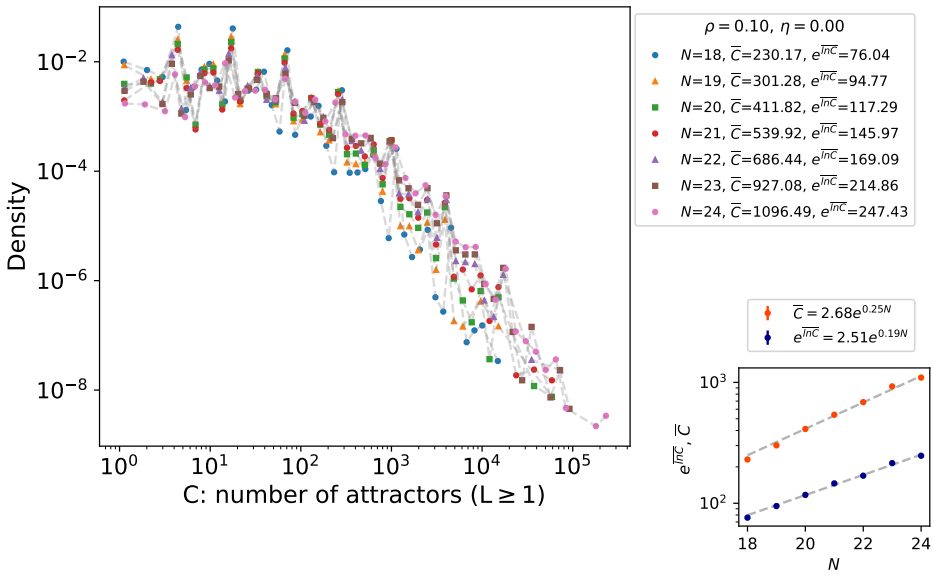


Figure C.7 Attractors in fully connected low asymmetric network. Distribution of the number of attractors C and relative fit of the mean and typical value as a function of the system size N in the inset. The correlation is $\eta = 0.8$.

Back to the Chapter



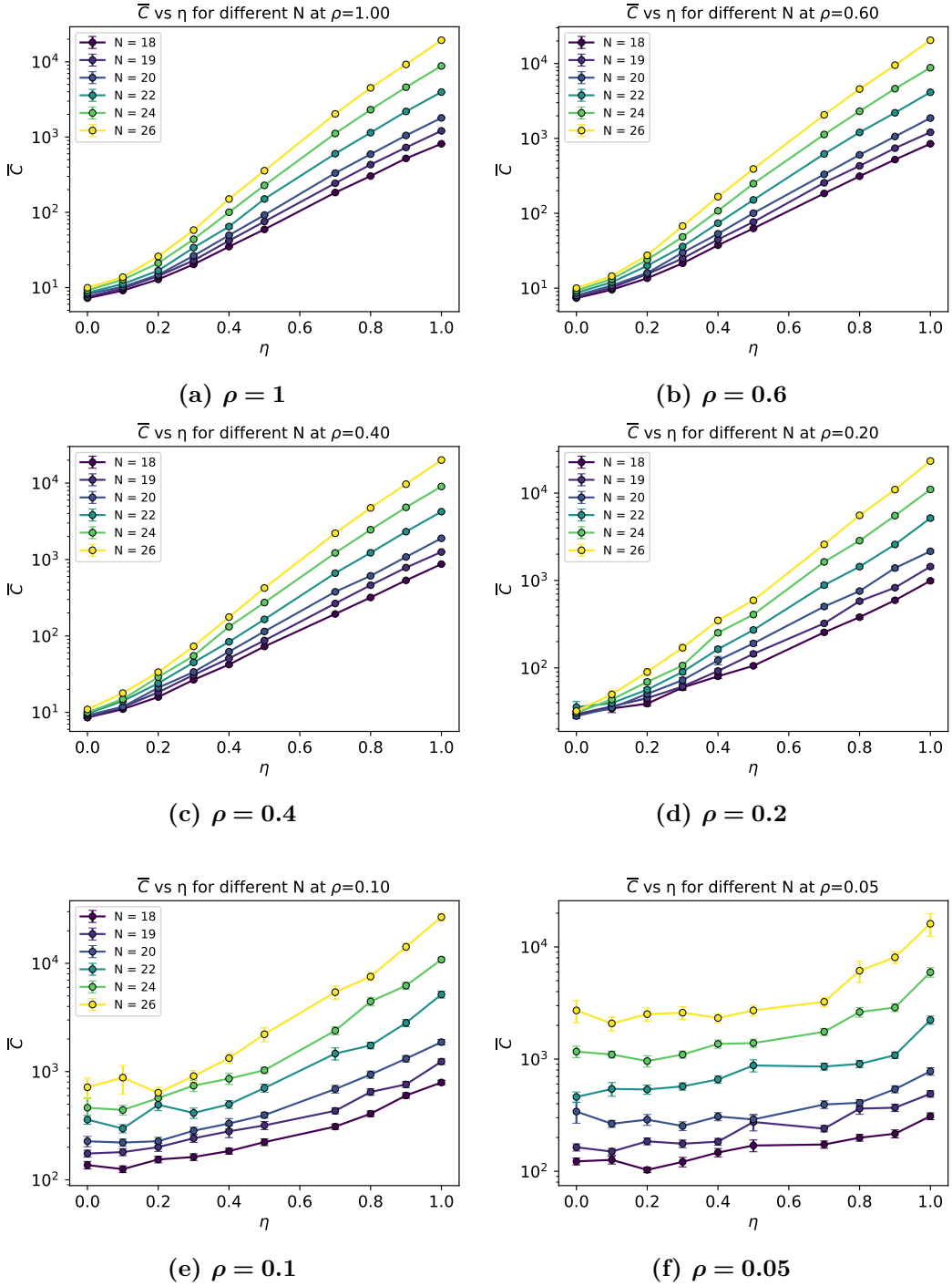
(a) $\rho = 0.23$



(b) $\rho = 0.10$

Figure C.8 Effects of dilution on the attractor distribution. Distribution of the number of attractors C for different values of density ρ and related fit of the mean and typical values. A changing from a power law to an exponential dependence is visible for $\rho \lesssim 0.2$. The exponential coefficient increases as the density decreases.

Back to the Chapter

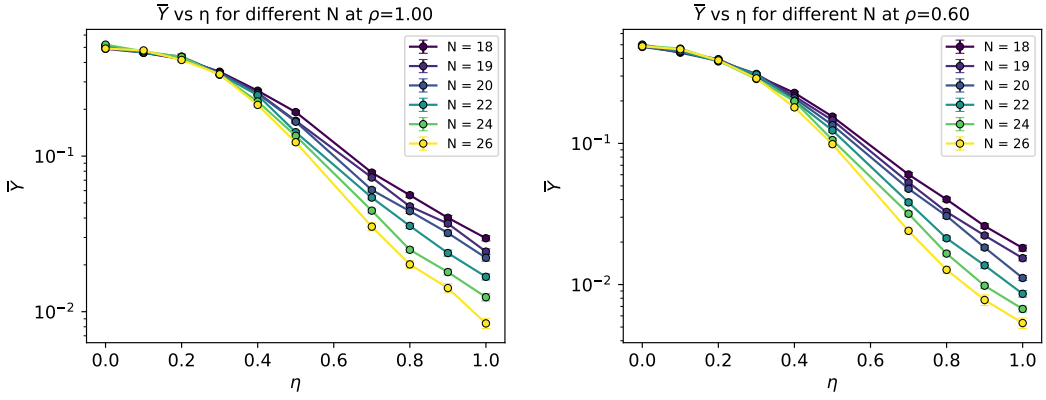


(a) $\boldsymbol{\rho = 1}$ **(b)** $\boldsymbol{\rho = 0.6}$

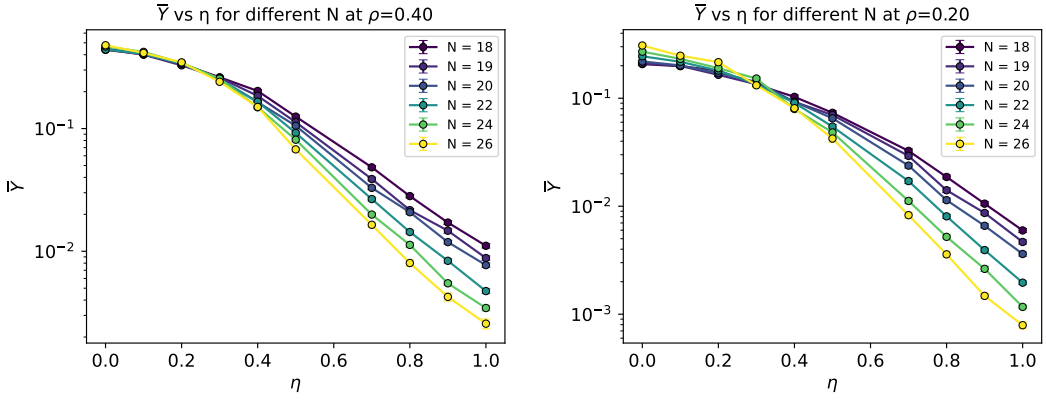
(c) $\boldsymbol{\rho = 0.4}$ **(d)** $\boldsymbol{\rho = 0.2}$

(e) $\boldsymbol{\rho = 0.1}$ **(f)** $\boldsymbol{\rho = 0.05}$

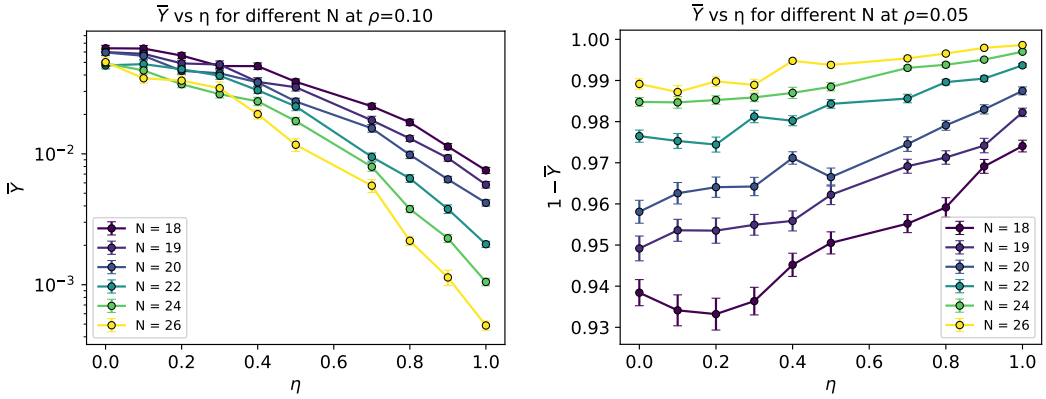
Figure C.9 Dependence of attractor count on synaptic correlation and network density. Mean number of attractors $\overline{C}$ as a function of synaptic correlation η for different connection densities ρ, ranging from fully connected (panel a) to highly diluted scenario (panels e-f). A consistent trend is observed for $\rho \gtrsim 0.2$ while for lower densities we can see that the counts remain largely independent of synaptic correlation η.



(a) $\rho = 1$ **(b)** $\rho = 0.6$

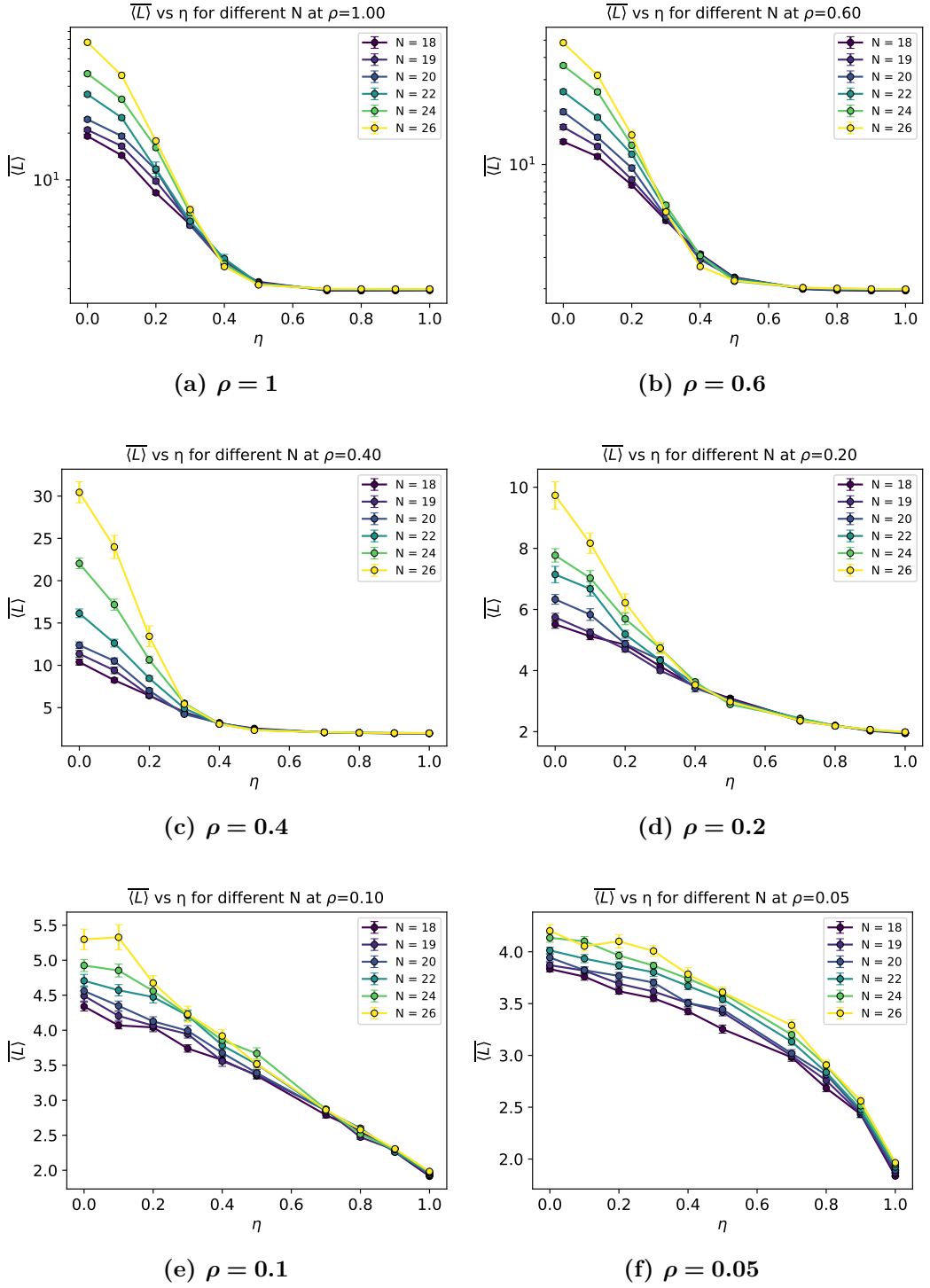


(c) $\rho = 0.4$ **(d)** $\rho = 0.2$



(e) $\rho = 0.1$ **(f)** $\rho = 0.05$

Figure C.10 Evolution of $\overline{Y}$ with synaptic correlation and network density. Trend of the mean value of the distribution $\Pi(Y)$ as a function of the correlation η at different levels of density ρ ranging from fully connected ($\rho = 1$, top-left) to highly diluted ($\rho = 0.05$, bottom-right, where we represents $1 - \overline{Y}$ for better readability). The variable represents of the probability that two randomly peaked configurations end in the same attractor and it also gives a measure of the mean attractor counts weighted through the basin size since $\overline{Y} \equiv 1/N_{\text{eff}}$.

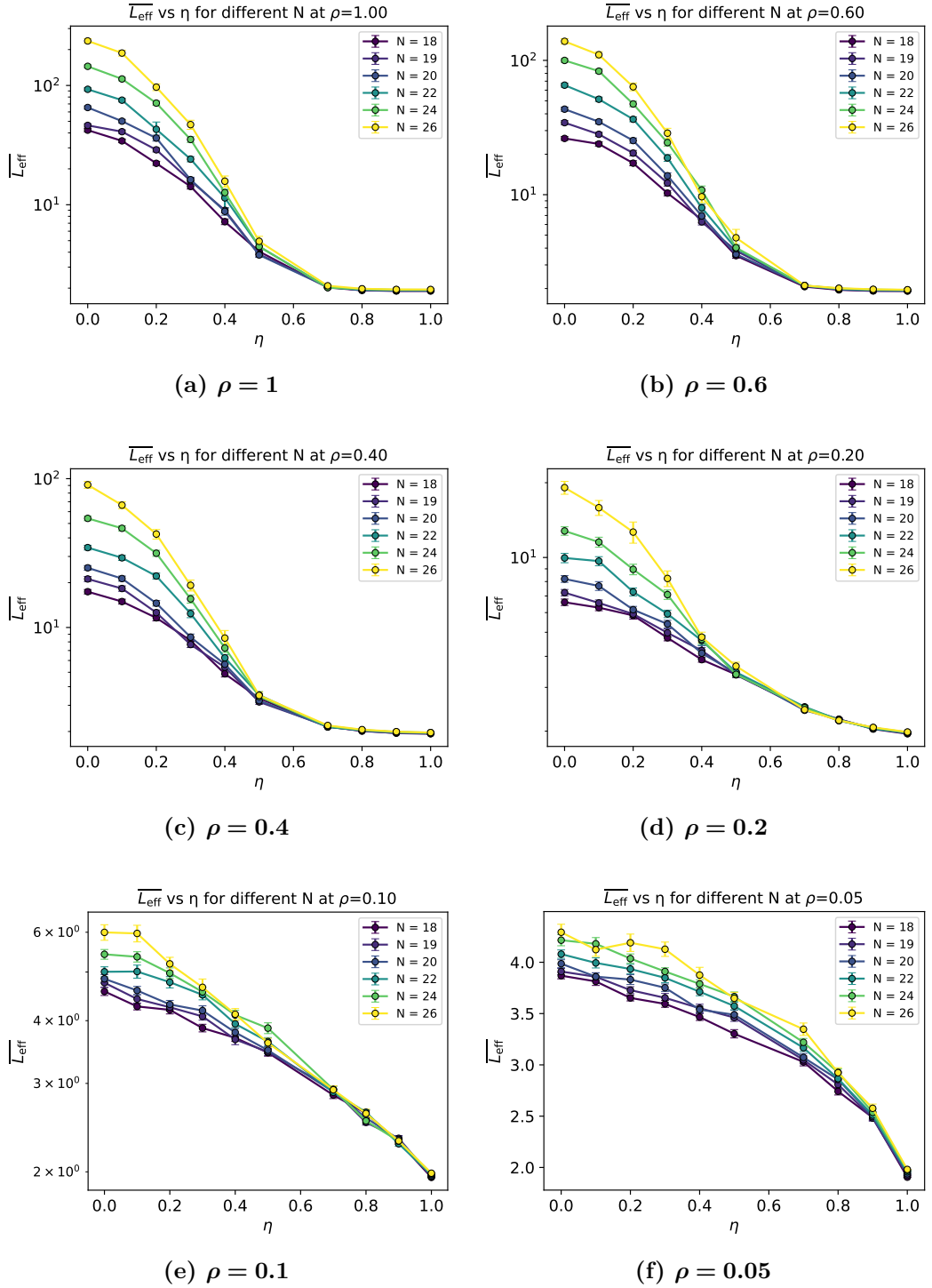


(a) $\rho = 1$ **(b)** $\rho = 0.6$

(c) $\rho = 0.4$ **(d)** $\rho = 0.2$

(e) $\rho = 0.1$ **(f)** $\rho = 0.05$

Figure C.11 Dependence of mean cycle lengths on synaptic correlation and network density. Mean cycle lengths $\overline{\langle L \rangle}$ as a function of synaptic correlation η for different connection densities ρ, ranging from fully connected (a) to highly diluted scenarios (e-f). For $\rho \gtrsim 0.2$, the crossover from exponential cycle growth at low correlation to an ordered regime at high η is clearly visible, where $\overline{\langle L \rangle} \sim 2$ across all system sizes N. In the highly diluted regime ($\rho \sim 0.05$), mean cycle lengths become nearly size-independent, ranging from $\overline{\langle L \rangle} \sim 4$ at low correlation to $\overline{\langle L \rangle} \sim 2$ in the fully symmetric case ($\eta = 1$). This demonstrates how high dilution suppresses chaotic dynamics and stabilizes short, biologically plausible cycle lengths regardless of synaptic correlation. For completeness we report in Fig. C.12 also the weighted mean cycle measures $\overline{L_{\text{eff}}}$.

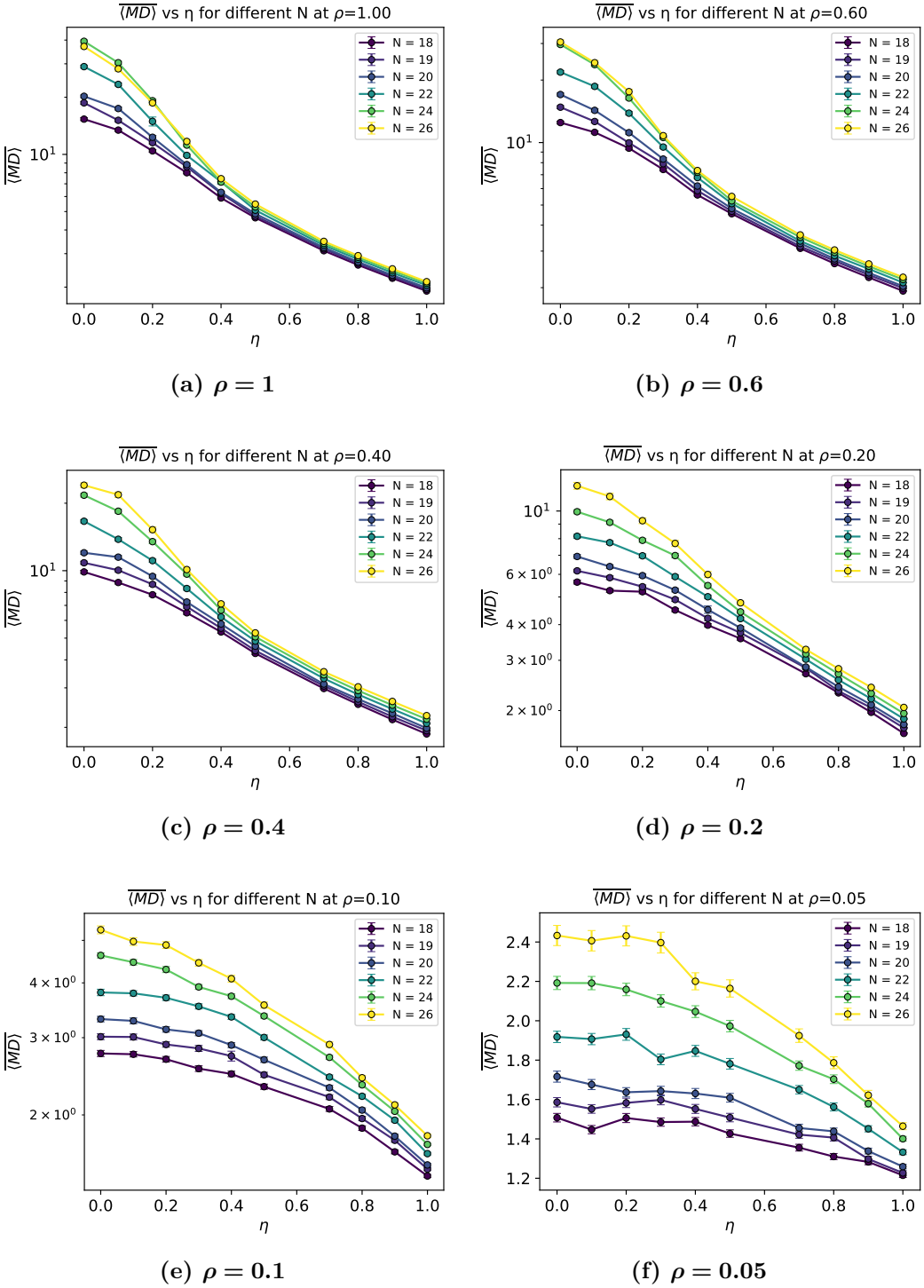


(a) $\rho = 1$ **(b)** $\rho = 0.6$

(c) $\rho = 0.4$ **(d)** $\rho = 0.2$

(e) $\rho = 0.1$ **(f)** $\rho = 0.05$

Figure C.12 Basin-weighted cycle lengths across synaptic correlation and network density. Mean effective cycle lengths $\overline{L_{\text{eff}}}$ as a function of synaptic correlation η for different connection densities ρ, ranging from fully connected (a) to highly diluted scenarios (e-f). Unlike the arithmetic average, $\overline{L_{\text{eff}}}$ weights cycle lengths by their basin sizes. For $\rho \gtrsim 0.2$, the shift to an ordered regime occurs at higher η values compared to the unweighted case, indicating that large basins stabilize shorter cycles earlier. In the highly diluted regime ($\rho \lesssim 0.1$), effective cycle lengths remain short and size-independent as in the previous case.



(a) $\rho = 1$ **(b) $\rho = 0.6$**

(c) $\rho = 0.4$ **(d) $\rho = 0.2$**

(e) $\rho = 0.1$ **(f) $\rho = 0.05$**

Figure C.13 Dependence of mean transient time on synaptic correlation and network density. Average distance to attractors $\overline{\langle \mathrm{MD} \rangle}$ as a function of the synaptic correlation η and the network density ρ from the dense case $\rho = 1$ (panel a) to the highly diluted regimes (e-f). The plots reveal a gradual shift from exponentially long transients in dense, uncorrelated networks to rapid convergence in the correlated cases. For $\rho \gtrsim 0.2$ and low η, the system exhibits chaotic-like behavior with extended transients, while for high synaptic correlation ($\eta \to 1$) we can see fast convergence even in dense networks. In the highly diluted regime ($\rho \lesssim 0.1$), convergence happens in few time steps across all values of η, demonstrating that network sparsity alone allows to reach the attractor in few steps.

Back to the Chapter

C.3.2 Phase Diagram

To fully characterize the phase diagram and the general behavior of key observables, we present in this section colormaps for additional quantities listed in Table 5.1. Fig. C.14 shows the scaling of the transient time, quantified by the mean distance to attractors $\overline{\langle MD \rangle}$.

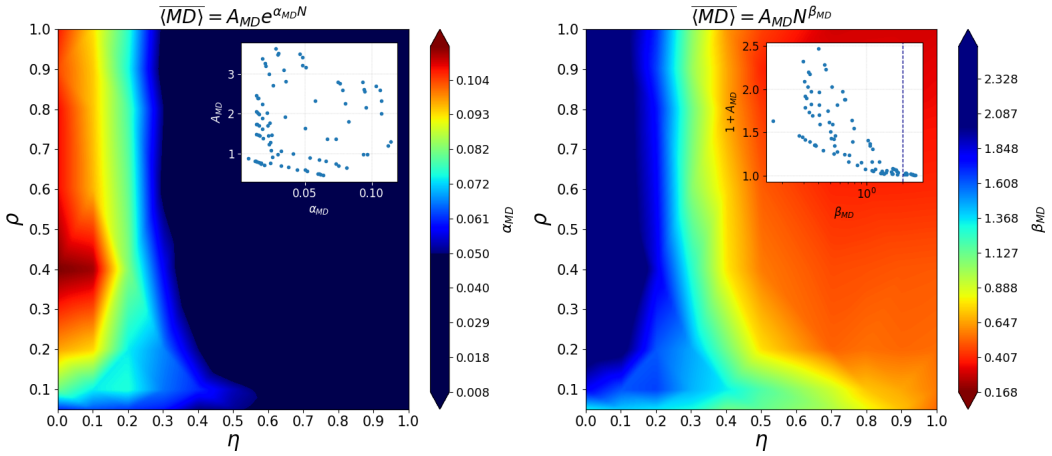


Figure C.14 Scaling behavior of mean distance to attractors $\overline{\langle MD \rangle}$. *On the left:* Phase diagram of the mean distance from the attractors with a colorbar defined by the exponent α_{MD}. In dark blue the region in which the exponential fit is considered no more valid. *On the right:* Phase diagram defined by the a power law fit. The threshold value for the validity of the fit is $\beta_{MD} = 2$: in blue the region in which the power law is no more acceptable. Insets show relationships between fitting parameters.

Back to the Chapter

For the exponential fit, parameters A_{MD} and α_{MD} show no clear correlation, suggesting independent variation across the (ρ, η) parameter space. This may indicate that the exponential model provides unstable fits in regions where α_{MD} is small and the growth becomes effectively linear. In contrast, the power-law fit exhibits a more structured parameter relationship and we have observed a clear inverse correlation between A_{MD} and β_{MD}, consistent with the expected trade-off between prefactor and exponent in power-law scaling. This trend suggests a more interpretable parametrization under the power-law model, particularly in the region where the fit is valid ($\beta_{\mathrm{MD}} \leq 2$).

Figures C.15 and C.16 present the effective quantities $\overline{N_{\mathrm{eff}}}$ and $\overline{L_{\mathrm{eff}}}$, where averages are weighted by basin sizes as defined in the main text. These effective measures provide insight into the attractor landscape accounting for basin dominance.

The phase diagrams for effective quantities show scaling behaviors consistent with those observed for their unweighted counterparts in Fig. 6.4 and 6.5. The power law scaling drops sharply in the same parameter region where the exponential fit dominates, suggesting a clear crossover between the two scaling regimes. The corresponding inset shows a negative correlation between amplitudes and growth rates, reinforcing the complementary nature of these scaling descriptions across the parameter space.

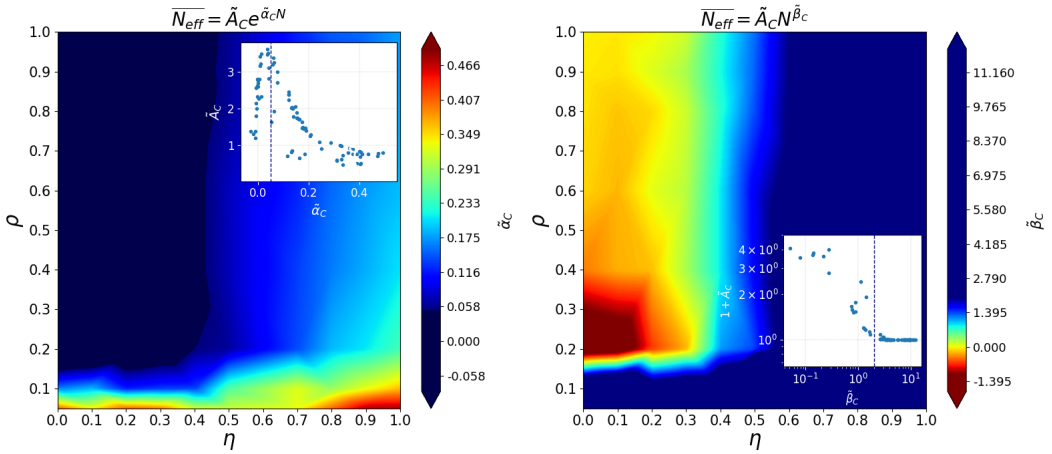


Figure C.15 Scaling behavior of effective attractor count $\overline{N_{\text{eff}}}$. *On the left:* Phase diagram of the mean effective number of attractors $\overline{N_{eff}}$ with a colorbar defined by the exponent $\tilde{\alpha}_C$. In dark blue the region in which the exponential fit is considered no more valid. *On the right:* Phase diagram defined by the a power law fit. The threshold value for the validity of the fit is $\tilde{\beta}_C = 2$: in blue the region in which the power law is no more acceptable.

Back to the Chapter

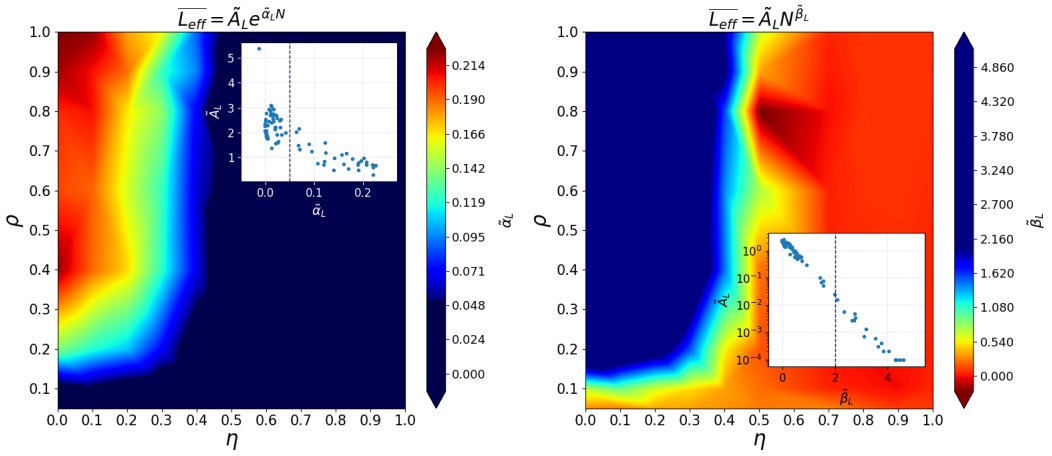


Figure C.16 Scaling behavior of effective cycle length $\overline{L_{\text{eff}}}$. *On the left:* Phase diagram of the mean effective cycle lengths $\overline{L_{eff}}$ with a colorbar defined by the exponent $\tilde{\alpha}_L$. In dark blue the region in which the exponential fit is considered no more valid. *On the right:* Phase diagram defined by the a power law fit. The threshold value for the validity of the fit is $\tilde{\beta}_L = 2$: in blue the region in which the power law is no more acceptable.

As a final description of the scaling behavior of attractors and cycle lengths, we present the colormaps of Figs. 6.4-6.5 using the nomenclature adopted in [54, 87, 76] to improve the readability of this work and contextualize our findings within this existing formulations. This parametrization facilitates direct comparison with prior results on asymmetric, diluted and intrinsically topologically symmetric networks. In order

to compare our results directly with previous studies, we adapt our parametrization to this special case as described in Sec. C.1.

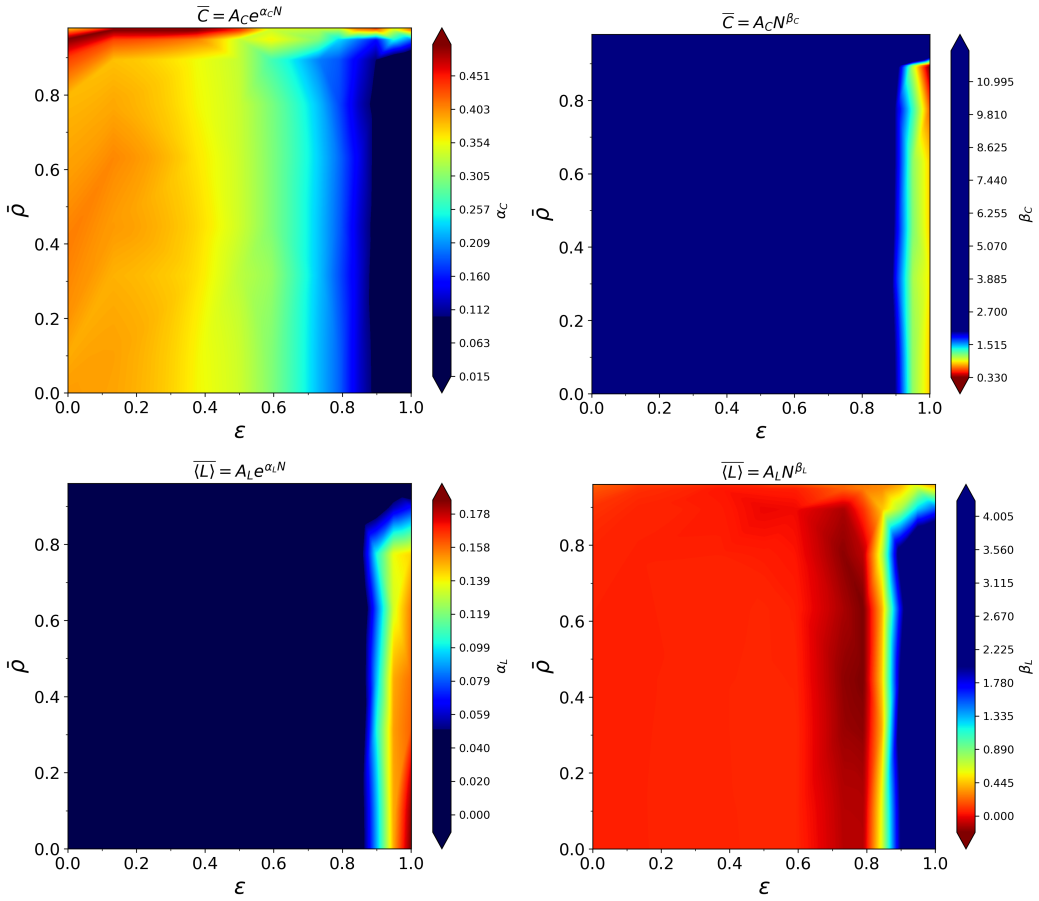
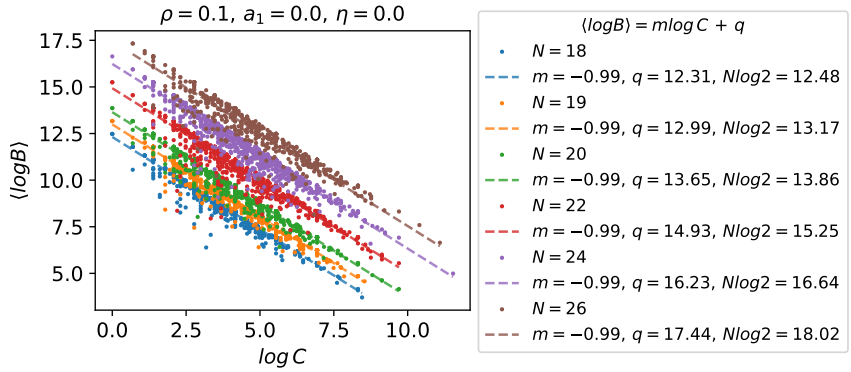


Figure C.17 Dynamical Regimes in Asymmetric Diluted Networks
Scaling exponents for attractor count and cycle length in the $(\epsilon, \bar{\rho})$ parameter space, following the nomenclature of [54, 87, 76]. **Top row:** Mean number of attractors $\overline{C}$ with (a) exponential scaling $\overline{C} \sim e^{\alpha_C N}$ and (b) power-law scaling $\overline{C} \sim N^{\beta_C}$. **Bottom row:** Mean cycle length $\overline{\langle L \rangle}$ with (c) exponential scaling $\overline{\langle L \rangle} \sim e^{\alpha_L N}$ and (d) power-law scaling $\overline{\langle L \rangle} \sim N^{\beta_L}$. The complementary structure between exponential and power-law regimes reveals distinct dynamical phases, with exponential attractor proliferation dominating at high dilution ($\bar{\rho} \to 1$) and power-law behavior emerging in denser uncorrelated regimes.

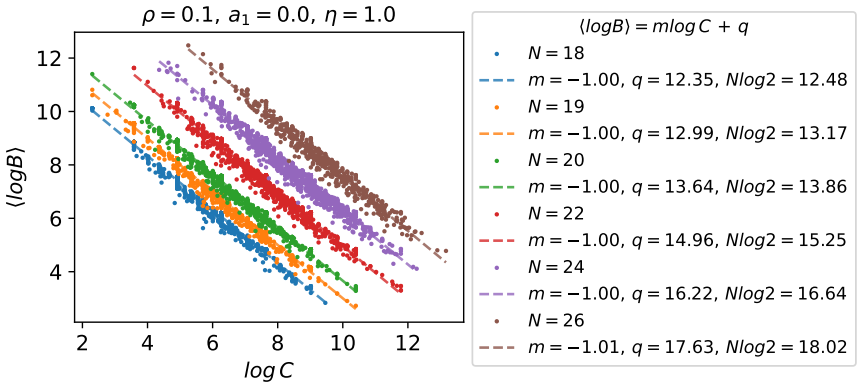
Back to the Chapter

C.3.3 High Diluted Regime



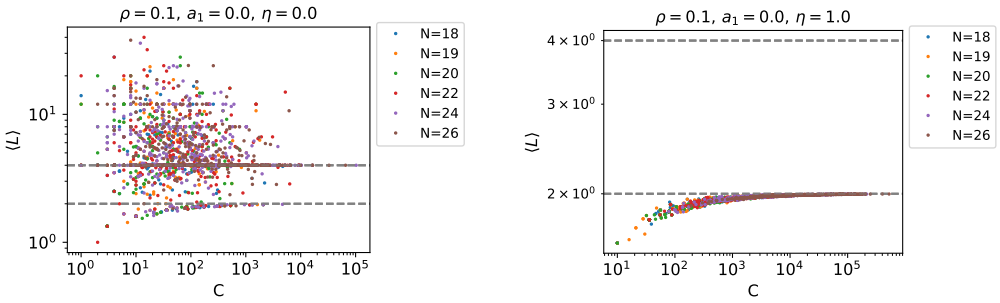
(a) Fully asymmetric case $\boldsymbol{\eta = 0}$

Figure C.18 Basin size distributions at high dilution ($\rho = 0.1$). Relationship between $\langle \log B \rangle$ and $\log C(\alpha)$ for networks with symmetric topology ($a_1 = 0$).



(b) Fully symmetric case $\boldsymbol{\eta = 1}$

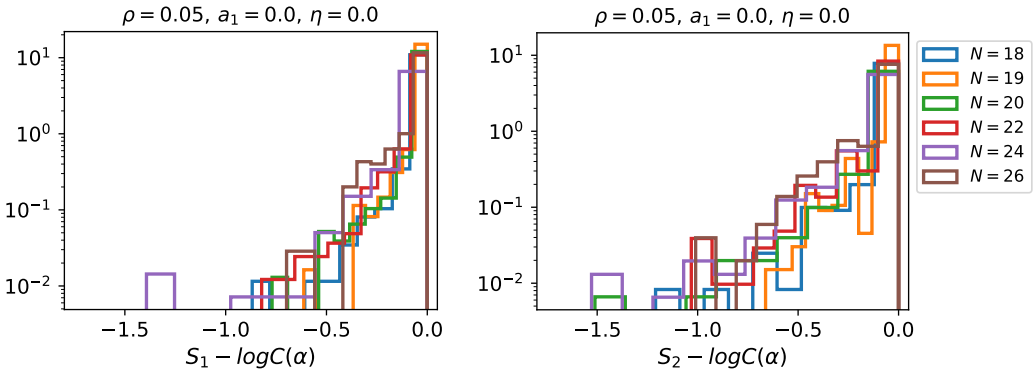
Figure C.18 Basin size distributions at high dilution ($\rho = 0.1$). Both correlation regimes show linear trend with increased fluctuations compared to $\rho = 0.05$ (Fig. 6.8).



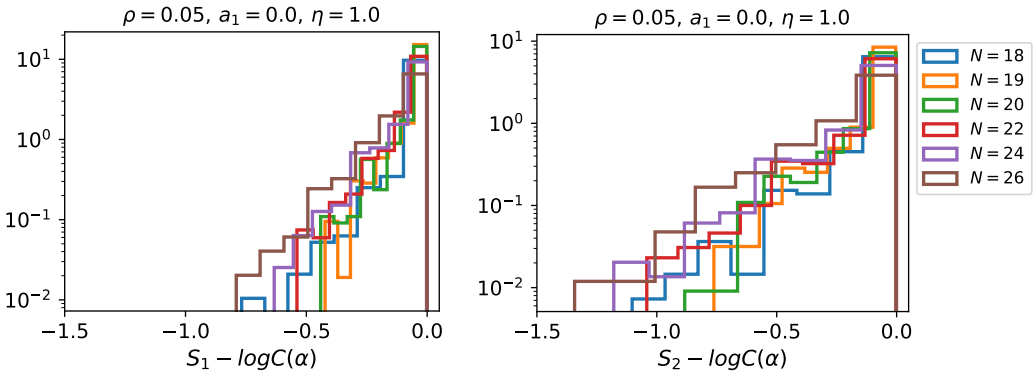
(a) $\boldsymbol{\eta = 0}$ **(b)** $\boldsymbol{\eta = 1}$

Figure C.19 Cycle length stability in highly diluted networks. Mean cycle length $\langle L \rangle$ versus attractor count $C(\alpha)$ for different correlation values η in the high dilution regime ($\rho = 0.1$, $a_1 = 0$). Cycle lengths remain short and stable across all correlation regimes, with the symmetric case ($\eta = 1$) converging exclusively to 2-cycles.

Back to the Chapter



(a) Fully asymmetric case ($\boldsymbol{\eta = 0}$)



(b) Fully symmetric case ($\boldsymbol{\eta = 1}$)

Figure C.20 Entropy fluctuations in highly diluted networks ($\rho = 0.05$). Histograms of fluctuations of Shannon entropy S_1 (left) and collision entropy S_2 (right) from the uniform distribution baseline $\log C(\alpha)$. (**a**) Fully asymmetric regime ($\eta = 0$) and (**b**) fully symmetric regime ($\eta = 1$). The sharp peaks around zero for both entropy measures and all system sizes N indicate nearly uniform basin distributions, demonstrating that uniform state space partitioning occurs independently of both system size and synaptic correlation.

Back to the Chapter

Bibliography

[1] Agliari, E. *Modelli di Reti Neurali* (2020).

[2] Agliari, E., Albanese, L., Alemanno, F., Alessandrelli, A., Barra, A., Giannotti, F., Lotito, D., and Pedreschi, D. Dense hebbian neural networks: A replica symmetric picture of unsupervised learning. *Physica A: Statistical Mechanics and its Applications*, **627** (2023), 129143. Available from: https://www.sciencedirect.com/science/article/pii/S0378437123006982, doi:https://doi.org/10.1016/j.physa.2023.129143.

[3] Agliari, E., Alemanno, F., Barra, A., and Fachechi, A. Dreaming neural networks: rigorous results. *Journal of Statistical Mechanics: Theory and Experiment*, **2019** (2019), 083503. doi:10.1088/1742-5468/ab371d.

[4] Agliari, E., Alemanno, F., Barra, A., and Fachechi, A. Generalized guerra's interpolation schemes for dense associative neural networks. *Neural Networks*, **128** (2020), 254. Available from: https://www.sciencedirect.com/science/article/pii/S0893608020301714, doi:https://doi.org/10.1016/j.neunet.2020.05.009.

[5] Agliari, E., Aquaro, M., Alemanno, F., and Fachechi, A. Regularization, early-stopping and dreaming: a hopfield-like setup to address generalization and overfitting. *arXiv preprint arXiv:2308.01421*, (2023).

[6] Agoritsas, E., Catania, G., Decelle, A., and Seoane, B. Explaining the effects of non-convergent mcmc in the training of energy-based models. In *International Conference on Machine Learning*, pp. 322–336. PMLR (2023).

[7] Aguirre-López, F., Pastore, M., and Franz, S. Satisfiability transition in asymmetric neural networks. *Journal of Physics A: Mathematical and Theoretical*, **55** (2022), 305001. Available from: https://dx.doi.org/10.1088/1751-8121/ac79e5, doi:10.1088/1751-8121/ac79e5.

[8] Albanese, L., Alessandrelli, A., Annibale, A., and Barra, A. About the de almeida–thouless line in neural networks. *Physica A: Statistical Mechanics and its Applications*, **633** (2024), 129372. Available from: https://www.sciencedirect.com/science/article/pii/S0378437123009275, doi:https://doi.org/10.1016/j.physa.2023.129372.

[9] Albeverio, S., Tirozzi, B., and Zegarlinski, B. Rigorous results for the free energy in the Hopfield model. *Communications in Mathematical Physics*, **150** (1992), 337. doi:10.1007/BF02096664.

[10] Alemanno, F., Aquaro, M., Kanter, I., Barra, A., and Agliari, E. Supervised hebbian learning. *Europhysics Letters*, **141** (2023), 11001.

[11] Amit, D. J. *Modeling brain function: The world of attractor neural networks*. Cambridge University Press (1989).

[12] Amit, D. J., Gutfreund, H., and Sompolinsky, H. Spin-glass models of neural networks. *Phys. Rev. A*, **32** (1985), 1007. Available from: https://link.aps.org/doi/10.1103/PhysRevA.32.1007, doi:10.1103/PhysRevA.32.1007.

[13] Amit, D. J., Gutfreund, H., and Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, **55** (1985), 1530.

[14] Amit, D. J., Gutfreund, H., and Sompolinsky, H. Information storage in neural networks with low levels of activity. *Physical Review A*, **35** (1987), 2293.

[15] Amit, D. J., Gutfreund, H., and Sompolinsky, H. Statistical mechanics of neural networks near saturation. *Annals of Physics*, **173** (1987), 30.

[16] Aubin, B., Perkins, W., and Zdeborová, L. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, **52** (2019), 294003. Available from: https://doi.org/10.1088/1751-8121/ab227a, doi:10.1088/1751-8121/ab227a.

[17] Badalotti, D., Baldassi, C., Mézard, M., Scardecchia, M., and Zecchina, R. Dynamical learning in deep asymmetric recurrent neural networks. (2025). Available from: https://arxiv.org/abs/2509.05041, arXiv:2509.05041.

[18] Baldassi, C., Gerace, F., Saglietti, L., and Zecchina, R. From inverse problems to learning: a statistical mechanics approach. **955** (2018), 012001.

[19] Barra, A., Genovese, G., Guerra, F., and Tantari, D. How glassy are neural networks? *Journal of Statistical Mechanics: Theory and Experiment*, **2012** (2012), P07009. Available from: https://dx.doi.org/10.1088/1742-5468/2012/07/P07009, doi:10.1088/1742-5468/2012/07/P07009.

[20] Bastolla, U. and Parisi, G. Attractors in fully asymmetric neural networks. *Journal of Physics A: Mathematical and General*, **30** (1997), 5613. Available from: https://dx.doi.org/10.1088/0305-4470/30/16/007, doi:10.1088/0305-4470/30/16/007.

[21] Bastolla, U. and Parisi, G. Relaxation, closing probabilities and transition from oscillatory to chaotic attractors in asymmetric neural networks. *Journal of Physics A: Mathematical and General*, **31** (1998), 4583. Available from: https://dx.doi.org/10.1088/0305-4470/31/20/003, doi:10.1088/0305-4470/31/20/003.

[22] Belyaev, M. and Velichko, A. Classification of handwritten digits using the hopfield network. **862** (2020), 052048.

[23] Benedetti, M., Carillo, L., Marinari, E., and Mézard, M. Eigenvector dreaming. *Journal of Statistical Mechanics: Theory and Experiment*, **2024** (2024), 013302. doi:10.1088/1742-5468/ad138e.

[24] Benedetti, M. and Ventura, E. Training neural networks with structured noise improves classification and generalization. *arXiv preprint arXiv:2302.13417*, (2023).

[25] Benedetti, M., Ventura, E., Marinari, E., Ruocco, G., and Zamponi, F. Supervised perceptron learning vs unsupervised hebbian unlearning: Approaching optimal memory retrieval in hopfield-like networks. *J Chem Phys*, **156** (2022), 104107. doi:10.1063/5.0084219.

[26] Bialek, W. *Biophysics: Searching for Principles*. Princeton University Press (2012). ISBN 9780691138916. Available from: https://books.google.it/books?id=5In_FKA2rmUC.

[27] Braunstein, A., Ramezanpour, A., Zecchina, R., and Zhang, P. Inference and learning in sparse systems with multiple states. *Phys. Rev. E*, **83** (2011), 056114. Available from: https://link.aps.org/doi/10.1103/PhysRevE.83.056114, doi:10.1103/PhysRevE.83.056114.

[28] Bray, A. J. and Moore, M. A. Metastable states in spin glasses. *Journal of Physics C: Solid State Physics*, **13** (1980), L469. Available from: https://dx.doi.org/10.1088/0022-3719/13/19/002, doi:10.1088/0022-3719/13/19/002.

[29] Brunel, N. Is cortical connectivity optimized for storing information? *Nature Neuroscience*, **19** (2016), 749. doi:10.1038/nn.4286.

[30] Carreira-Perpiñán, M. A. and Hinton, G. On contrastive divergence learning. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics* (edited by R. G. Cowell and Z. Ghahramani), vol. R5 of *Proceedings of Machine Learning Research*, pp. 33–40. PMLR (2005). Reissued by PMLR on 30 March 2021. Available from: https://proceedings.mlr.press/r5/carreira-perpinan05a.html.

[31] Charbonneau, P., Marinari, E., Parisi, G., Mézard, M., Ricci-Tersenghi, F., Sicuro, G., and Zamponi, F. *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking After 40 Years*. G - Reference,Information and Interdisciplinary Subjects Series. World Scientific (2023). ISBN 9789811273919. Available from: https://books.google.it/books?id=pYK5zwEACAAJ.

[32] Christos, G. A. Investigation of the crick-mitchison reverse-learning dream sleep hypothesis in a dynamical setting. *Neural Networks*, **9** (1996), 427. Available from: https://www.sciencedirect.com/science/article/pii/0893608095000720, doi:https://doi.org/10.1016/0893-6080(95)00072-0.

[33] Coolen, A. C. C., Kühn, R., and Sollich, P. *Theory of Neural Information Processing Systems*. Oxford University Press (2005).

[34] Crick, F. and Mitchison, G. The function of dream sleep. *Nature*, **304** (1983), 111. doi:10.1038/304111a0.

[35] Crisanti, A., Amit, D. J., and Gutfreund, H. Saturation level of the hopfield model for neural network. *Europhysics Letters*, **2** (1986), 337. Available from: https://dx.doi.org/10.1209/0295-5075/2/4/012, doi:10.1209/0295-5075/2/4/012.

[36] Crisanti, A., Falcioni, M., and Vulpiani, A. Transition from regular to complex behaviour in a discrete deterministic asymmetric neural network model. *Journal of Physics A: Mathematical and General*, **26** (1993), 3441. Available from: https://dx.doi.org/10.1088/0305-4470/26/14/011, doi:10.1088/0305-4470/26/14/011.

[37] Crisanti, A. and Sompolinsky, H. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Phys. Rev. A*, **36** (1987), 4922. Available from: https://link.aps.org/doi/10.1103/PhysRevA.36.4922, doi:10.1103/PhysRevA.36.4922.

[38] Crisanti, A. and Sompolinsky, H. Dynamics of spin systems with randomly asymmetric bonds: Ising spins and glauber dynamics. *Phys. Rev. A*, **37** (1988), 4865. Available from: https://link.aps.org/doi/10.1103/PhysRevA.37.4865, doi:10.1103/PhysRevA.37.4865.

[39] de Almeida, J. R. L. and Thouless, D. J. Stability of the sherrington-kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, **11** (1978), 983. Available from: https://dx.doi.org/10.1088/0305-4470/11/5/028, doi:10.1088/0305-4470/11/5/028.

[40] Decelle, A. and Furtlehner, C. Restricted boltzmann machine: Recent advances and mean-field theory. *Chinese Physics B*, **30** (2021), 040202.

[41] Decelle, A., Furtlehner, C., and Seoane, B. Equilibrium and non-equilibrium regimes in the learning of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, **34** (2021), 5345.

[42] Demircigil, M., Heusel, J., Löwe, M., Upgang, S., and Vermet, F. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, **168** (2017), 288.

[43] Der, R., Dotsenko, V. S., and Tirozzi, B. Modified pseudo-inverse neural networks storing correlated patterns. *Journal of Physics A: Mathematical and General*, **25** (1992), 2843.

[44] Derrida, B. and Flyvbjerg, H. The random map model: a disordered model with deterministic dynamics. *http://dx.doi.org/10.1051/jphys:01987004806097100*, **48** (1987). doi:10.1051/jphys:01987004806097100.

[45] Derrida, B. and Flyvbjerg, H. Statistical properties of randomly broken objects and of multivalley structures in disordered systems. *Journal of Physics*

A: Mathematical and General, **20** (1987), 5273. Available from: https://dx.doi.org/10.1088/0305-4470/20/15/039, doi:10.1088/0305-4470/20/15/039.

[46] Derrida, B., Gardner, E., and Zippelius, A. An exactly solvable asymmetric neural network model. *Europhysics Letters*, **4** (1987), 167. Available from: https://doi.org/10.1209/0295-5075/4/2/007, doi:10.1209/0295-5075/4/2/007.

[47] Dotsenko, V. and Tirozzi, B. Replica-symmetry breaking in neural networks. *Physica A: Statistical Mechanics and its Applications*, **185** (1992), 385. Available from: https://www.sciencedirect.com/science/article/pii/037843719290479A, doi:https://doi.org/10.1016/0378-4371(92)90479-A.

[48] Dotsenko, V. S., Yarunin, N. D., and Dorotheyev, E. A. Statistical mechanics of hopfield-like neural networks with modified interactions. *Journal of Physics A: Mathematical and General*, **24** (1991), 2419.

[49] Edwards, S. F. and Anderson, P. W. Theory of spin glasses. *Journal of Physics F: Metal Physics*, **5** (1975), 965. Available from: https://dx.doi.org/10.1088/0305-4608/5/5/017, doi:10.1088/0305-4608/5/5/017.

[50] Eissfeller, H. and Opper, M. New method for studying the dynamics of disordered spin systems without finite-size effects. *Phys. Rev. Lett*, **68** (1992), 2094. Available from: https://link.aps.org/doi/10.1103/PhysRevLett.68.2094, doi:10.1103/PhysRevLett.68.2094.

[51] Eissfeller, H. and Opper, M. Mean-field monte carlo approach to the sherrington-kirkpatrick model with asymmetric couplings. *Phys. Rev. E*, **50** (1994), 709. Available from: https://link.aps.org/doi/10.1103/PhysRevE.50.709, doi:10.1103/PhysRevE.50.709.

[52] Engel, A. and Van den Broeck, C. *Statistical Mechanics of Learning*. Cambridge University Press, Cambridge (2001).

[53] Fachechi, A., Agliari, E., and Barra, A. Dreaming neural networks: forgetting spurious memories and reinforcing pure ones. *Neural Networks*, **112** (2019), 24.

[54] Folli, V., Gosti, G., Leonetti, M., and Ruocco, G. Effect of dilution in asymmetric recurrent neural networks. *Neural Networks*, **104** (2018), 50. Available from: https://www.sciencedirect.com/science/article/pii/S0893608018301230, doi:https://doi.org/10.1016/j.neunet.2018.04.003.

[55] Fontanari, J. F. and Theumann, W. On the storage of correlated patterns in hopfield's model. *Journal de Physique*, **51** (1990), 375.

[56] Fournier, S. J., Pacco, A., Ros, V., and Urbani, P. Non-reciprocal interactions and high-dimensional chaos: comparing dynamics and statistics of equilibria in a solvable model. (2025). Available from: https://arxiv.org/abs/2503.20908, arXiv:2503.20908.

[57] Gardner, E. Structure of metastable states in the hopfield model. *Journal of Physics A: Mathematical and General*, **19** (1986), L1047. Available from: https://doi.org/10.1088/0305-4470/19/16/017, doi:10.1088/0305-4470/19/16/017.

[58] Gardner, E. Multiconnected neural network models. *Journal of Physics A: Mathematical and General*, **20** (1987), 3453. Available from: https://dx.doi.org/10.1088/0305-4470/20/11/046, doi:10.1088/0305-4470/20/11/046.

[59] Gardner, E. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, **21** (1988), 257.

[60] Gardner, E. and Derrida, B. Optimal storage properties of neural network models. **21** (1988), 271. Available from: https://dx.doi.org/10.1088/0305-4470/21/1/031, doi:10.1088/0305-4470/21/1/031.

[61] Gardner, E., Gutfreund, H., and Yekutieli, I. The phase space of interactions in neural networks with definite symmetry. *Journal of Physics A: Mathematical and General*, **22** (1989), 1995. Available from: https://dx.doi.org/10.1088/0305-4470/22/12/005, doi:10.1088/0305-4470/22/12/005.

[62] Gardner, E., Wallace, D., and Stroud, N. Training with noise and the storage of correlated patterns in a neural network model. *Journal of Physics A: Mathematical and General*, **22** (1989), 2019.

[63] Glauber, R. J. Time-dependent statistics of the ising model. *Journal of Mathematical Physics*, **4** (1963), 294. doi:10.1063/1.1703954.

[64] Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, **10** (2020), 041044.

[65] Guerra, F. Broken replica symmetry bounds in the mean field spin glass model. *Communications in Mathematical Physics*, **233** (2002). doi:10.1007/s00220-002-0773-5.

[66] Guerra, F. and Toninelli, F. Quadratic replica coupling in the sherrington-kirkpatrick mean field spin glass model. *Journal of Mathematical Physics*, **43** (2002). doi:10.1063/1.1483378.

[67] Guerra, F. and Toninelli, F. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, **230** (2002), 71. doi:10.1007/s00220-002-0699-y.

[68] Gutfreund, H., Reger, J. D., and Young, A. P. The nature of attractors in an asymmetric spin glass with deterministic dynamics. *Journal of Physics A: Mathematical and General*, **21** (1988), 2775. Available from: https://dx.doi.org/10.1088/0305-4470/21/12/020, doi:10.1088/0305-4470/21/12/020.

[69] Guzmán, S. J., Schlögl, A., Frotscher, M., and Jonas, P. Synaptic mechanisms of pattern completion in the hippocampal CA3 network. *Science*,

353 (2016), 1117. Available from: https://www.science.org/doi/10.1126/science.aaf1836, doi:10.1126/science.aaf1836.

[70] Hadsell, R., Rao, D., Rusu, A. A., and Pascanu, R. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, **24** (2020), 1028.

[71] Hayashi, K., Hashimoto, C., Kimoto, T., and Uezu, T. Unlearning of mixed states in the hopfield model—extensive loading case—. *Journal of the Physical Society of Japan*, **87** (2018), 054004. Available from: https://journals.jps.jp/doi/abs/10.7566/JPSJ.87.054004, doi:10.7566/JPSJ.87.054004.

[72] Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York (1949).

[73] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, **79** (1982), 2554. doi:10.1073/pnas.79.8.2554.

[74] Hopfield, J. J., Feinstein, D. I., and Palmer, R. G. Unlearning has a stabilizing effect in collective memories. *Nature*, **304** (1983), 158. Available from: https://doi.org/10.1038/304158a0, doi:10.1038/304158a0.

[75] Horas, J. A. and Pasinetti, P. M. On the unlearning procedure yielding a high-performance associative memory neural network. *Journal of Physics A: Mathematical and General*, **31** (1998), L463. Available from: https://doi.org/10.1088/0305-4470/31/25/001, doi:10.1088/0305-4470/31/25/001.

[76] Hwang, S., Folli, V., Lanza, E., Parisi, G., Ruocco, G., and Zamponi, F. On the number of limit cycles in asymmetric neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2019** (2019). Cited by: 7; All Open Access, Green Open Access. doi:10.1088/1742-5468/ab11e3.

[77] Hwang, S., Lanza, E., Parisi, G., Rocchi, J., Ruocco, G., and Zamponi, F. On the number of limit cycles in diluted neural networks. *Journal of Statistical Physics*, **181** (2020), 2304. doi:10.1007/s10955-020-02664-3.

[78] Kac, M. Mathematical mechanisms of phase transitions. *pp 241-305 of Statistical Physics, Phase Transitions, and Superfluidity. Vol. I. Chretien, M. Gross, E. P. Deser, S. (eds.). New York, Gordon and Breach, Science Publishers, 1968.*, (1969). Available from: https://www.osti.gov/biblio/4819765.

[79] Kalaj, S., Lauditi, C., Perugini, G., Lucibello, C., Malatesta, E. M., and Negri, M. Random features hopfield networks generalize retrieval to previously unseen examples. *arXiv preprint arXiv:2407.05658*, (2024).

[80] Kanter, I. and Sompolinsky, H. Associative recall of memory without errors. *Phys. Rev. A*, **35** (1987), 380. Available from: https://link.aps.org/doi/10.1103/PhysRevA.35.380, doi:10.1103/PhysRevA.35.380.

[81] Kleinfeld, D. and Pendergraft, D. "unlearning" increases the storage capacity of content addressable memories. *Biophysical Journal*, **51** (1987), 47. Available from: https://www.sciencedirect.com/science/article/pii/S0006349587833106, doi:https://doi.org/10.1016/S0006-3495(87)83310-6.

[82] Kojima, T., Nonaka, H., and Da-te, T. Capacity of the associative memory using the boltzmann machine learning. vol. 5, p. 2572 – 2577 (1995). Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0029452426&partnerID=40&md5=aee815aa3d92889c3254a466c034a7d5.

[83] Krauth, W. and Mézard, M. Storage capacity of memory networks with binary couplings. *Journal de Physique*, **50** (1989), 3057. Available from: https://hal.science/jpa-00211124, doi:10.1051/jphys:0198900500200305700.

[84] Krotov, D. and Hopfield, J. J. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, **29** (2016). arXiv:1606.01164.

[85] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, **521** (2015), 436.

[86] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86** (1998), 2278 . doi:10.1109/5.726791.

[87] Leonetti, M., Folli, V., Milanetti, E., Ruocco, G., and Gosti, G. Network dilution and asymmetry in an efficient brain. *Philosophical Magazine*, **100** (2020), 2544. Available from: https://doi.org/10.1080/14786435.2020.1750726, doi:10.1080/14786435.2020.1750726.

[88] Little, W. A. The existence of persistent states in the brain. *Mathematical Biosciences*, **19** (1974), 101. doi:10.1016/0025-5564(74)90031-5.

[89] Löwe, M. On the storage capacity of hopfield models with correlated patterns. *Annals of Applied Probability*, **8** (1998), 1216. Available from: https://api.semanticscholar.org/CorpusID:119824254.

[90] Lucibello, C. and Mézard, M. Exponential capacity of dense associative memories. *Phys. Rev. Lett.*, **132** (2024), 077301. Available from: https://link.aps.org/doi/10.1103/PhysRevLett.132.077301, doi:10.1103/PhysRevLett.132.077301.

[91] MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge University Press (2003).

[92] Marinari, E. Forgetting memories and their attractiveness. *Neural Comput*, **31** (2019), 503. doi:10.1162/neco_a_01162.

[93] Martorell, C., Calvo, R., Annibale, A., and Muñoz, M. A. Dynamically selected steady states and criticality in non-reciprocal networks.

Chaos, Solitons & Fractals, **182** (2024), 114809. Available from: https://www.sciencedirect.com/science/article/pii/S0960077924003618, doi:https://doi.org/10.1016/j.chaos.2024.114809.

[94] McCulloch, W. S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, **5** (1943), 115 – 133. Cited by: 11956. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-51249194645&doi=10.1007%2fBF02478259&partnerID=40&md5=edb67afceee33d22eaabbf1f8c1dca90, doi:10.1007/BF02478259.

[95] McEliece, R. J., Posner, E. C., Rodemich, E. R., and Venkatesh, S. S. The capacity of the hopfield associative memory. (2023). doi:10.1109/TIT.1987.1057328.

[96] Mezard, M. and Montanari, A. *Information, Physics, and Computation*. Oxford University Press, Inc., USA (2009). ISBN 019857083X.

[97] Mezard, M., Parisi, G., Virasoro, M., and Hopfield, J. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*. World Scientific Lecture Notes In Physics. World Scientific Publishing Company (1987). ISBN 9789813103917. Available from: https://books.google.it/books?id=DwY8DQAAQBAJ.

[98] Minsky, M. and Papert, S. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA (1969).

[99] Morris, R. and Hebb, D. The organization of behavior. *Wiley: New York, Brain Res Bull. 1999;50(5-6):437*, (1949). doi:10.1016/s0361-9230(99)00182-3.

[100] Mézard, M., Nadal, J. P., and Toulouse, G. Solvable models of working memories. *Journal de Physique*, **47** (1986), 1457. arXiv:jpa-00210340, doi:10.1051/jphys:019860047090145700.

[101] Negri, M., Lauditi, C., Perugini, G., Lucibello, C., and Malatesta, E. Storage and learning phase transitions in the random-features hopfield model. *Physical Review Letters*, **131** (2023), 257301.

[102] Nishimori, H. *Statistical Physics of Spin Glasses and Information Processing* (2001).

[103] Nokura, K. Paramagnetic unlearning in neural network models. *Phys. Rev. E*, **54** (1996), 5571. Available from: https://link.aps.org/doi/10.1103/PhysRevE.54.5571, doi:10.1103/PhysRevE.54.5571.

[104] Nokura, K. Unlearning in the paramagnetic phase of neural network models. *Journal of Physics A: Mathematical and General*, **29** (1996), 3871. Available from: https://doi.org/10.1088/0305-4470/29/14/013, doi:10.1088/0305-4470/29/14/013.

[105] Nouri, A. and Seyyedsalehi, S. A. Eigen value based loss function for training attractors in iterated autoencoders. *Neural Networks*, (2023).

[106] Nutzel, K. The length of attractors in asymmetric random neural networks with deterministic dynamics. *Journal of Physics A: Mathematical and General*, **24** (1991), L151. Available from: https://dx.doi.org/10.1088/0305-4470/24/3/010, doi:10.1088/0305-4470/24/3/010.

[107] Nutzel, K. and Krey, U. Subtle dynamic behaviour of finite-size sherrington-kirkpatrick spin glasses with nonsymmetric couplings. *Journal of Physics A: Mathematical and General*, **26** (1993), L591. Available from: https://dx.doi.org/10.1088/0305-4470/26/14/001.

[108] Parisi, G. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, **43** (1979), 1754.

[109] Parisi, G. Order parameter for spin-glasses. *Physical Review Letters*, **50** (1983), 1946.

[110] Parisi, G. Asymmetric neural networks and the process of learning. *Journal of Physics A: Mathematical and General*, **19** (1986), L675. Available from: https://dx.doi.org/10.1088/0305-4470/19/11/005, doi:10.1088/0305-4470/19/11/005.

[111] Parisi, G. A memory which forgets. *Journal of Physics A: Mathematical and General*, **19** (1986), L617.

[112] Pastur, L., Shcherbina, M., and Tirozzi, B. The replica-symmetric solution without replica trick for the Hopfield model. *Journal of Statistical Physics*, **74** (1994), 1161. doi:10.1007/BF02188221.

[113] Peretto, P. *An Introduction to the Modeling of Neural Networks*. Collection Alea-Saclay: Monographs and Texts in Statistical Physics. Cambridge University Press (1992).

[114] Peretto, P. and Niez, J. Stochastic dynamics of neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, **16** (1986), 73.

[115] Personnaz, L., Guyon, I., and Dreyfus, G. Information storage and retrieval in spin-glass like neural networks. *Journal de Physique Lettres*, **46** (1985), 359.

[116] Pfenning, T., Rieger, H., and Schreckenberg, M. Numerical investigation of the asymmetric sk-model with deterministic dynamics. *http://dx.doi.org/10.1051/jp1:1991134*, **1** (1991). doi:10.1051/jp1:1991134.

[117] Plakhov, A. and Semenov, S. The modified unlearning procedure for enhancing storage capacity in hopfield network. (1992), 242.

[118] Plakhov, A. Y. The converging unlearning algorithm for the hopfield neural network: optimal strategy. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, vol. 2, pp. 104–106 (1994).

[119] Pöppel, G. and Krey, U. Dynamical learning process for recognition of correlated patterns in symmetric spin glass models. *Europhysics Letters*, **4** (1987), 979.

[120] Ramsauer, H., et al. Hopfield networks is all you need (2021). Available from: https://arxiv.org/abs/2008.02217, arXiv:2008.02217.

[121] Robins, A. and McCallum, S. Catastrophic Forgetting and the Pseudorehearsal Solution in Hopfield-type Networks. *Connection Science*, **10** (1998), 121. Available from: https://www.academia.edu/101394251/Catastrophic_Forgetting_and_the_Pseudorehearsal_Solution_in_Hopfield_type_Networks.

[122] Rolls, E. T. *Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach.* Oxford University Press (2007). ISBN 9780199232703. Available from: https://doi.org/10.1093/acprof:oso/9780199232703.001.0001, doi:10.1093/acprof:oso/9780199232703.001.0001.

[123] Rolls, E. T. Advantages of dilution in the connectivity of attractor networks in the brain. *Biologically Inspired Cognitive Architectures*, **1** (2012), 44. Available from: https://www.sciencedirect.com/science/article/pii/S2212683X12000047, doi:https://doi.org/10.1016/j.bica.2012.03.003.

[124] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65, 386.*, (1958). doi:10.1037/h0042519.

[125] Saglietti, L., Gerace, F., Ingrosso, A., Baldassi, C., and Zecchina, R. From statistical inference to a differential learning rule for stochastic neural networks. *Interface Focus*, **8** (2018), 20180033.

[126] Scellier, B. and Bengio, Y. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, **11** (2017), 24. doi:10.3389/fncom.2017.00024.

[127] Serricchio, L., Bocchi, D., Chilin, C., Marino, R., Negri, M., Cammarota, C., and Ricci-Tersenghi, F. Daydreaming hopfield networks and their surprising effectiveness on correlated data. *Neural Networks*, **186** (2025), 107216. Available from: https://www.sciencedirect.com/science/article/pii/S0893608025000954, doi:https://doi.org/10.1016/j.neunet.2025.107216.

[128] Sherrington, D. and Kirkpatrick, S. Solvable model of a spin-glass. *Phys. Rev. Lett.*, **35** (1975), 1792. Available from: https://link.aps.org/doi/10.1103/PhysRevLett.35.1792, doi:10.1103/PhysRevLett.35.1792.

[129] Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLOS Biology*, **3** (2005), null. Available from: https://doi.org/10.1371/journal.pbio.0030068, doi:10.1371/journal.pbio.0030068.

[130] Spitzner, P. and Kinzel, W. Freezing transition in asymmetric random neural networks with deterministic dynamics. *Zeitschrift für Physik B Condensed Matter*, **77** (1989), 511. Available from: https://doi.org/10.1007/BF01453802, doi:10.1007/BF01453802.

[131] Storkey, A. Increasing the capacity of a hopfield network without sacrificing functionality. (1997), 451. doi:10.1007/BFb0020196.

[132] Storkey, A. J. and Valabregue, R. The basins of attraction of a new hopfield learning rule. *Neural Networks*, **12** (1999), 869. doi:10.1016/S0893-6080(99)00038-6.

[133] Talagrand, M. Rigorous results for the Hopfield model with many patterns. *Probability Theory and Related Fields*, **110** (1998), 177. doi:10.1007/s004400050148.

[134] Talagrand, M. *Spin glasses : a challenge for mathematicians ; cavity and mean field models*. Ergebnisse der Mathematik und ihrer Grenzgebiete, 3. Folge, v. 46. Springer, Berlin ; (2003). ISBN 3540003568.

[135] Talagrand, M. The parisi formula. *Annals of Mathematics*, **163** (2006), 221. Available from: https://api.semanticscholar.org/CorpusID:56101828.

[136] Tanaka, F. and Edwards, S. F. Analytic theory of the ground state properties of a spin glass. i. ising spin glass. *Journal of Physics F: Metal Physics*, **10** (1980), 2769. Available from: https://dx.doi.org/10.1088/0305-4608/10/12/017, doi:10.1088/0305-4608/10/12/017.

[137] Treves, A. and Rolls, E. T. Computational constraints suggest the need for two distinct input systems to the hippocampal ca3 network. *Hippocampus*, **2** (1992), 189. doi:10.1002/hipo.450020209.

[138] Tsodyks, M. V. Associative memory in asymmetric diluted network with low level of activity. *Europhysics Letters*, **7** (1988), 203. Available from: https://doi.org/10.1209/0295-5075/7/3/003, doi:10.1209/0295-5075/7/3/003.

[139] Tsodyks, M. V. and Feigel'man, M. V. The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, **6** (1988), 101. Available from: https://doi.org/10.1209/0295-5075/6/2/002, doi:10.1209/0295-5075/6/2/002.

[140] Van Hemmen, J. Hebbian learning, its correlation catastrophe, and unlearning. *Network: Computation in Neural Systems*, **8** (1997), V1.

[141] Van Hemmen, J., Ioffe, L., Kühn, R., and Vaas, M. Increasing the efficiency of a neural network through unlearning. *Physica A: Statistical Mechanics and its Applications*, **163** (1990), 386.

[142] Van Hemmen, J. and Klemmer, N. Unlearning and its relevance to rem sleep: Decorrelating correlated data. In *Neural Network Dynamics: Proceedings of the Workshop on Complex Dynamics in Neural Networks, June 17–21 1991 at IIASS, Vietri, Italy*, pp. 30–43. Springer (1992).

[143] Van Hemmen, J. L., Keller, G., and Kühn, R. Forgetful memories. *Europhysics Letters*, **5** (1988), 663. Available from: https://dx.doi.org/10.1209/0295-5075/5/7/016, doi:10.1209/0295-5075/5/7/016.

[144] Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2024).

[145] Watkin, T. L. H., Rau, A., and Biehl, M. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, **65** (1993), 499. doi:10.1103/RevModPhys.65.499.

[146] Witter, M. P. Connectivity of the hippocampus. In *Hippocampal Microcircuits: A Computational Modeler's Resource Book* (edited by V. Cutsuridis, B. Graham, S. Cobb, and I. Vida), pp. 5–26. Springer New York, New York, NY (2010). ISBN 978-1-4419-0996-1. Available from: https://doi.org/10.1007/978-1-4419-0996-1_1, doi:10.1007/978-1-4419-0996-1_1.