EPFL

Thèse n° 10 094

A theory of memory consolidation and synaptic pruning in cortical circuits

Présentée le 13 décembre 2023

Faculté des sciences de la vie
Projet Bluebrain
Programme doctoral en neurosciences

pour l'obtention du grade de Docteur ès Sciences

par

Georgios IATROPOULOS

Acceptée sur proposition du jury

Prof. C. Petersen, président du jury
Prof. H. Markram, Prof. W. Gerstner, directeurs de thèse
Prof. C. Clopath, rapporteuse
Prof. W. Senn, rapporteur
Prof. A. Mathis, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

Acknowledgements

The work presented in this thesis is the product of a four-year collaboration between the Blue Brain Project (BBP) and the Laboratory of Computational Neuroscience (LCN) at EPFL. I would therefore like to begin by expressing my gratitude for the technical and financial support that has been made available to me through the BBP and by my thesis director Prof. Henry Markram, who has has allowed to me to independently pursue my research interests. I am equally grateful to my co-director Prof. Wulfram Gerstner, for taking on a supervisory role and providing me with a place in his lab, together with feedback on my ideas and my writing throughout the doctoral studies.

I would like to thank my colleagues at the BBP Connectomics Group and the LCN for all their encouraging and helpful comments on my work, and for creating a welcoming working environment with lively discussions on science and society.

At the BBP: Dr. James Isbister, András Ecker, Sirio Bolaños Puchet, Dr. Giuseppe Chindemi, Dr. Michael Reimann, Daniela Egas Santander, Joseph Tharayil, Dr. Max Nolte.

At the LCN: Dr. Valentin Schmutz, Sophia Becker, Louis Pezon, Flavio Martinelli, Alireza Modirshanechi, Christos Sourmpis, Dr. Bernd Illing, Dr. Martin Barry, Dr. Berfin Şimşek, Shuqi Wang, Dr. Chiara Gastaldi, Dr. Guillaume Bellec.

I would like to especially thank Dr. Johanni Brea, a senior scientist and lecturer at the LCN, for being almost like a second supervisor, and whose expertise, encouragement, and assistance has been invaluable.

Finally, I express my deepest gratitude to the people whose patient support, more than anything else, has been essential for the completion of this thesis, namely my family: my mother Lambrini Theodossiou, my father Fotios Iatropoulos, and my brother Terry Vassiliadis.

Abstract

Over the course of a lifetime, the human brain acquires an astonishing amount of semantic knowledge and autobiographical memories, often with an imprinting strong enough to allow detailed information to be recalled many years after the initial learning experience took place. The formation of such long-lasting memories is known to primarily involve cortex, where it is accompanied by a wave of synaptic growth, pruning, and fine-tuning that stretches across several nights of sleep. This process, broadly referred to as consolidation, gradually stabilizes labile information and moves it into permanent storage. It has a profound impact on connectivity and cognitive function, especially during development. Though extensively studied in terms of behavior and neuroanatomy, it is still unclear how this interplay between structural adaptation and long-term memory consolidation can be explained from a theoretical and computational perspective.

In this thesis, we take a top-down approach to develop a mathematical model of consolidation and pruning within the context of recurrent neural networks, by combining recent techniques from the fields of optimization, machine learning, and statistics. The first part of the thesis treats the problem of maximally noise-robust memory without synaptic resource constraints. Using kernel methods, we derive a compact description of networks with optimal weight configuration. This unifies many of the classical memory models under a common mathematical framework, and formalizes the relationship between active dendritic processing on the single-neuron level, and the storage capacity of the circuit as a whole.

In the second part of the thesis, we treat the problem of maximal memory robustness under conditions of sparse connectivity. We combine our unconstrained model with an implicit regularization, by endowing the network with bi- and tri-partite synapses, instead of the usual scalar weights. This allows us to derive a simple synaptic learning rule that simultaneously consolidates memories and prunes weights, while incorporating memory replay, multiplicative homeostatic scaling, and weight-dependent plasticity. We also use the synapse model to derive scaling properties of intrinsic synaptic noise, which we test in a meta-analysis of experimental data on dendritic spine dynamics.

In the concluding sections, we briefly discuss the implication of our results with regards to current memory-inspired machine learning methods, the function of sleep, and the environmental effects on structural plasticity in development.

Keywords: artificial neural networks, attractor networks, Hopfield networks, Hebbian learning, associative learning, kernel machines, support vector machines, REM sleep, dot-product attention, fractional norm, pyramidal cells, declarative memory, connectome, engram.

Résumé

Au cours d'une vie, le cerveau humain acquiert une quantité étonnante de connaissances sémantiques et de souvenirs autobiographiques, souvent avec une empreinte suffisamment forte pour que des informations détaillées puissent être rappelées de nombreuses années après l'expérience d'apprentissage initiale. La formation de ces souvenirs durables implique principalement le cortex, où elle s'accompagne d'une vague de croissance synaptique, d'élagage et de réglage fin qui s'étend sur plusieurs nuits de sommeil. Ce processus, généralement appelé consolidation, stabilise progressivement les informations labiles et a un impact profond sur la connectivité et les fonctions cognitives, en particulier au cours du développement. Bien qu'elle ait été largement étudiée en termes de comportement et de neuroanatomie, la manière dont cette interaction entre l'adaptation structurelle et la consolidation peut être expliquée d'un point de vue théorique n'est toujours pas claire.

Dans cette thèse, nous adoptons une approche descendante pour développer un modèle mathématique de consolidation et d'élagage dans le contexte des réseaux neuronaux récurrents, en combinant des techniques récentes issues des domaines de l'optimisation et de l'apprentissage automatique. La première partie de la thèse traite le problème de la mémoire maximalement robuste au bruit sans contraintes de ressources synaptiques. En utilisant des méthodes de noyau, nous dérivons une description compacte des réseaux avec une configuration optimale des poids. Cela permet d'unifier de nombreux modèles de mémoire classiques dans un cadre mathématique commun et de formaliser la relation entre le traitement dendritique et la capacité de stockage.

Dans la deuxième partie de la thèse, nous traitons le problème de la robustesse maximale de la mémoire dans des conditions de connectivité éparse. Nous combinons notre modèle sans contrainte avec une régularisation implicite, en utilisant des synapses bi- et tripartites, au lieu des poids scalaires habituels. Cela nous permet de dériver une règle d'apprentissage synaptique qui consolide les mémoires et élague les poids, tout en incorporant le rejeu de mémoire, l'échelonnement homéostatique multiplicatif et la plasticité dépendante du poids. Nous utilisons également le modèle de synapse pour dériver des propriétés d'échelle du bruit synaptique intrinsèque, que nous testons dans une méta-analyse de données expérimentales sur la dynamique des épines dendritiques.

Mots clés: réseaux de neurones artificiels, réseaux d'attracteurs, réseaux Hopfield, règle de Hebb, apprentissage associatif, astuce du noyau, machines à vecteur de support, sommeil paradoxal, cellules pyramidales, mémoire déclarative, connectome, engramme.

Contents

Acknowledgements i

Abstract (English/Français) iii

List of figures xi

List of tables xiii

1 Introduction 1
- 1.1 A brief history of memory modeling . . . 2
- 1.2 A primer on synaptic structural plasticity . . . 4
- 1.3 Scope and terminology . . . 7
- 1.4 Summary of the thesis . . . 8

2 Kernel memory networks: a unifying framework 11
- 2.1 Introduction . . . 12
  - 2.1.1 Our contributions . . . 12
  - 2.1.2 Related work . . . 12
- 2.2 Background . . . 13
- 2.3 Kernel memory networks for binary patterns . . . 14
  - 2.3.1 Hetero-associative memory as a feed-forward SVM network . . . 14
  - 2.3.2 Auto-associative memory as a recurrent SVM network . . . 14
  - 2.3.3 The Kanerva network is a feed-forward SVM network . . . 16
  - 2.3.4 The Hopfield network is a recurrent SVM network . . . 18
- 2.4 Kernel memory networks for continuous patterns . . . 20
  - 2.4.1 Auto-associative memory as a recurrent interpolation network . . . 20
  - 2.4.2 A recurrent interpolation network with exponential capacity . . . 21
- 2.5 Discussion . . . 23

Appendix . . . 25
- A2.1 Derivation of storage capacity scaling . . . 25
- A2.2 Kernel of an infinite SDM on the hypersphere . . . 26
- A2.3 Iterative learning in an SVM network . . . 28
- A2.4 Generalized pseudoinverse rule . . . 29
- A2.5 The kernel memory network for continuous patterns . . . 30
- A2.6 Comparison to neuron models with active dendrites . . . 38

3 Optimal memory consolidation and pruning **41**
3.1 Introduction . . . 42
3.2 Results . . . 43
3.2.1 Memory consolidation as sparse optimization . . . 43
3.2.2 Learning with complex synapses and memory replay . . . 45
3.2.3 Multiplicative synapses produce optimal storage efficiency . . . 47
3.2.4 Simultaneous consolidation and pruning with multiplicative synapses in sleep . . . 50
3.2.5 Preferential consolidation of weakly encoded memories in sleep . . . 54
3.2.6 Intrinsic synaptic noise scales sublinearly with weight . . . 55
3.2.7 Estimating homeostatic scaling from intrinsic synaptic noise . . . 59
3.3 Discussion . . . 59
3.3.1 The robustness-redundancy trade-off . . . 60
3.3.2 Replay and sleep . . . 61
3.3.3 Implications for life-long learning . . . 62
3.3.4 Biological interpretation of the synapse model . . . 63
3.3.5 Predictions and future work . . . 64
3.4 Methods . . . 65
3.4.1 Network model . . . 65
3.4.2 Memory patterns . . . 65
3.4.3 SNR and error margin . . . 66
3.4.4 Theoretical solutions . . . 67
3.4.5 Synapse model . . . 67
3.4.6 General learning rule . . . 67
3.4.7 Numerical optimization and evaluation . . . 68
3.4.8 Simulating wakefulness and sleep . . . 69
3.4.9 Simulating synaptic volatility . . . 71
3.4.10 Experimental data: Connectivity . . . 72
3.4.11 Experimental data: Synaptic volatility . . . 72
3.4.12 Experimental data: CV of synapse norm . . . 73
3.4.13 Experimental data: Synaptic pruning . . . 73
3.4.14 Experimental data: Memory consolidation in sleep . . . 74
Appendix . . . 76
A3.1 Derivation of consolidation algorithm . . . 76
A3.2 Derivation of homeostatic scaling laws . . . 77
A3.3 Derivation of input current statistics . . . 79
A3.4 Theoretical solution for maximal SNR . . . 79
A3.5 Theoretical solution for maximal pruning . . . 81
A3.6 Theoretical solution for $q=1$ . . . 82
A3.7 Learning rule for $q=1$ and fixed inhibition . . . 82
A3.8 Simulation parameters . . . 84
A3.9 Synapse metadata . . . 85
A3.10 Supplementary figures . . . 86

4 Conclusion **95**

Bibliography **99**

Curriculum Vitae **113**

List of Figures

1.1 Hierarchical summary of the thesis. 9

2.1 Schematic of kernel memory networks. 15
2.2 Attractor basins of pattern on a sphere. 24
A2.1 Exact and approximate kernel of an infinite SDM. 28
A2.2 Attractor basin size after iterative and one-shot learning. 29
A2.3 Storage capacity scaling of a continuous kernel memory network. 34

3.1 Schematic of the circuit and synapse model. 44
3.2 Attractor networks at dense and sparse optimality 48
3.3 Schematic of encoding and consolidation over a day 51
3.4 Simulations of encoding and consolidation over a day 53
3.5 Scaling of synaptic fluctuations 56
3.6 Simulated and experimental synaptic fluctuation statistics over time . . . 58
3.7 Schematic of predicted consolidation over a lifetime 62
A3.1 Recall performance of optimal attractor networks 86
A3.2 Additional properties of optimal attractor networks 87
A3.3 Efficiency of optimal attractor networks with isotropic noise 88
A3.4 Simulated synaptic volatility 89
A3.5 Different noise models 90
A3.6 Theoretical and numerical storage optimization 91
A3.7 The effect of tripartite weights on the optimization landscape 92
A3.8 Geometrical explanation of $K_{q=1}$ maximization 93
A3.9 Control models for SNR maximization 94

List of Tables

A3.1 Simulation parameters for Fig. 3.2 84
A3.2 Simulation parameters for Figs. 3.5 and 3.6 84
A3.3 Simulation parameters for Fig. 3.4 84
A3.4 Description of synaptic data with short sampling intervals 85
A3.5 Description of synaptic data with long sampling intervals 85

Chapter 1

Introduction

Some 2400 years ago, in what could perhaps be considered to be one of history's earliest recorded attempts at modeling the brain, Socrates famously argued that the memory of a human can be likened to a block of wax (Plato, 1990, sec. 191-195). Whenever an external stimulus is strongly perceived, it imprints a pattern into the wax, like the seal on a letter, and a memory is formed. When the same stimulus is encountered again at a later occasion, the perception is fitted to the imprint in the wax, and recognition takes place. This analogy was also used to explain how some people can posses a lot of knowledge and be wise, while others easily make errors of judgment. In wise people, Socrates explained, the block of wax is large and can be imprinted many times, without any risk that overlap occurs. The wax is clean and of good consistency, which allows each imprint to be deep and distinct. Conversely, in ignorant people, the block is small and the imprints easily become crowded and obscured. The wax might also be too soft, too hard, or impure, causing the imprints to become blurry, shallow, or distorted. This, Socrates concluded, is ultimately what produces misunderstanding and false recollection of past events.

Even though this conceptual model may seem old-fashioned and elementary to a modern audience, it is strikingly similar to the way in which neuroscientists have been thinking about the mechanisms of memory and learning for the past 70 years. While current models of the brain are dominated by concepts from digital information technology, with the wax tablet having been replaced by magnetic spin lattices and computer hard-drives, the biological processes that underlie the formation, retrieval, and forgetting of memories are still understood in terms of imprinting and matching of patterns in a malleable substance. The question, therefore, of exactly how memorization and recall is implemented in real neuronal circuits remains an active area of research to this day.

The gap between our theoretical understanding of memory and learning, and the current state of knowledge in neurobiology and -physiology, has become particularly pronounced in the modern era of experimental neuroscience, spanning roughly the last 20-30 years. During this time, the development of new microscopy and nanoscopy techniques (see, e.g., Holtmaat et al., 2009; Berning et al., 2012) has produced a wealth of new data on the dynamics of not only large populations of neurons, but also individual synapses, in living animals. One of the most important insights to come out of these studies has been the fact that the synaptic configuration of cortical circuits is remarkably volatile over time. Indeed, while the shape and arborization of entire dendrites and axons is relatively stable, synaptic boutons

and dendritic spines undergo a constant process of formation and retraction (Holtmaat & Svoboda, 2009). This turnover is modulated both by the cycle of wakefulness and sleep (e.g., Xu et al., 2009; Chen et al., 2015) as well as by the stage of development (Petanjek et al., 2011). Even existing synapses are inherently unreliable, and display a substantial degree of activity-independent size fluctuations (Kasai et al., 2021). How should this constant rewiring and remodeling of the brain be incorporated into current connectionist models of learning and memory?

In order to begin answering this question, we will first provide a brief review of previous work on memory modeling with neural networks. This will be followed by a short summary of the experimental literature on synaptic structural plasticity. Finally, with the historical background in mind, we will define a set of mathematical problems regarding memory consolidation and structural plasticity, which will form the basis of the next two chapters. We conclude by summarizing the results and contributions of the thesis.

1.1 A brief history of memory modeling

The consensus among today's neuroscientists is that the primary physical correlate (and cause) of learning in nervous systems is synaptic plasticity. The general idea, however, that information storage is expressed in the modification of neural connections can be traced back at least to the late 19th century and the work by Santiago Ramón y Cajal (Yuste, 2015). In fact, the term *engram*, which is commonly used today to refer to essentially any physical change in the brain induced by learning, was coined in the early years of the 20th century by one of Cajal's contemporaries, Richard Semon, who posited that such imprints generally are dormant but can be awakened, in a retrieval process, by partial cues (Josselyn et al., 2017).

Despite the prescience of these early ideas, the question of how, exactly, brain activity can produce memorization and recall, and how this is underpinned by alterations in physiology, remained debated by psychologists and biologists until the middle of the 20th century. The contention between hypotheses centered on neural activity vis-à-vis neural structure, was finally reconciled by the theory of *cell assembly* formation, which today is credited to Donald Hebb (1949, pp. 60-66).[1] Hebb postulated that memory activation entails two, mutually reinforcing, neurophysiological components: First, the perception of a stimulus is reflected in the transient reverberation of neural activity, which, if allowed to continue long enough, induces a lasting structural change by strengthening the synaptic connections between the underlying co-active neurons (i.e., the assembly), thus making them more likely to reverberate and awaken the memory of the stimulus in the future. The strengthening of connections, Hebb specified, could be accomplished both by the formation of new synaptic "knobs", as well as by the enlargement of existing ones.

[1] Note, however, that a very similar theory had been published a year earlier by Konorski (1948) to explain the learning of conditioned reflexes (see review by Zieliński, 2006).

First-generation network models. The theory of cell assemblies and synaptic plasticity found fertile ground in the nascent field of artificial intelligence, and, in particular, connectionism. As a predecessor of what we today refer to as *artificial neural networks* and *deep learning*, the foundational idea of this discipline was to model cognition and behavior as a function of interconnected simple, binary units, called perceptrons, that collectively operate much like a switchboard or a transistor circuit (see, e.g., Rosenblatt, 1962, ch. 3, and references therein). It was soon understood that these models also could be interpreted in terms of magnetic spin lattices, which suggested that the reverberating activity that underlies memory recall could be an emergent phenomenon, where neurons mutually activate each other and self-organize into stable patterns of activity (Amari, 1972; Nakano, 1972), analogously to how long-range correlations or symmetry breaks appear in Ising models (Little, 1974; Hopfield, 1982). Each memory could, in the terminology of dynamical systems, be described as an *attractor* in the state space of the network.

Studies of attractor networks in the 1980s and 90s resulted in a series of influential publications, which today are considered classics in the field of memory modeling. These were partly enabled by a set of new mathematical techniques borrowed from statistical physics, which made it possible to characterize the optimal storage properties of recurrent networks, both without constraints (Cover, 1965; Venkatesh, 1986; Gardner, 1987a), but also with brain-inspired parameter restrictions, such as sparse activity (Gardner, 1988), binary connections (Krauth & Mézard, 1989), discrete connections (Gutfreund & Stein, 1990; Baldassi et al., 2016), sign-constrained connections (Amit et al., 1989; Kanter & Eisenstein, 1990; Nadal, 1990; Viswanathan, 1993), pruned connections (Gardner, 1989; Bouten et al., 1990), and higher-order connections (Lee et al., 1986; Peretto & Niez, 1986; Abbott & Arian, 1987; Gardner, 1987b).[2] Many of these results were obtained using mean-field theory under a synaptic weight scaling of $\mathcal{O}(1/\sqrt{N})$. In chapter 2, we propose a new way of understanding this family of models, using a general, normative framework based on optimal storage robustness.

Second-generation network models. The interest in attractor networks was revived in the middle of the 2000s by a new wave of theoretical findings, this time obtained by analyzing networks with an $\mathcal{O}(1/N)$ scaling (for an early example, see Köhler & Widmaier, 1991). This variant was shown to be particularly interesting as a normative model of not only memory function, but also of cortical anatomy. At optimal storage, these attractor networks display a very sparse overall connection probability (Brunel et al., 2004) with realistic differences between excitatory and inhibitory neurons (Chapeton et al., 2012), as well as an over-representation of bi-directional connections (Brunel, 2016) and higher-order connection motifs (Brunel, 2016; Zhang et al., 2019); all in agreement with experimental data (Song et al., 2005; Perin et al., 2011).

By and large, the conclusion drawn from these results has been that long-term memory in adult neocortex operates as an attractor network at optimal storage. This, however, has raised the question of how a neural circuit can reach such a state of optimality and,

[2] We show in chapter 2 that higher-order connections can be interpreted as models of synaptic cross-talk and non-linear dendritic integration.

in particular, how a synaptic plasticity mechanism would be capable of endowing a circuit with optimal memory storage. A satisfactory solution to this problem is still lacking. The correlational, one-shot learning rules that were explored in the early literature (Amari, 1972; Kohonen, 1972; Nakano, 1972; Hopfield, 1982; Tsodyks & Feigel'man, 1988) have a sub-optimal performance (Amit et al., 1985; McEliece et al., 1987; Amari, 1989), and the iterative rules that have been derived with optimization methods (e.g., Gardner, 1988; Frieß et al., 1998; Alemi et al., 2015; Sacramento et al., 2015) require plasticity mechanisms that, in general, are either biologically implausible, incompatible with homeostatic plasticity mechanisms (Turrigiano, 2008), or inconsistent with synaptic dynamics (Yasumatsu et al., 2008; Loewenstein et al., 2011). We propose a solution to this problem in chapter 3.

Third-generation network models. Over the past six to seven years, the remarkable success of transformers (Vaswani et al., 2017) and related attention-based deep learning applications has sparked a renewed interest in memory network research. The result has been a new generation of models, variably referred to as dense associative memory (Krotov & Hopfield, 2016, 2020), modern Hopfield networks (Ramsauer et al., 2021; Millidge et al., 2022), or key-value memory networks (Sukhbaatar et al., 2015; Tyulmankov et al., 2021). The derivation of these models still relies on energy minimization, but requires a new formulation of the Hamiltonian, which substantially enhances storage capacity (see, e.g. Demircigil et al., 2017). In chapter 2, however, we show that the third- and first-generation models belong to the same family, and can be obtained by maximizing the same robustness-based objective, albeit with different neuron models; while first-generation networks consist of neurons that only perform linear integration, third-generation neurons are non-linear, and contain an added "dendritic" compartment.

In spite of this recent progress in memory modeling, the development of brain-inspired learning rules with structural synaptic change, such as pruning, appears to be particularly challenging, and has generally received less attention by the theoretical community than problems regarding functional and homeostatic plasticity. Moreover, the few models that have been published have predominantly been phenomenological, without an algorithmic basis or performance guarantees (e.g., Levy, 2004; Knoblauch et al., 2014; Gallinaro et al., 2022). This is partly a consequence of the fact that experimental tools for observing and quantifying synaptic rewiring have, until recently, been unavailable. We briefly review this literature in the next section.

1.2 A primer on synaptic structural plasticity

While it has long been possible to measure the functional strength of synaptic connections using electrophysiological techniques, direct observation of the dynamics of synaptic anatomy and structure has only been possible since the beginning of the 2000s and the advent of live tissue imaging at spatial resolutions of single micrometers. The effort to quantify structural plasticity properties have been further complicated by the fact that the stability of an individual synapse is a latent variable that has to be inferred from population statistics, by measuring, for example, synaptic lifetime, density, or rate of formation and retraction

(Loewenstein et al., 2015). This typically requires experimental setups capable of tracking large numbers of synapses over time periods spanning several hours to multiple days.

Early experiments. The earliest observations of synaptic structural plasticity in cortex were done in post-mortem studies, by estimating the synaptic density in fixed brain samples taken from rodents and cats that were reared in environments enriched in or deprived of stimuli (Valverde, 1967; Fifková, 1968; Globus et al., 1973; Parnavelas et al., 1973; Cragg, 1975; Turner & Greenough, 1985). Previous studies had inferred macroscopic effects of structural plasticity from measurements of either cortical weight (Rosenzweig et al., 1962; Bennett et al., 1964) or thickness (Diamond et al., 1967). The conclusion was that sensory enrichment generally enhanced synaptogenesis and caused an elevated synaptic density, presumably to accommodate the additional learning and information processing required to navigate complex environments. These findings, in combination with the discovery of electrically induced synaptic potentiation (Bliss & Lømo, 1973), served as the first pieces of evidence in support of the idea that sensory experiences could leave lasting neurophysiological and neuroanatomical traces in the brain.

Rewiring over years. Modifications of connectivity are not only regulated by experience but also by the ontogenetic stage of an animal. In a series of experiments during the 1970s and 80s, synaptic density was measured at different ages in rodents (Aghajanian & Bloom, 1967; Feldman & Dowd, 1975), cats (Winfield, 1981; O'Kusky, 1985), monkeys (O'Kusky & Colonnier, 1982; Rakic et al., 1986), and humans (Huttenlocher, 1979; Huttenlocher et al., 1982). These studies resulted in what is today a broadly accepted understanding of the brain's innate ability to modulate the turnover of synapses throughout development. In early infancy, synaptogenesis is dramatically ramped up, causing the density of synapses to quickly reach a maximum. During adolescence, the rate of synapse elimination slightly outbalances the rate of formation (Zuo et al., 2005a), causing synapse density to slowly decrease and plateau in adulthood (Petanjek et al., 2011).

Rewiring over days. By the late 1990s, the recent development of two-photon fluorescence microscopy enabled experimentalists to image dendritic spines *in vivo* over timescales of several days. This technique, combined with serial electron microscopy and three-dimensional reconstruction, have established most of what is currently known about the synaptic lifecycle.[3] Presynaptic activity, mediated by glutamate release, promotes the growth of both dendritic spines (Maletic-Savatic et al., 1999) and entire connections (Le Bé & Markram, 2006) on the postsynaptic neuron. New spines do not initially form synapses (Knott et al., 2006; Nägerl et al., 2007). This structure is added later, in an activity-dependent maturation process (De Roo et al., 2008), preferentially onto already existing axonal boutons (Knott et al., 2006), which also undergo structural changes, though they tend to be more stable than spines (De Paola et al., 2006; Majewska et al., 2006; Qiao et al., 2016). Spino- and synaptogenesis is reversible, as stimuli that induce functional long-term depression eventually cause spine retraction and elimination (Nägerl et al., 2004; Hayama et al., 2013; Oh et al., 2013; Wiegert & Oertner, 2013).

[3] We focus here primarily on excitatory neurons and, in particular, cortical pyramidal cells.

Experiential modulation of dendritic spine stability and turnover has been demonstrated in experiments involving both sensory deprivation (Lendvai et al., 2000; Trachtenberg et al., 2002; Zuo et al., 2005b; Holtmaat et al., 2006) and active task-learning (Xu et al., 2009; Yang et al., 2009; Chen et al., 2015). The link between spine dynamics and learning is, in fact, not only correlational but also causal; for example, spine formation in motor cortex is directly necessary for the acquisition of new motor skills (Hayashi-Takagi et al., 2015). Both experience-driven spine formation and elimination occurs to a larger extent during sleep[4] than during wakefulness (Yang et al., 2014; Chen et al., 2015; Li et al., 2017; Zhou et al., 2020). These types of studies have concluded that novel sensory input promotes rewiring of neural circuits by destabilizing old, existing, dendritic spines and stabilizing new ones.

Statistics of dendritic spine turnover have successfully been reproduced in simulations of multi-synaptic connections (Fauth et al., 2015; Deger et al., 2018). Although this work is based on phenomenological modeling, it has demonstrated that flexible synaptic motility can serve a computational purpose, by allowing a network to quickly learn new information and maintain it for long periods of time, in spite of internal noise. A second set of theoretical studies have suggested that dendritic rewiring could be a way for the brain to implement Bayesian inference and particle filtering (Kappel et al., 2015; Hiratani & Fukai, 2018). This work, however, is based on normative assumptions and has not been corroborated with experimental data.

The model presented in chapter 3 differs from past work in that it focuses on whole connections and exclusively on the problem of optimal pruning, in a single bout of consolidation. Moreover, we refrain from modeling dendritic spine turnover, due to this being a relatively slow process. Instead, we focus on structural modifications in existing spines over shorter time-scales, and the phenomenon of intrinsic synaptic noise.

Intrinsic synaptic noise. Over the last decade, experiments involving the observation of exceptionally large numbers of individual dendritic spines have revealed yet another interesting aspect about structural plasticity. The size of a spine exhibits constant, state-dependent fluctuations, even over such small time windows as 10 minutes. (Yasumatsu et al., 2008; Loewenstein et al., 2011; Statman et al., 2014; Ishii et al., 2018). Surprisingly, these fluctuations have been found to only partially be caused by neural activity. Instead, they are, to a large extent, driven by internal, activity-independent noise sources, and can be seen to persist even as all glutamatergic transmission has been completely silenced (Yasumatsu et al., 2008; Minerbi et al., 2009; Kaufman et al., 2012; Fisher-Lavie & Ziv, 2013; Hazan & Ziv, 2020). This structural volatility is directly reflective of a functional volatility, given that most morphological metrics of spine size, such as spine head volume, head area, and post-synaptic density size, are strongly correlated with each other (Arellano et al., 2007) and highly predictive of synaptic conductance (Holler et al., 2021).

Most theoretical work on this topic has taken a descriptive approach, whereby synaptic change is assumed to behave as a random walk with additive and multiplicative components (Yasumatsu et al., 2008; Loewenstein et al., 2011; Statman et al., 2014; Hazan & Ziv,

[4] Specifically REM-sleep.

2020; Dorkenwald et al., 2022). These noise sources are, in turn, assumed to be caused by a mixture of different plasticity mechanisms that are governed by the dynamics of ongoing neural activity (Zheng et al., 2013). This view has primarily been motivated by the seemingly linear relationship between spine size and fluctuation amplitude (Loewenstein et al., 2011; Statman et al., 2014), as well as by the ubiquity of log-normal distributions that has been found for various proxies of synaptic strength (Song et al., 2005; Loewenstein et al., 2011; Dorkenwald et al., 2022). In chapter 3, we contest this description of structural fluctuations, and propose a different way of modeling internal synaptic noise.

1.3 Scope and terminology

As this thesis is based solely on theoretical and computational work, it employs a neuroscientific terminology that is somewhat simpler and more abstract than that typically found in the experimental literature. We therefore clarify some of our definitions below.

First, while we make a distinction between a *synapse* and an entire inter-neuronal *connection* when reviewing biological data, we use these terms synonymously in theoretical discussions, as we do not include multi-synaptic connections in any of the models.

Second, we use the term *synaptic plasticity* in its broadest sense, encompassing all processes that alter the functional state of a synapse, whether it be through anatomical, biochemical, or biophysical means. Synaptic plasticity is divided into two different subcategories, based on the form of expression: *functional plasticity*, meaning long-term potentiation (LTP) and depression (LTD) in the efficacy of existing synapses, and *structural plasticity*, which we will use to refer to purely anatomical synaptic changes, and, in particular, synaptic formation and retraction. Note, therefore, that we will also use the term structural plasticity, for lack of a better phrase, when discussing changes in dendritic spine anatomy, even if no formation or pruning occurs.

The term *consolidation* will be used to refer to any general process that, following initial memory encoding, strengthens or stabilizes a memory trace or engram through additional plasticity. Although the term consolidation sometimes is used to refer to optimal memory encoding and decay across multiple sessions or datasets, for example in the context of continual, sequential, or lifelong learning, this definition is not related to our work, as we only treat the problem of consolidating a single set of memories, over, at most, a single day.

In terms of the temporal specification of memory, we restrict our work to *long-term* memory and plasticity. Transient learning processes, such as short-term plasticity and working memory, are outside the scope of the thesis.

When modeling circuits of neurons, we will typically be working with recurrent artificial neural networks with binary activation functions. Although we never explicitly specify the type or modality of the information stored in these models, they are perhaps best understood as models of declarative memory, and, in particular, visual or semantic memory, where each pattern of neural activity represents the embedding of an image, symbol, word, or linguistic morpheme. It is, with this interpretation in mind, easier to intuitively grasp notions such as

pattern similarity, completion, and distortion, as well as memory strength and vividness.

1.4 Summary of the thesis

The work presented in the next two chapters is centered on the following questions:

i. What does it mean to optimally encode or consolidate a memory, and how can we mathematically define this concept within a neural network framework?

ii. How can we mathematically formalize the concept of synaptic pruning within the context of memory encoding and consolidation?

iii. Can we derive a biologically plausible synaptic learning rule for consolidating memory and/or pruning synapses in a neural network?

iv. What is the relationship between consolidation, synaptic pruning, homeostatic scaling, and multiplicative synaptic plasticity, and is there some way of reconciling these seemingly disparate aspects of memory formation and learning?

Chapter 2. The first question is treated in chapter 2 and our work on *kernel memory networks*. We begin our analysis by considering each individual neuron in a neural network as a classifier endowed with a kernel function that pre-processes and transforms the synaptic input before it reaches the neuron. Within this mathematical framework, we define the problem of optimal storage, or consolidation, as the maximization of the signal-to-noise ratio of all associations learned by each neuron (Fig. 1.1, blue boxes).

Using well-established theoretical results from the literature on kernel methods, we are able to derive closed-form solutions for the optimal weights and state-update rules for both hetero- and auto-associative networks. Interestingly, we find that this family of models, which we term *kernel memory networks*, generalize many well-known memory models, such as classical and modern Hopfield networks, as well as the Kanerva network (also known as the sparse distributed memory). Our kernel-based approach offers a simple and intuitive understanding of how the classical Hopfield learning rule produces sub-optimal storage, and why non-linear Hamiltonians and higher-order synaptic interactions enhance the capacity. We are also able to extend and generalize previous results on iterative learning rules and derive closed-form solutions for Kanerva networks of very large size.

In the second half of the chapter, we apply the kernel-based approach to the case of continuous patterns, and design a simple example of an attractor network with exponential storage capacity. We compare this to a particular variant of the modern Hopfield network which is closely related to the attention mechanism in transformers.

Finally, we comment on the biological relevance of our results by highlighting fact that kernel classifiers generalize many previously published two-stage neuron models with separate, non-linear dendritic and somatic processing.

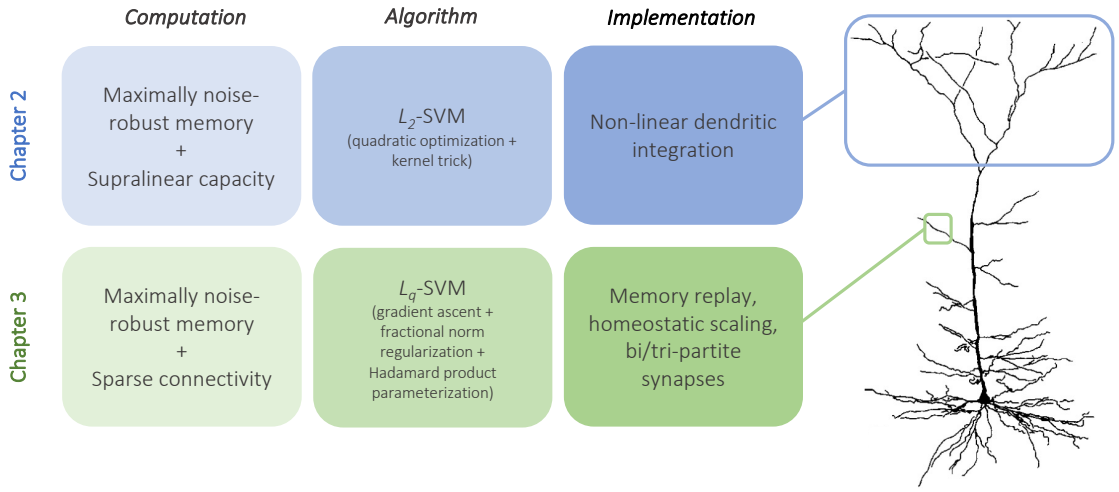


Figure 1.1: An overview of the thesis organized according to David Marr's three levels of analysis. The results in chapter 2 (blue boxes) are primarily closed-form solutions, with biological interpretations focusing on dendritic processing. The solutions in chapter 3 (green boxes) cannot be expressed in closed form; instead, we analyze iterative synaptic learning rules and dendritic spine dynamics. SVM stands for support vector machine. Pyramidal cell reconstruction adapted from Chen et al. (2003).

The main contributions of this work are two-fold. First, from the perspective of neuroscience and memory modeling, we derive a general, compact expression for the structure of optimal hetero- and auto-associative memory networks based on a simple, intuitive definition of engram strength. This directly establishes a link between neuron models with non-linear dendrites and the storage capacity of a memory circuit. It also demonstrates that most models of memory and active dendritic processing fundamentally belong to a single class, and that they differ only with respect to two properties: model complexity, which depends on the dendritic function, and model precision, which is determined by the level of fine-tuning of the synaptic weights.

Secondly, from a machine learning perspective, we clarify the mathematical commonalities between kernel methods, memory network models, and their relation to the dot-product attention mechanism. We also demonstrate that memory networks can be derived from first principles, out of an assumption of optimal storage, without the need to assign a Hamiltonian to the system.

Chapter 3. The results on kernel memory networks naturally set the stage for chapter 3, where we tackle questions (ii), (iii), and (iv). In order to extend our previous definition of consolidation to also include synaptic pruning, we regularize the optimization problem of maximizing memory SNR so that the solution becomes sparse. To achieve this without the need to invoke unrealistic weight scaling or manual thresholding, we employ a recently developed machine learning method for implicit regularization. By reparameterizing each synaptic weight as a product of multiple sub-synaptic components, we are able to derive a synaptic learning rule that performs sparse consolidation in a way that naturally incorporates features like memory replay, multiplicative homeostatic scaling, and state-dependent homosynaptic

plasticity (Fig. 1.1, green boxes).

An important property of our learning rule is that intrinsic synaptic noise scales sub-linearly with the weight. To test this prediction, we perform a meta-analysis of several published datasets on the volatility of dendritic spines and find that spine size fluctuations, indeed, scale sublinearly with size.

Our simulations of consolidated attractor networks at different levels of sparsity indicate that there exists a trade-off between memory robustness and synaptic pruning, and suggests that an optimal compromise can be found when each synapse consists of only two to three plasticity expression sites. This configuration maximizes the amount of retrievable information per synapse. In the second half of the chapter, we demonstrate how such an optimum can be reached in a cortical circuit by implementing our learning rule in a network that encodes and consolidates memory across wakefulness and sleep. We compare the results of our sleep-based consolidation algorithm with human behavioral data and discuss its merits relative other theories on the function of sleep.

The most significant contribution of this work is arguably the idea of leveraging the internal complexity of synapses to implicitly bias a consolidation algorithm to find sparse solutions. The learning rule that emerges from our derivation naturally reconciles the seemingly contradictory notions of sparsification with multiplicative plasticity and homeostatic scaling. Our approach is also compatible with other models of internal synaptic machinery, such as the tagging-and-capture model and the cascade model for optimal memory decay.

A noteworthy implication of our definition of pruning is the fact that it predicts that synaptic density changes over the course of development in a way that, we argue, better explains the dependence on environmental enrichment, compared to previous models. We discuss this point in greater detail towards the end of the chapter.

From the perspective of statistical physics and the classical literature on attractor networks, it is important to note that kernel memory networks, which are finite-sized networks at maximal robustness, have a correspondence, in the mean-field limit, to saturated networks with a weight scaling of $\mathcal{O}(1/\sqrt{N})$. They are therefore, historically speaking, *first-generation* memory models. In contrast, the optimally pruned networks can be described, somewhat informally, as corresponding to saturation in the mean-field limit under the scaling $\mathcal{O}(1/N^q)$, where $q \geq \frac{1}{2}$. As such, these models subsume the $\mathcal{O}(1/N)$ case, and can therefore be seen as generalized *second-generation* models.

Thesis structure. The writing style in chapters 2 and 3 is oriented to different audiences. Chapter 2 is based on a manuscript published in a machine learning journal, and is therefore more focused on mathematical results, with biological interpretations being secondary. Chapter 3, on the other hand, is an adaptation of a manuscript currently being prepared for submission to a neuroscience journal. The main text therefore contains fewer equations, and instead focuses on key concepts, simulation results, and analysis of experimental data.

Chapter 2

Kernel memory networks: a unifying framework

This chapter is based on the following article:

> "Kernel memory networks: A unifying framework for memory modeling"
> **Georgios Iatropoulos**, Johanni Brea*, Wulfram Gerstner*
> *Advances in Neural Information Processing Systems* 35 (2022)

Abstract. We consider the problem of training a neural network to store a set of patterns with maximal noise robustness. A solution, in terms of optimal weights and state update rules, is derived by training each individual neuron to perform either kernel classification or interpolation with a minimum weight norm. By applying this method to feed-forward and recurrent networks, we derive optimal models, termed kernel memory networks, that include, as special cases, many of the hetero- and auto-associative memory models that have been proposed over the past years, such as modern Hopfield networks and Kanerva's sparse distributed memory. We modify Kanerva's model and demonstrate a simple way to design a kernel memory network that can store an exponential number of continuous-valued patterns with a finite basin of attraction. The framework of kernel memory networks offers a simple and intuitive way to understand the storage capacity of previous memory models, and allows for new biological interpretations in terms of dendritic non-linearities and synaptic cross-talk.

Author contributions. GI created the model and produced the theoretical results. JB assisted in writing the proofs. GI and JB performed the simulations. GI, JB, and WG wrote the article.

*JB and WG were co-senior authors.

Acknowledgements. This study was supported by funding from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology, to the Blue Brain Project, a research center of the École Polytechnique Fédérale de Lausanne (EPFL).

2.1 Introduction

Although the classical work on attractor neural networks reached its peak in the late 1980's, with the publication of a number of seminal works (e.g., Hopfield, 1982; Amit et al., 1985; Gardner, 1987a, 1988), recent years have seen a renewed interest in the topic, motivated by the popularity of the attention mechanism (Vaswani et al., 2017), external memory-augmented neural networks (Graves et al., 2014; Weston et al., 2015), as well as a new generation of energy-based attractor networks models, termed modern Hopfield networks (MHNs), capable of vastly increased memory storage (Krotov & Hopfield, 2016; Demircigil et al., 2017). Recent efforts to understand the theoretical foundation of the attention mechanism have, in fact, shown that it can be linked to Hopfield networks (Krotov & Hopfield, 2020; Ramsauer et al., 2021), but also to Kanerva's sparse distributed memory (SDM) (Kanerva, 1988; Bricken & Pehlevan, 2021), and to the field of kernel machines (Tsai et al., 2019; Wright & Gonzalez, 2021). The last connection is particularly intriguing, in light of the many theoretical commonalities between neural networks and kernel methods (Neal, 1996; Williams, 1996; Cho & Saul, 2009; Jacot et al., 2018; Chen & Xu, 2020). Overall, these results suggest that a unified view can offer new insights into memory modeling and new tools for leveraging memory in machine learning.

In this work, we aim to clarify some of the overlap between the fields of memory modeling and statistical learning, by integrating and formalizing a set of theoretical connections between Hopfield networks, the SDM, kernel machines, and neuron models with non-linear dendritic processing.

2.1.1 Our contributions

First, we derive a set of normative kernel-based models that describe the general mathematical structure of feed-forward (i.e., hetero-associative) and recurrent (i.e., auto-associative) memory networks that can perform error-free recall of a given set of patterns with maximal robustness to noise.

Second, we show that the normative models include, as special cases, the classical and modern Hopfield network, as well as the SDM.

Third, we derive a simple attractor network model for storing an exponential number of continuous-valued patterns with a finite basin of attraction. We discuss its similarity to attention.

Finally, we explain how classifiers with non-linear kernels can be interpreted as general forms of neuron models with non-linear dendritic activation functions and synaptic cross-talk.

2.1.2 Related work

Our work is primarily related to the studies by Casali et al. (2006), Krotov & Hopfield (2016, 2020), Bricken & Pehlevan (2021), Ramsauer et al. (2021), and Millidge et al. (2022). While MHNs are extensively analyzed by Krotov & Hopfield (2016, 2020), Ramsauer et al.

(2021), and Millidge et al. (2022), the approach is energy-based and makes no statements about the relation between MHNs and kernel methods; a brief comment by Ramsauer et al. (2021) mentions some similarity to SVMs, but this is not further explained. The work by Bricken & Pehlevan (2021) focuses on the SDM and its connection to attention. It observes that the classical Hopfield network is a special case of the SDM, but no further generalization is made, and kernel methods are not mentioned. In our work, we place MHNs and the SDM in a broader theoretical context by showing that *both* models are special suboptimal cases of a family of memory networks that can be derived with a normative kernel-based approach.

2.2 Background

Consider the following simple model of hetero-associative memory: a single-layer feed-forward network consisting of a single output neuron connected to N_{in} inputs with the weights $\boldsymbol{w} \in \mathbb{R}^{N_\phi}$. The output $s_{\text{out}} \in \{\pm 1\}$ is given by

$$s_{\text{out}} = \text{sgn}\left[\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{s}_{\text{in}}) - \theta\right] \tag{2.1}$$

where $\mathbf{s}_{\text{in}}$ is the input vector (also called query), θ the threshold, and $\boldsymbol{\phi}$ a function that maps the "raw" input to a N_ϕ-dimensional feature space, where typically $N_\phi \gg N_{\text{in}}$. Suppose that we are given a set of M input-output patterns $\{\boldsymbol{\xi}^\mu_{\text{in}}, \xi^\mu_{\text{out}}\}_{\mu=1}^M$, in which every entry ξ is randomly drawn from $\{\pm 1\}$ with sparseness $f := \mathbb{P}(\xi = 1)$. In order for the neuron to store the patterns in a way that maximizes the amount of noise it can tolerate while still being able to recall all patterns without errors, one needs to find the weights that produce the output ξ^μ_{out} in response to the input $\boldsymbol{\xi}^\mu_{\text{in}}$, $\forall \mu$, and that maximize the smallest Euclidean distance between the inputs and the neuron's decision boundary. Using Gardner's formalism (Gardner, 1987a, 1988), this problem can be expressed as

$$\begin{aligned} \arg\max_{\mathbf{w}} \ \kappa \quad \text{s.t.} \quad & \xi^\mu_{\text{out}} \left(\mathbf{w}^\top \boldsymbol{\phi}(\boldsymbol{\xi}^\mu_{\text{in}}) - \theta\right) \geq \kappa, \ \forall \mu \\ & \|\mathbf{w}\|_2 = \bar{w} \end{aligned} \tag{2.2}$$

where $\bar{w} > 0$ is a constant. This is equivalent to solving

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2 \quad \text{s.t.} \quad \xi^\mu_{\text{out}} \left(\mathbf{w}^\top \boldsymbol{\phi}(\boldsymbol{\xi}^\mu_{\text{in}}) - \theta\right) \geq 1, \ \forall \mu \tag{2.3}$$

which can be directly identified as the support vector machine (SVM) problem for separable data (Cortes & Vapnik, 1995). The solution to Eq. 2.3 can today be found in any textbook on basic machine learning methods, and yields an optimal output rule that can be written in a *feature* and *kernel* form

$$s_{\text{out}} = \text{sgn}\left[\sum_\mu^M \alpha^\mu \xi^\mu_{\text{out}} \boldsymbol{\phi}(\boldsymbol{\xi}^\mu_{\text{in}})^\top \boldsymbol{\phi}(\mathbf{s}_{\text{in}}) - \theta\right] \qquad \text{(feature form)} \tag{2.4}$$

$$= \text{sgn}\left[\sum_\mu^M \alpha^\mu \xi^\mu_{\text{out}} K(\boldsymbol{\xi}^\mu_{\text{in}}, \mathbf{s}_{\text{in}}) - \theta\right] \qquad \text{(kernel form)} \tag{2.5}$$

where we, in the latter expression, have used the "kernel-trick" $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_j)$. The solution depends on the Lagrange coefficients $\alpha^\mu \geq 0$, many of which are typically zero. Patterns with $\alpha^\mu > 0$ are called *support vectors*.

2.3 Kernel memory networks for binary patterns

2.3.1 Hetero-associative memory as a feed-forward SVM network

We begin by considering a hetero-associative memory network with an arbitrary number N_{out} output neurons, whose combined state we denote $\mathbf{s}_{\text{out}}$. In order for the network as a whole to be able to tolerate a maximal level of noise and still successfully recall its stored memories, we solve Eq. 2.3 for each neuron independently. As each neuron can have a different classification boundary along with a different set of support vectors, its weights will, in general, be characterized by an independent set of M Lagrange coefficients. To simplify the notation, we represent these coefficients α_i^μ, across neurons i and patterns μ, as entries in the matrix $\mathbf{A}$, where $(\mathbf{A})_{i\mu} = \alpha_i^\mu$. We also combine all thresholds in the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{N_{\text{out}}})$, and all input and output patterns as columns in the matrices $\mathbf{X}_{\text{in}} = (\boldsymbol{\xi}_{\text{in}}^1, \ldots, \boldsymbol{\xi}_{\text{in}}^M)$ and $\mathbf{X}_{\text{out}} = (\boldsymbol{\xi}_{\text{out}}^1, \ldots, \boldsymbol{\xi}_{\text{out}}^M)$. Finally, we assume that all neurons have the same feature map, so that $\boldsymbol{\phi}_i = \boldsymbol{\phi}$, $\forall i$ (see Fig. 2.1). All functions are applied column-wise when the argument is a matrix, for example $\boldsymbol{\phi}(\mathbf{X}_{\text{in}}) = (\boldsymbol{\phi}(\boldsymbol{\xi}_{\text{in}}^1), \ldots, \boldsymbol{\phi}(\boldsymbol{\xi}_{\text{in}}^M))$. The optimal response of the network can now be compactly summarized as follows.

Property 1 (Robust hetero-associative memory network)**.** *A single-layer hetero-associative memory network trained to recall the patterns* $\mathbf{X}_{\text{out}}$ *in response to the inputs* $\mathbf{X}_{\text{in}}$ *with maximal noise robustness, has an optimal output rule that can be written as*

$$\mathbf{s}_{\text{out}} = \operatorname{sgn}\left[(\mathbf{A} \odot \mathbf{X}_{\text{out}})\boldsymbol{\phi}(\mathbf{X}_{\text{in}})^\top \boldsymbol{\phi}(\mathbf{s}_{\text{in}}) - \boldsymbol{\theta}\right] \qquad \textit{(feature form)} \tag{2.6}$$

$$= \operatorname{sgn}\left[(\mathbf{A} \odot \mathbf{X}_{\text{out}})K(\mathbf{X}_{\text{in}}, \mathbf{s}_{\text{in}}) - \boldsymbol{\theta}\right] \qquad \textit{(kernel form)} \tag{2.7}$$

where $\odot$ *denotes the Hadamard product.*

2.3.2 Auto-associative memory as a recurrent SVM network

The hetero-associative network can be made auto-associative by setting $N_{\text{out}} = N_{\text{in}}$ and $\mathbf{X}_{\text{out}} = \mathbf{X}_{\text{in}}$. The network is now effectively recurrent, as each neuron can serve both as an input and output simultaneously (see Fig. 2.1). Consider a recurrent network with N neurons, whose state at time point t is denoted $\mathbf{s}^{(t)} \in \{\pm 1\}^N$, and whose dynamics evolve according to the update rule

$$s_i^{(t+1)} = \operatorname{sgn}\left[\mathbf{w}_i^\top \boldsymbol{\phi}(\mathbf{s}^{(t)}) - \theta_i\right] \tag{2.8}$$

where $\mathbf{w}_i \in \mathbb{R}^{N_\phi}$ is the weight vector to neuron $i = 1, \ldots, N$. In order to make the patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^M$ fixed points of the network dynamics, we train each neuron i independently on every pattern μ to, again, produce the response ξ_i^μ when the rest of the network is initialized in $\boldsymbol{\xi}^\mu$. Moreover, we maximize the amount of noise that can be tolerated by the network

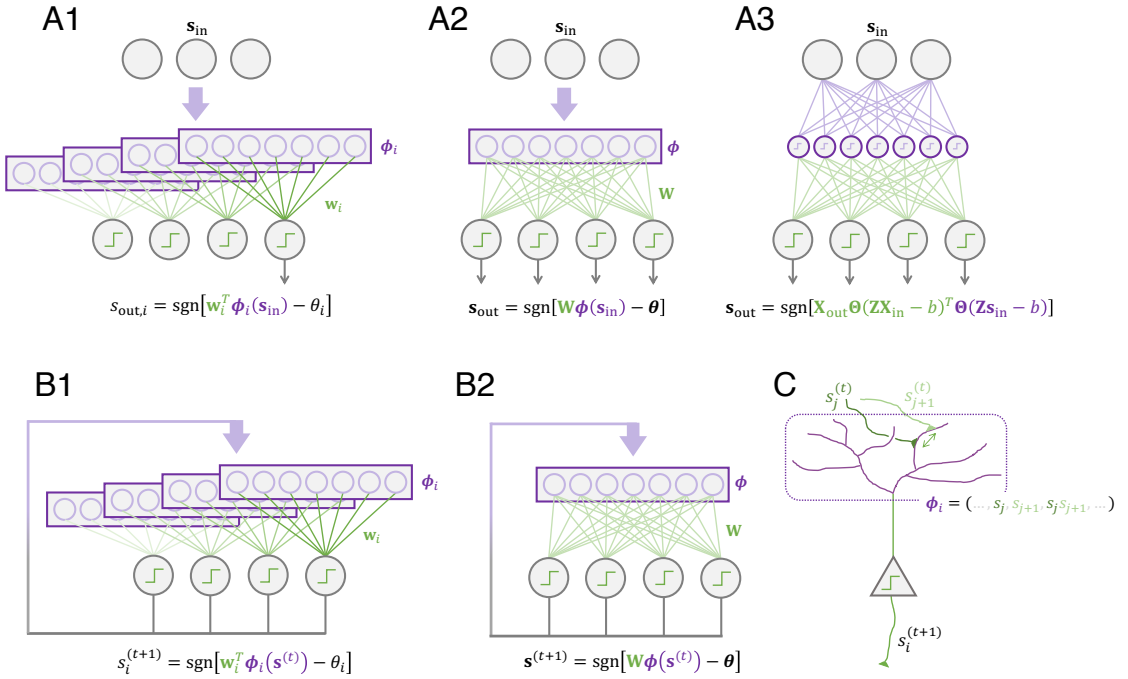


Figure 2.1: Graphical representation of (A1-A2) the feed-forward SVM network, (A3) the SDM, (B1-B2) the recurrent SVM network, and (C) an SVM mapped to the anatomy of a pyramidal cell (see Sec. 2.5). Circles represent neurons, while boxes represent the input transformation by the feature map $\boldsymbol{\phi}$, which can be dependent (A1, B1) or independent (A2, B2) of neuron index i.

while maintaining error-free recall by maximizing the smallest Euclidean distance between each neuron's decision boundary and its inputs. This maximizes the size of the attractor basins (Forrest, 1988; Kepler & Abbott, 1988). The problem of training the entire network is, in this way, transformed into the problem of training N separate classifiers according to

$$\min_{\mathbf{w}_i} \|\mathbf{w}_i\|_2 \quad \text{s.t.} \quad \xi_i^\mu \left(\mathbf{w}_i^\top \boldsymbol{\phi}(\boldsymbol{\xi}^\mu) - \theta_i\right) \geq 1, \ \forall \mu, i \ . \tag{2.9}$$

The solution can be obtained by slightly modifying Property 1, and is stated below.

Property 2.1 (Robust auto-associative memory)**.** *A recurrent auto-associative memory network trained to recall the patterns* $\mathbf{X}$ *with maximal noise robustness has an optimal synchronous update rule that can be written as*

$$\mathbf{s}^{(t+1)} = \operatorname{sgn}\left[(\mathbf{A} \odot \mathbf{X})\boldsymbol{\phi}(\mathbf{X})^\top \boldsymbol{\phi}(\mathbf{s}^{(t)}) - \boldsymbol{\theta}\right] \qquad \textit{(feature form)} \tag{2.10}$$

$$= \operatorname{sgn}\left[(\mathbf{A} \odot \mathbf{X})K(\mathbf{X}, \mathbf{s}^{(t)}) - \boldsymbol{\theta}\right] \qquad \textit{(kernel form)} \tag{2.11}$$

Remark. With a linear feature map $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, the optimal update is reduced to

$$\mathbf{s}^{(t+1)} = \operatorname{sgn}\left[(\mathbf{A} \odot \mathbf{X})\mathbf{X}^\top \mathbf{s}^{(t)} - \boldsymbol{\theta}\right] \tag{2.12}$$

where $(\mathbf{A} \odot \mathbf{X})\mathbf{X}^\top$ can be identified as the general form of the optimal weight matrix.

The solution described by Property 2.1 does not, in general, prohibit a neuron from having self-connections. Applying this constraint yields the following result.

Property 2.2 (Robust auto-associative memory without self-connections)**.** *A recurrent auto-associative memory network without self-connections, with the inner-product kernel* $K(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i^\top \mathbf{x}_j)$*, that has been trained to recall the patterns* $\mathbf{X}$ *with maximal noise robustness, has an optimal asynchronous update rule that can be written in the kernel form*

$$s_i^{(t+1)} = \operatorname{sgn}\left[\sum_{\mu}^{M} \alpha_i^\mu \xi_i^\mu k\left(\sum_{j\neq i}^{N} \xi_j^\mu s_j^{(t)}\right) - \theta_i\right] . \tag{2.13}$$

Storage capacity. An intuition for the storage capacity scaling of the hetero- and auto-associative memory networks can be gained by observing that the network as a whole will be able to successfully recall patterns as long as each neuron is able to correctly classify its inputs (or is very unlikely to produce an error). The capacity of the network can thereby be derived from the capacity of each individual neuron. It is well-known that a linear binary classifier can learn to correctly discriminate a maximum of $M_{\text{max}} \approx 2D_{\text{VC}}$ random patterns, where D_{VC} is the Vapnik-Chervonenkis dimension of the classifier (Cover, 1965; Gardner, 1987a; MacKay, 2003, ch. 40). For a neuron with N inputs and a linear feature map $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, this results in $D_{\text{VC}} = N$ and, thus, the capacity $M_{\text{max}} \approx 2N$. Suppose, on the other hand, that the kernel is a homogeneous polynomial of degree p, so that $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j)^p$. In this case, $\boldsymbol{\phi}$ will contain all monomials of degree p composed of the entries in $\mathbf{x}$. As there are $\mathcal{O}(N^p)$ unique p-degree monomials (see Appendix A2.1.1), the input dimensionality and M_{max} will be $\mathcal{O}(N^p)$. For the exponential kernel, which we can write as $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(\mathbf{x}_i^\top \mathbf{x}_j) = \sum_{p=0}^{\infty} (\mathbf{x}_i^\top \mathbf{x}_j)^p / p!$, the dimensionality of $\boldsymbol{\phi}$ will be $\sum_{p=0}^{N} \binom{N}{p} = 2^N$, which yields $M_{\text{max}} \sim \mathcal{O}(e^N)$.

Special cases. In the following sections, we will show that many of the models of hetero- and auto-associative memory that have been proposed in the past are special cases of the solutions in Properties 1, 2.1, and 2.2, characterized by specific choices of $\mathbf{A}$, $\boldsymbol{\phi}$, and K.

2.3.3 The Kanerva network is a feed-forward SVM network

The Kanerva network (Kanerva, 1988), originally referred to as the sparse distributed memory (SDM), is one of the most famous examples of a hetero-associative memory model. It has lately received much attention in the context of generative memory models (Wu et al., 2018) and attention layers in transformers (Bricken & Pehlevan, 2021).

The SDM consists of a register of N_ϕ memory slots, each associated with an address $\mathbf{z}_i \in \{\pm 1\}^{N_{\text{in}}}$, $i = 1, \ldots, N_\phi$. All addresses are listed as rows in the matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{N_\phi})^\top$. The content of each slot is represented by an N_{out}-dimensional vector, initialized at zero. Suppose that we wish to store the M patterns $\mathbf{X}_{\text{out}} = (\boldsymbol{\xi}_{\text{out}}^1, \ldots, \boldsymbol{\xi}_{\text{out}}^M)$ in the addresses $\mathbf{X}_{\text{in}} = (\boldsymbol{\xi}_{\text{in}}^1, \ldots, \boldsymbol{\xi}_{\text{in}}^M)$, where all entries are random and bipolar. The basic idea of the SDM is to write the data to, and later read it from, multiple memory slots at once (hence the distributed storage); this ensures a degree of noise-robustness. In mathematical terms, the

read-out of the SDM provided with a query $\mathbf{s}_{\text{in}}$, is given by

$$\mathbf{s}_{\text{out}} = \text{sgn}\left[\mathbf{X}_{\text{out}}\,\Theta(\mathbf{Z}\mathbf{X}_{\text{in}} - b)^{\top}\,\Theta(\mathbf{Z}\mathbf{s}_{\text{in}} - b)\right] \tag{2.14}$$

where Θ the Heaviside function with bias $b = N_{\text{in}} - 2r$, and r is a parameter that determines the precision of the writing and reading process. Upon comparing Eqs. 2.14 and 2.6, the SDM can be directly identified as a special case of a suboptimal feed-forward SVM network in the feature form, with $\mathbf{A} = \mathbf{1}$, $\boldsymbol{\theta} = \mathbf{0}$, and the feature map $\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}) = \Theta(\mathbf{Z}\mathbf{x} - b)$. When viewed as a kernel method, the function of the SDM is to store the dense addresses $\mathbf{X}_{\text{in}}$ as sparse high-dimensional representations $\boldsymbol{\phi}_{\text{SDM}}$, to make it easier to later determine the slots closest to a query $\mathbf{s}_{\text{in}}$, and retrieve the relevant data.

Capacity. As the SDM is linear in $\boldsymbol{\phi}_{\text{SDM}}$, with $D_{\text{VC}} \approx N_{\phi}$, it follows from the analysis in Sec. 2.3.2 that one should expect the capacity to scale as $M_{\text{max}} \sim \mathcal{O}(N_{\phi})$. Moreover, one should expect a proportionality constant $\sim$0.1, since the SDM is suboptimal relative to the feed-forward SVM network, analogously to how the classical Hopfield network is suboptimal relative to the recurrent SVM network (see Sec. 2.3.4). This is consistent with earlier proofs (Keeler, 1988; Chou, 1989).

Kernel of an infinite SDM. In practice, an SDM with a large number of memory slots N_{ϕ} requires calculations involving a large address matrix $\mathbf{Z}$. This can be avoided by applying the kernel-trick to Eq. 2.14 in the limit $N_{\phi} \to \infty$, which allows for the output to be computed with

$$\mathbf{s}_{\text{out}} = \text{sgn}\left[\mathbf{X}_{\text{out}} K_{\text{SDM}}(\mathbf{X}_{\text{in}}, \mathbf{s}_{\text{in}})\right] \tag{2.15}$$

where we have defined the kernel as

$$K_{\text{SDM}}(\mathbf{x}_i, \mathbf{x}_j) = \lim_{N_{\phi}\to\infty} \frac{\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}_i)^{\top}\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}_j)}{N_{\phi}} \tag{2.16}$$

in order to ensure convergence. In this section, we will derive this kernel for two different variants of the SDM and demonstrate that both are translation-invariant. It is interesting to note here that $\boldsymbol{\phi}_{\text{SDM}}$ is equivalent to a single-layer neural network with N_{ϕ} neurons, weights $\mathbf{Z}$, and bias b. This means that K_{SDM} is equivalent to the kernel of an infinitely wide neural network (Neal, 1996; Williams, 1996; Cho & Saul, 2009).

We begin by noticing that $\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x})$ has a geometrical interpretation (Keeler, 1988; Bricken & Pehlevan, 2021). It is a binary vector that indicates those memory addresses in $\mathbf{Z}$ that differ by at most r bits compared to $\mathbf{x}$. For any two bipolar vectors $\mathbf{z}$ and $\mathbf{x}$, the bit-wise difference can be computed as $\frac{1}{2}|\mathbf{z} - \mathbf{x}| = \frac{1}{4}\|\mathbf{z} - \mathbf{x}\|_2^2$. This means that $\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x})$ indicates all addresses that lie within a sphere centered at $\mathbf{x}$ with radius $2\sqrt{r}$. Consequently, the inner product $\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}_i)^{\top}\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}_j)$ is the number of addresses located in the overlapping volume of two spheres centered at $\mathbf{x}_i$ and $\mathbf{x}_j$. Although an exact calculation of this quantity can be found in the literature (Kanerva, 1988; Bricken & Pehlevan, 2021), its connection to the SDM kernel has, to the best of our knowledge, not previously been made. We therefore modify the previously published expression with a normalization factor $1/2^{N_{\text{in}}}$ and state the

following property.

Property 3.1 (Kernel of an infinite SDM on the hypercube)**.** *In the limit $N_\phi \to \infty$, the kernel of an SDM with N_ϕ memory slots, whose addresses are randomly drawn from $\{\pm 1\}^{N_{\text{in}}}$, is given by*

$$K_{\text{SDM}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2^{N_{\text{in}}}} \sum_{i=N_{\text{in}}-r-\lfloor \frac{\Delta}{2} \rfloor}^{N_{\text{in}}-\Delta} \sum_{j=[N_{\text{in}}-r-i]_+}^{\Delta-(N_{\text{in}}-r-i)} \binom{N_{\text{in}}-\Delta}{i} \cdot \binom{\Delta}{j} \tag{2.17}$$

where r is the bit-wise error threshold and Δ is the bit-wise difference between $\mathbf{x}_i$ and $\mathbf{x}_j$, given by $\Delta = \frac{1}{2}|\mathbf{x}_i - \mathbf{x}_j| = \frac{1}{4}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$.

The SDM can also be implemented with continuous addresses, randomly placed on a unit hypersphere of $(N_{\text{in}} - 1)$ dimensions, denoted $\mathbb{S}^{N_{\text{in}}-1}$. The vector $\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x})$ now indicates all addresses that lie within a hyperspherical cap centered at $\mathbf{x}$ with an angle $\arccos(b)$ between its central axis and the rim. The inner product $\boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}_i)^\top \boldsymbol{\phi}_{\text{SDM}}(\mathbf{x}_j)$ is the number of addresses located in the overlapping area of two spherical caps centered at $\mathbf{x}_i$ and $\mathbf{x}_j$. A calculation of this quantity can, again, be found in the literature (Bricken & Pehlevan, 2021), but has not previously been connected to the kernel of an SDM. We simplify the previously published result and also derive a closed-form approximation, valid for highly sparse $\boldsymbol{\phi}_{\text{SDM}}$ (see Appendix A2.2 for details). The results are summarized below.

Property 3.2 (Kernel of an infinite SDM on the hypersphere)**.** *In the limit $N_\phi \to \infty$, the kernel of an SDM with N_ϕ memory slots, whose addresses are randomly drawn from $\mathbb{S}^{N_{\text{in}}-1}$, is given by*

$$K_{\text{SDM}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{N_{\text{in}} - 2}{2\pi} \int_{\alpha_x}^{\alpha_b} \sin(\varphi)^{N_{\text{in}}-2} B\left[1 - \frac{\tan^2(\alpha_x)}{\tan^2(\varphi)}; \frac{N_{\text{in}} - 2}{2}, \frac{1}{2}\right] \mathrm{d}\varphi \tag{2.18}$$

where $\alpha_x = \frac{1}{2}\arccos(\mathbf{x}_i^\top \mathbf{x}_j)$, $\alpha_b = \arccos(b)$, and B is the incomplete Beta function. In the highly sparse regime, when $0.9 \lesssim b < 1$ and $\frac{1}{N_\phi}\|\boldsymbol{\phi}_{\text{SDM}}\|_0 \ll 1$, the kernel can be approximated with

$$K_{\text{SDM}}(\mathbf{x}_i, \mathbf{x}_j) \approx \frac{\hat{b}^{N_{\text{in}}-1}}{2\pi} B\left[1 - \left(\frac{\Delta}{\hat{b}}\right)^2; \frac{N_{\text{in}}}{2}, \frac{1}{2}\right] \tag{2.19}$$

where $\Delta = \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|_2$ and $\hat{b} = \sin(\arccos(b))$.

In conclusion, an infinitely large SDM with sparse internal representations $\boldsymbol{\phi}_{\text{SDM}}$, can be represented as a suboptimal case of a feed-forward SVM network with a translation-invariant kernel.

2.3.4 The Hopfield network is a recurrent SVM network

The Hopfield network (Hopfield, 1982) is, arguably, the most well-known model of auto-associative memory. In its modern form (Krotov & Hopfield, 2016), it is a recurrent network

of N neurons with the state $\mathbf{s}^{(t)}$, whose dynamics are governed by the energy

$$E = -\sum_{\mu}^{M} F\left(\sum_{i}^{N} \xi_i^\mu s_i^{(t)}\right) \tag{2.20}$$

and state update rule

$$s_i^{(t+1)} = \text{sgn}\left[\sum_{\mu}^{M} \xi_i^\mu F'\left(\sum_{j\neq i}^{N} \xi_j^\mu s_j^{(t)}\right)\right] \tag{2.21}$$

where F is a smooth function, typically a sigmoid, polynomial, or exponential. This "generalized" Hopfield model has a long history (see, e.g., Hintzman, 1984; Lee et al., 1986; Abbott & Arian, 1987; Gardner, 1987b) but has received renewed attention in recent years under the name *modern Hopfield network* (MHN) or *dense associative memory* (Krotov & Hopfield, 2016; Demircigil et al., 2017). By comparing Eq. 2.21 with Eq. 2.13, the state update of the MHN can be identified as a special case of a suboptimal recurrent SVM network in the kernel form, with $k = F'$, $\mathbf{A} = \mathbf{1}$, and $\boldsymbol{\theta} = \mathbf{0}$ (since $f = 0.5$). With a linear $F'(x) = x$, the MHN reduces to the classical Hopfield network, which is a special case of the recurrent SVM network with the linear kernel $k(\mathbf{x}_i^\top \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$.

Capacity. The storage capacity of the MHN has been shown to depend on the shape of F'. In the linear case, the capacity is famously limited to $\sim 0.1N$ patterns, depending on the precision of retrieval (Amit et al., 1985; McEliece et al., 1987). If, on the other hand, F' is polynomial with degree p, the capacity scales as $M_{\text{max}} \sim \mathcal{O}(N^p)$ (Krotov & Hopfield, 2016), while an exponential F' endows the network with a capacity $M_{\text{max}} \sim \mathcal{O}(e^N)$ (Demircigil et al., 2017). From the perspective of the kernel memory framework, this scaling directly follows from the analysis in Sec. 2.3.2 with $k = F'$.

In the regime of low errors, the kernel memory framework can also be used to derive a more precise capacity scaling for the classical Hopfield network. We first note that any one-shot learning rule that implies $\mathbf{A} > 0$ is equivalent to an SVM network where every stored pattern is a support vector. Such a heuristic is only likely to be close to the optimal solution and perform well in large networks with very few patterns, as high-dimensional linear SVMs trained on few patterns are highly likely to find solutions where all patterns are support vectors; this effect has been termed *support vector proliferation* (Ardeshir et al., 2021). Restricting the network to this regime limits the capacity to $M_{\text{max}} \sim \mathcal{O}(\frac{N}{2 \log N})$, consistent with the result by McEliece et al. (1987) (see Appendix A2.1.2).

Iterative learning rules. The problem of iteratively training MHNs with biologically plausible online learning rules has recently been studied (Tyulmankov et al., 2021), with a resulting storage capacity ranging from $\sim 0.16N$ to $\sim N$, depending on the exact implementation. The aim, in general, of such studies is to find a learning rule capable of producing a capacity close to the theoretical maximum $\sim 2N$. For this purpose, the perspective of kernel memory networks can be particularly helpful, as many of the algorithms that have been developed over the past two decades to optimize SVMs can be utilized for MHNs as well. For example,

a network formulated in the feature form can be trained with the stochastic batch perceptron rule (Krauth & Mezard, 1987; Cotter et al., 2012), the passive aggressive rules (Crammer et al., 2006), the minnorm rule (Bansal et al., 2018), as well as with likelihood maximization applied to logistic regression (Soudry et al., 2018; Nacson et al., 2019; Ji et al., 2021). In the kernel form, two of the most well-known online algorithms for training linear and non-linear SVMs are the Adatron (Anlauf & Biehl, 1989) and the Kernel-Adatron (Frieß et al., 1998). A performance comparison between iterative learning and the modern Hopfield learning rule can be found in Appendix A2.3.

Generalization. Viewing the MHN as a recurrent network of SVMs can also facilitate a more intuitive understanding of its ability to generalize, when used as a conventional classifier. In this setting, one designates a subset of the neurons as input units, and the remaining neurons as outputs. Given a set of input-output associations, one optimizes the memory patterns $\boldsymbol{\xi}^\mu$ using, for example, gradient descent. Such an experiment was performed by Krotov & Hopfield (2016) on the MNIST data set, using a polynomial non-linearity $F(x) = x^p$. Results showed that the test error first improved as p increased from 2 to 3, but later deteriorated for high degrees, like $p = 20$. While it may be difficult to explain this behavior within an energy-based framework, it is entirely expected when viewed from the SVM perspective: a kernel of low polynomial degree has too few degrees of freedom to fit the classification boundary in the training set, causing *underfitting*, while a polynomial of too high degree grants the model too much flexibility, which results in *overfitting*.

The pseudoinverse learning rule. The coefficients in $\mathbf{A}$ are, in general, computed numerically, and cannot be written in closed form. However, in the special case when Eq. 2.9 is underdetermined, meaning $M < N_\phi$, a closed-form (but suboptimal) solution can be obtained using the *least-squares SVM* method (Suykens & Vandewalle, 1999). The result is a generalized form of the *pseudoinverse learning rule* (Personnaz et al., 1986). See Appendix A2.4 for details.

2.4 Kernel memory networks for continuous patterns

2.4.1 Auto-associative memory as a recurrent interpolation network

So far, we have considered memory models designed to store only bipolar patterns. We now relax this constraint and allow patterns to be continuous-valued. We first observe that any set of patterns $\mathbf{X} \in \mathbb{R}^{N \times M}$ can be made fixed points of the dynamics by training each neuron i to interpolate ξ_i^μ when the rest of the network is initialized in $\boldsymbol{\xi}^\mu$, for every pattern μ. Assuming that the model is equipped with a kernel that allows for each fixed point to also be attracting, we can ensure that a lower bounding estimate of the size of the attractor basin is maximized by finding the interpolation with minimum weight norm (see Appendix A2.5.1 for proof). These results are summarized below.

Property 4 (Robust auto-associative memory with continuous patterns)**.** *Suppose that the dynamics of a recurrent auto-associative memory network evolve according to the*

synchronous update rule

$$\mathbf{s}^{(t+1)} = \mathbf{X}\mathbf{K}^{\dagger}K(\mathbf{X}, \mathbf{s}^{(t)}) \tag{2.22}$$

where $\mathbf{K} = K(\mathbf{X}, \mathbf{X}) = \boldsymbol{\phi}(\mathbf{X})^{\top}\boldsymbol{\phi}(\mathbf{X})$ *is the kernel matrix and* $\mathbf{K}^{\dagger}$ *its Moore-Penrose pseudoinverse, where* $\mathbf{K}^{\dagger} = \mathbf{K}^{-1}$ *if* $\boldsymbol{\phi}(\mathbf{X})$ *is full column rank. Then, the dynamics of the network is guaranteed to have the fixed points* $\mathbf{X}$*. Moreover, if the points are attracting, Eq. 2.22 maximizes a lower bound of the attractor basin sizes.*

2.4.2 A recurrent interpolation network with exponential capacity

Memory models for continuous data (e.g., Hopfield, 1984; Koiran, 1994; Nowicki & Siegelmann, 2010) have generally received less attention than their binary counterparts. Recently, however, Ramsauer et al. (2021) proposed an energy-based model capable of storing an exponential number of continuous-valued patterns (we will refer to this model as the softmax network). While the structure of this model is similar to Eq. 2.22, it cannot be analyzed within the framework of Property 4, as it involves a kernel that is neither symmetric nor positive-definite (Wright & Gonzalez, 2021).

Nonetheless, we will in this section demonstrate that it is possible to use conventional kernel methods to design an attractor network with exponential capacity for continuous patterns. We utilize the properties of the SDM by using a translation-invariant kernel with a fixed spatial scale r. For the sake of simplicity, we choose the exponential power kernel (EPK)

$$K_{\mathrm{EPK}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\left(\frac{1}{r}\|\mathbf{x}_i - \mathbf{x}_j\|_2\right)^{\beta}\right] \tag{2.23}$$

where $\beta, r > 0$. These parameters determine the shape of the attractor basin that surrounds each pattern. While r roughly sets the radius of attraction, β represents an inverse temperature which changes the steepness of the boundary of the attractor basin. Moreover, as long as the patterns are unique, the kernel matrix is invertible and we have $\mathbf{K}_{\mathrm{EPK}}^{\dagger} = \mathbf{K}_{\mathrm{EPK}}^{-1}$ (Micchelli, 1986).

We will now analyze the noise robustness and storage capacity of this model. To make the analysis tractable, we will operate in the regime of low temperatures, meaning the limit $\beta \to \infty$. We first establish the following three properties.

Property 5.1 (EPK network at zero temperature)**.** *Given a set of unique patterns* $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1}^{M}$ *with* $\min_{\mu,\nu\neq\mu}\|\boldsymbol{\xi}^{\mu} - \boldsymbol{\xi}^{\nu}\|_2 > r$*, the state update rule for the EPK network at* $\beta \to \infty$ *reduces to*

$$\mathbf{s}^{(t+1)} = \mathbf{X}\,\Theta(r - \|\mathbf{X} - \mathbf{s}^{(t)}\|_2) \tag{2.24}$$

where $\Theta(\cdot)$ *is the Heaviside function with* $\Theta(0) = e^{-1}$ *(see Appendix A2.5.2).*

Remark. In geometrical terms, Property 5.1 states that the boundary of the basin of attraction surrounding each pattern becomes a sharp $(N-1)$-dimensional hypersphere with radius r in the limit $\beta \to \infty$. For lower, finite β, the spherical boundary becomes increasingly fuzzy. From the perspective of an energy landscape, each pattern lies in an N-dimensional energy minimum with infinitely steep walls when $\beta \to \infty$. As β is lowered, the barriers

become progressively smoother.

Property 5.2 (Convergence in one step)**.** *Given a set of unique patterns* $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^M$ *with* $\min_{\mu,\nu\neq\mu}\|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2 > 2r$*, the EPK network at* $\beta\to\infty$*, initialized at* $\mathbf{s}^{(0)} = \boldsymbol{\xi}^\mu + \Delta\boldsymbol{\xi}$*, will converge to* $\boldsymbol{\xi}^\mu$ *in one step if* $\|\Delta\boldsymbol{\xi}\|_2 < r$*.*

Property 5.3 (No spurious attractors)**.** *Given a set of unique patterns* $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^M$ *with* $\min_{\mu,\nu\neq\mu}\|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2 > 2r$ *and* $\nexists\mu : \|\boldsymbol{\xi}^\mu\|_2 = r/(1-e^{-1})$*, the only attractors of the dynamics of the EPK network at* $\beta\to\infty$ *are the points* $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^M$*, together with* $\mathbf{0}$ *if* $\nexists\mu : \|\boldsymbol{\xi}^\mu\|_2 \leq r$*.*

Remark. Properties 5.2 and 5.3 can be shown to be true simply by inserting the expression $\mathbf{s}^{(0)} = \boldsymbol{\xi}^\mu + \Delta\boldsymbol{\xi}$ in Eq. 2.24. Assuming no overlaps between the basins of attraction, a quick calculation shows that $\mathbf{s}^{(1)} = \boldsymbol{\xi}^\mu$ if $\|\Delta\boldsymbol{\xi}\|_2 < r$. If, on the other hand, the network is initialized such that $\|\mathbf{s}^{(0)} - \boldsymbol{\xi}^\mu\|_2 > r$, $\forall\mu$, one always obtains $\mathbf{s}^{(2)} = \boldsymbol{\xi}^0$, where $\boldsymbol{\xi}^0$ is either $\mathbf{0}$ or the pattern closest to $\mathbf{0}$. In other words, the network recalls a pattern only if the initialization is close enough to it. If located far from *all* patterns, the network assumes an "agnostic" state, represented either by the origin or the pattern closest to the origin (if the origin happens to be located within a basin of attraction).

In the following two properties, we evaluate how the radius of attraction r determines the maximum input noise tolerance and storage capacity.

Property 6 (Robustness to white noise)**.** *Assume that we are given a set of unique patterns* $\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^M \sim \mathcal{N}(\mathbf{0},\mathbf{I}_N)$ *with* $\min_{\mu,\nu\neq\mu}\|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2 > 2r$*, and that the EPK network is initialized in a distorted pattern* $\mathbf{s}^{(0)} = \boldsymbol{\xi}^\mu + \boldsymbol{\epsilon}$*, where* $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_N)$*. Then, at* $\beta\to\infty$*, the maximum noise variance* σ^2_{max} *with which* $\boldsymbol{\xi}^\mu$ *can be recovered in at least 50% of trials is*

$$\sigma^2_{\text{max}} = r^2/N\ . \tag{2.25}$$

Property 7 (Exponential storage capacity)**.** *At* $\beta\to\infty$*, and for* $N\gg 1$*, the average maximum number of patterns sampled from* $\mathcal{N}(\mathbf{0},\mathbf{I}_N)$ *that the EPK network can store and recall without errors is lower-bounded according to*

$$M_{\text{max}} \geq \sqrt{2\sqrt{\pi N}(1-2\sigma^2_{\text{max}})}\exp\left[\frac{N(1-2\sigma^2_{\text{max}})^2}{8}\right] \tag{2.26}$$

where σ^2_{max} *is the maximum white noise variance tolerated by the network.*

Remark. Proofs can be found in Appendix A2.5.2. Note that Property 7 is valid in the range $\sigma^2_{\text{max}} \lesssim 1/2$. While the bounds are fairly tight at the upper end of the range, they become loose when $\sigma^2_{\text{max}}\to 0$. In this limit, which is equivalent to $r\to 0$, the storage capacity tends to infinity, as the risk of interference between patterns vanishes when their radius of attraction becomes infinitesimal.

Comparison to the softmax network. If patterns are randomly placed on a hypersphere instead of being normally distributed, the state update rule in Eq. 2.24 reduces to the form $\mathbf{s}^{(t+1)} = \mathbf{X}\,\Theta(\mathbf{X}^\top\mathbf{s}^{(t)} - \theta)$, where θ is a fixed threshold. While the capacity remains exponential (see Appendix A2.5.3), the basin of attraction surrounding each pattern now forms a spherical cap instead of a ball.

We can compare this to the softmax network at zero temperature, given by $\mathbf{s}^{(t+1)} = \lim_{\beta\to\infty} \mathbf{X}\,\text{softmax}(\beta\mathbf{X}^\top\mathbf{s}^{(t)}) = \mathbf{X}\arg\max(\mathbf{X}^\top\mathbf{s}^{(t)})$. This model differs from the EPK only in a replacement of Θ with arg max. This changes the shape of the attractor basins from spherical caps to Voronoi cells, which parcellate the entire surface of the hypersphere into a Voronoi diagram (see Fig. 2.2). The boundary of each basin is now no longer radially symmetric around a pattern, but instead extends as far as possible in all directions. Consequently, at $\beta \to \infty$, the softmax network has larger attractor basins and always converges to one of the stored patterns, regardless of the initialization point (assuming this is not precisely on a boundary). In contrast, the EPK network may converge to the origin if initialized far from all patterns. This can be interpreted as an agnostic response, which indicates that the model cannot associate the input query with any of its stored patterns.

2.5 Discussion

Biological interpretation. Kernel memory networks can be mapped to the anatomical properties of biological neurons. Consider an individual neuron in the feature form of the recurrent network (Eq. 2.10). The state of neighboring neurons $\mathbf{s}$ is first transformed through $\boldsymbol{\phi}(\mathbf{s})$ and thereafter projected to the neuron through the weight matrix $(\mathbf{A}\odot\mathbf{X})\boldsymbol{\phi}(\mathbf{X})^\top$. When the kernel is polynomial of degree p, so that $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top\mathbf{x}_j + 1)^p$, the transformation $\boldsymbol{\phi}(\mathbf{s})$ consists of all elements in $\mathbf{s}$ and their cross-terms, up to degree p. The input to each neuron, in other words, consists of the states of all other neurons, as well as all possible combinations of their multiplicative interactions. This neuron model can be viewed as a generalized form of, for example, the multiconnected neuron (Peretto & Niez, 1986), the clusteron (Mel, 1991), or the sigma-pi unit (Rumelhart & McClelland, 1986, p. 73). These are all perceptrons that include multiplicative input interactions as a means to model synaptic cross-talk and cluster-sensitivity on non-linear dendrites (Polsky et al., 2004) (see Fig. 2.1).

In the kernel form (Eq. 2.11), each neuron is, again, implicitly comprised of a two-stage process, whereby the raw input $\mathbf{s}$ is first transformed through the function $K(\mathbf{X}, \mathbf{s})$ and then projected through the weight matrix $\mathbf{A}\odot\mathbf{X}$. For any inner-product kernel $K = k(\mathbf{x}_i^\top\mathbf{x}_j)$, this representation can be directly identified as a two-layer neural network, where the hidden layer is defined by the weights $\mathbf{X}$ and the activation function k. This interpretation of the recurrent network was recently proposed by Krotov & Hopfield (2016, 2020) and discussed in relation to hippocampal-cortical interactions involved in memory storage and recall; it is particularly reminiscent of the hippocampal indexing theory (Teyler & Rudy, 2007; Barry & Maguire, 2019).

However, the kernel form can also be viewed as a network in which each *individual* neuron is a generalized form of the two-layered pyramidal cell model (Poirazi & Mel, 2001; Poirazi et al., 2003). This was originally proposed as an abstract neuron model augmented with non-linear dendritic processing (Major et al., 2013). It should be noted, however, that the idea of interpreting kernel methods as neural networks has a longer history, and has been extensively analyzed in the case of, for example, radial basis functions (Poggio & Girosi, 1990a,b). For further details, see Appendix A2.6.

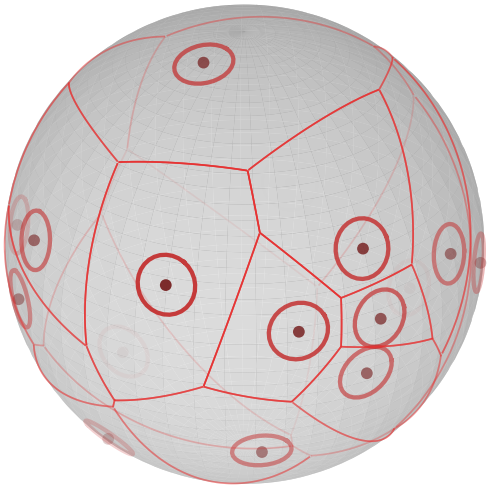


Figure 2.2: Plot of random patterns on $\mathbb{S}^2$ together with attractor basins at $\beta \to \infty$. Dots represent patterns ($M = 17$) while thick and thin red lines correspond to the boundaries of the attractor basins according to the EPK network and the softmax network, respectively. The radius of the circular boundaries has been set to half the minimum pairwise distance between the patterns.

Summary. We have shown that conventional kernel methods can be used to derive the weights for hetero- and auto-associative memory networks storing binary or continuous-valued patterns with maximal noise tolerance. The result is a family of optimal memory models, which we call *kernel memory networks*, which includes the SDM and MHN as special cases. This unifying framework facilitates an intuitive understanding of the storage capacity of memory models and offers new ways to biologically interpret these in terms of non-linear dendritic integration. This work formalizes the links between kernel methods, attractor networks, and models of dendritic processing.

Future work. A unifying theoretical framework for memory modeling can be useful for the development of both improved bio-plausible memory models and for machine learning applications. First, recognizing that there exists algorithms for training optimally noise-robust classifiers and adapting these to biological constraints can aid in the development of normative synaptic three-factor learning rules (Gerstner et al., 2018).

Second, the theoretical link between neuron models, kernel functions, and storage capacity enables one to fit kernel memory networks to neurophysiological data and to analyze the computational properties of biophysically informed memory models.

Finally, our unifying framework reveals that most memory models differ only in the choice of the kernel (model complexity) and the Lagrange parameters (model precision). This categorization simplifies the tailoring of memory models to specific applications, and allows for the design of models whose properties fundamentally can depart from classical networks, by, for example, choosing kernels not associated with a reproducing kernel Hilbert space.

Appendix

A2.1 Derivation of storage capacity scaling

A2.1.1 Optimal storage: The scaling of effective input dimensionality

Suppose that the kernel is a homogeneous polynomial of degree $p \ll N$, meaning $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j)^p$. This implies that the associated feature map $\boldsymbol{\phi}(\mathbf{x})$ contains all monomials of degree p composed of the entries in $\mathbf{x}$. Moreover, given that each entry x_i is ± 1, each monomial in $\boldsymbol{\phi}$ can be written as an interaction term of the form $x_1^{p_1} x_2^{p_2} \cdots x_N^{p_N}$, where $p_i \in \{0, 1\}$ and $\sum_i^N p_i \leq p$ (i.e., no factor x_i has an exponent higher than 1 and the sum of exponents is $\leq p$). The reason for this is that

$$x_i^{n_i} = \begin{cases} 1\,, & \text{if } n_i \text{ even} \\ x_i\,, & \text{if } n_i \text{ odd}\,. \end{cases} \tag{A2.1}$$

The number of unique interaction terms of precisely degree p (the highest degree) is $\binom{N}{p}$. As the binomial coefficient is known to be bounded according to

$$\left(\frac{N}{p}\right)^p \leq \binom{N}{p} \leq \left(\frac{Ne}{p}\right)^p \tag{A2.2}$$

we obtain $\binom{N}{p} \sim \mathcal{O}(N^p)$, for p fixed. Thus, the effective dimensionality of $\boldsymbol{\phi}$ scales like $\mathcal{O}(N^p)$.

For the exponential kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(\mathbf{x}_i^\top \mathbf{x}_j) = \sum_{p=0}^{\infty} (\mathbf{x}_i^\top \mathbf{x}_j)^p / p!$, we first note that the monomials in $\boldsymbol{\phi}$ now will be interaction terms of all degrees $p = 0, \ldots, N$. No monomial of degree $p > N$ will be possible. The total number of unique interaction terms will therefore be

$$\sum_{p=0}^{N} \binom{N}{p} = 2^N = e^{N \log 2} \tag{A2.3}$$

where the first equality can be found in Boros & Moll (2004, p. 14). This gives us an effective dimensionality of $\boldsymbol{\phi}$ that scales like $\mathcal{O}(e^N)$.

A2.1.2 One-shot storage: The scaling of the support vector proliferation regime

As shown recently by Ardeshir et al. (2021), support vector proliferation for an SVM trained on M random patterns drawn uniformly from $\{\pm 1\}^N$ occurs in the regime $N \gtrsim 2M \log M$. Solving for M gives us the scaling

$$M \lesssim \frac{N}{2W_0(\frac{N}{2})} \tag{A2.4}$$

where W_0 is the principal branch of the Lambert function. The largest number of patterns that can be stored in this regime is thus

$$M_{\max} \approx \frac{N}{2W_0(\frac{N}{2})} \,. \tag{A2.5}$$

Using the property $W_0(x) = \log x - \log\log x + o(1)$, we can write

$$W_0\left(\frac{N}{2}\right) \sim \mathcal{O}(\log N) \tag{A2.6}$$

which yields

$$M_{\max} \sim \mathcal{O}\left(\frac{N}{2\log N}\right) \,. \tag{A2.7}$$

A2.2 Kernel of an infinite SDM on the hypersphere

A2.2.1 Derivation of Eq. 2.18.

We follow the same steps as Bricken & Pehlevan (2021), with additional simplifications towards the end. As stated in the main text, we seek to calculate the overlapping area of two hyperspherical caps. A formula for this is provided by Lee & Kim (2014), and can be written as

$$A_\cap = A_\triangledown(R, \alpha_{\min}, \alpha_2) + A_\triangledown(R, \alpha_v - \alpha_{\min}, \alpha_1) \tag{A2.8}$$

where

$$A_\triangledown(R, \alpha_{\min}, \alpha_2) = \frac{\pi^{\frac{N_{\text{in}}-1}{2}}}{\Gamma(\frac{N_{\text{in}}-1}{2})} R^{N_{\text{in}}-1} \int_{\alpha_{\min}}^{\alpha_2} \sin(\varphi)^{N_{\text{in}}-2} I_{1-\frac{\tan^2(\alpha_{\min})}{\tan^2(\varphi)}}\left[\frac{N_{\text{in}}-2}{2}, \frac{1}{2}\right] \mathrm{d}\varphi \tag{A2.9}$$

where R is the radius, I is the regularized incomplete Beta function, and

$$\alpha_1 = \alpha_2 = \arccos(b) \tag{A2.10}$$

$$\alpha_v = \arccos(\mathbf{x}_i^\top \mathbf{x}_j) \tag{A2.11}$$

$$\alpha_{\min} = \arctan\left(\frac{\cos(\alpha_1)}{\cos(\alpha_2)\sin(\alpha_v)} - \frac{1}{\tan(\alpha_v)}\right) \tag{A2.12}$$

$$R = 1 \,. \tag{A2.13}$$

We insert Eq. A2.10 in A2.12 and obtain

$$\alpha_{\min} = \arctan\left(\frac{1}{\sin(\alpha_v)} - \frac{\cos(\alpha_v)}{\sin(\alpha_v)}\right) = \arctan\left(\tan\left(\frac{\alpha_v}{2}\right)\right) = \frac{\alpha_v}{2} \tag{A2.14}$$

where we have used $\tan(\alpha/2) = (1-\cos(\alpha))/\sin(\alpha)$. Eq. A2.14 also follows from the symmetry of the problem. This result yields

$$A_\cap = A_\triangledown(R, \alpha_{\min}, \alpha_2) + A_\triangledown(R, \alpha_v - \alpha_{\min}, \alpha_1) = 2A_\triangledown(R, \alpha_{\min}, \alpha_1) \,. \tag{A2.15}$$

We rewrite the regularized incomplete Beta function as

$$I_{1-\frac{\tan^2(\alpha_{\min})}{\tan^2(\varphi)}}\left[\frac{N_{\text{in}}-2}{2},\frac{1}{2}\right]=\frac{\Gamma(\frac{N_{\text{in}}-1}{2})}{\Gamma(\frac{N_{\text{in}}-2}{2})\Gamma(\frac{1}{2})}B\left[1-\frac{\tan^2(\alpha_{\min})}{\tan^2(\varphi)};\frac{N_{\text{in}}-2}{2},\frac{1}{2}\right] \tag{A2.16}$$

where B is the incomplete Beta function. The area of an $(N_{\text{in}}-1)$-dimensional hypersphere is

$$A_\circ=\frac{2\pi^{\frac{N_{\text{in}}}{2}}}{\Gamma(\frac{N_{\text{in}}}{2})}R^{N_{\text{in}}-1}\,. \tag{A2.17}$$

We insert Eq. A2.16 in A2.9 and use the result in Eq. A2.15. Using the notation $\alpha_b=\alpha_1$ and $\alpha_x=\alpha_{\min}$, and the identities

$$\frac{\Gamma(\frac{N_{\text{in}}}{2})}{\Gamma(\frac{N_{\text{in}}-2}{2})}=\frac{N_{\text{in}}-2}{2},\quad \Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}\,, \tag{A2.18}$$

the ratio between the overlapping area of the hyperspherical caps and the complete are of the hypersphere can now be calculated as

$$\frac{A_\cap}{A_\circ}=\frac{N_{\text{in}}-2}{2\pi}\int_{\alpha_x}^{\alpha_b}\sin(\varphi)^{N_{\text{in}}-2}B\left[1-\frac{\tan^2(\alpha_x)}{\tan^2(\varphi)};\frac{N_{\text{in}}-2}{2},\frac{1}{2}\right]\mathrm{d}\varphi\,. \tag{A2.19}$$

A2.2.2 Derivation of Eq. 2.19.

For a large bias $b\gtrsim 0.9$, which is equivalent to a small angle $\alpha_b=\arccos(b)$, the hyperspherical caps surrounding $\mathbf{x}_i$ and $\mathbf{x}_j$ will be very small in relation to the whole hypersphere. In this case, we can neglect the curvature of the hyperspherical surface and project the area of the hyperspherical cap to the plane that cuts through the rims of the cap. This projection is a $(N_{\text{in}}-1)$-dimensional hyperball (we will refer to it as a mini-ball). In three dimensions, for example, the projection of a spherical cap to the plane constitutes a disk, which is a 2-dimensional ball. The radius of each mini-ball is

$$\hat{b}=\sin(\arccos(b)) \tag{A2.20}$$

and the half-distance between the centers of the mini-balls is

$$\Delta=\frac{1}{2}\|\mathbf{x}_i-\mathbf{x}_j\|_2\,. \tag{A2.21}$$

We estimate the overlapping area of the hyperspherical caps by calculating the overlapping volume of the mini-balls in $(N_{\text{in}}-1)$ dimensions. The overlapping volume of two hyperballs has been computed by Li (2011) and is

$$V_\cap=\frac{\pi^{\frac{N_{\text{in}}-1}{2}}}{\Gamma(\frac{N_{\text{in}}+1}{2})}\hat{b}^{N_{\text{in}}-1}I_{1-\left(\frac{\Delta}{\hat{b}}\right)^2}\left[\frac{N_{\text{in}}}{2},\frac{1}{2}\right]\,. \tag{A2.22}$$

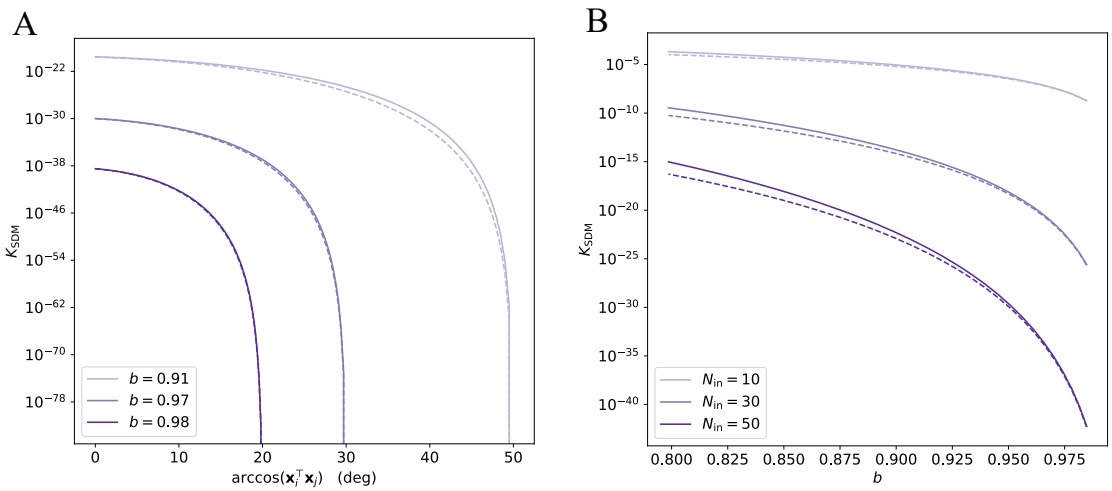


Figure A2.1: Plot of the kernel of an infinite SDM on the hypersphere, $K_{\mathrm{SDM}}(\mathbf{x}_i, \mathbf{x}_j)$, as a function of (A) the angle between $\mathbf{x}_i$ and $\mathbf{x}_j$, and (B) the bias b. Solid lines represent the exact solution in Eq. A2.19, and dashed lines the approximation in Eq. A2.24. Parameter values: (A) $N_{\mathrm{in}} = 50$; (B) $\arccos(\mathbf{x}_i^\top \mathbf{x}_j) = \arccos(b)$.

We rewrite the regularized incomplete Beta function as

$$I_{1-\left(\frac{\Delta}{2\hat{b}}\right)^2}\left[\frac{N_{\mathrm{in}}}{2}, \frac{1}{2}\right] = \frac{\Gamma\left(\frac{N_{\mathrm{in}}+1}{2}\right)}{\Gamma\left(\frac{N_{\mathrm{in}}}{2}\right)\Gamma\left(\frac{1}{2}\right)} B\left[1 - \left(\frac{\Delta}{\hat{b}}\right)^2; \frac{N_{\mathrm{in}}}{2}, \frac{1}{2}\right] \tag{A2.23}$$

and insert Eq. A2.23 in A2.22. The ratio between the overlapping area of the hyperspherical caps and the complete area of the hypersphere can now be estimated as

$$\frac{A_\cap}{A_\circ} \approx \frac{V_\cap}{A_\circ} = \frac{\hat{b}^{N_{\mathrm{in}}-1}}{2\pi} B\left[1 - \left(\frac{\Delta}{\hat{b}}\right)^2; \frac{N_{\mathrm{in}}}{2}, \frac{1}{2}\right] . \tag{A2.24}$$

A comparison of the exact solution in Eq. A2.19 and the approximation in Eq. A2.24 can be seen in Fig. A2.1.

A2.3 Iterative learning in an SVM network

We will in this section compare the noise tolerance of a single neuron in an SVM network when trained with an iterative learning rule, and when configured according to the MHN. First, we choose to equip the neuron with the feature map $\boldsymbol{\phi}_{\mathrm{pairs}}(\mathbf{x})$, which consists of all unique pairs of cross-terms $x_i x_j$, $i \neq j$. This yields a storage capacity scaling of $\mathcal{O}(N^2)$, and we therefore parameterize the storage load as M/N^2. We train the weights of the neuron either with the stochastic batch perceptron rule (Cotter et al., 2012) or with the one-shot learning rule of the MHN, which is obtained by setting $\alpha^\mu = 1$, $\forall \mu$, in Eq. 2.4, that is

$$\mathbf{w} = \sum_\mu^M \xi_{\mathrm{out}}^\mu \boldsymbol{\phi}(\boldsymbol{\xi}_{\mathrm{in}}^\mu) . \tag{A2.25}$$

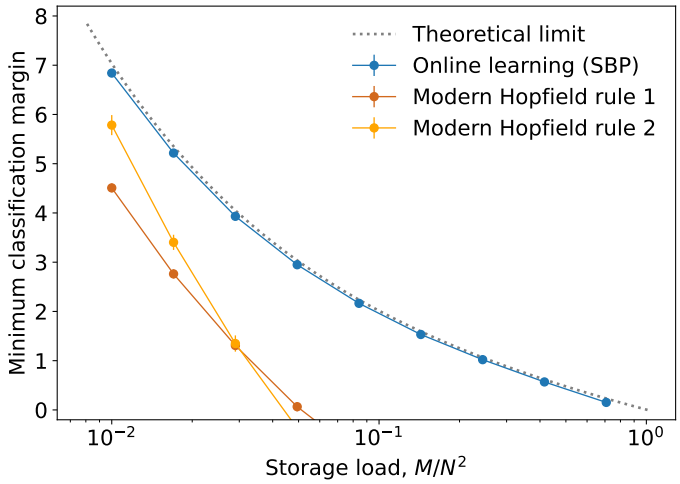


Figure A2.2: Plot of the noise tolerance γ (mean $\pm$ s.e.m. over 20 simulations) as a function of the storage load for a single SVM neuron with $N = 10^2$ inputs, trained with the stochastic batch perceptron (SBP) and the modern Hopfield rule. The SBP uses the feature map $\boldsymbol{\phi}_{\text{pairs}}$, while the modern Hopfield rule is applied to both $\boldsymbol{\phi}_{\text{pairs}}$ and $\boldsymbol{\phi}_{\text{poly2}}$, corresponding to the kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j)^2$. SBP hyperparameters: learning rate $= 10^{-5}$, iterations $= 20M$.

Finally, we quantify the noise tolerance as the smallest Euclidean distance between the neuron's decision boundary and the patterns $\{\boldsymbol{\xi}_{\text{in}}^\mu\}_{\mu=1}^M$. This is equivalent to the minimum classification margin, defined as

$$\gamma = \min_\mu \frac{\xi_{\text{out}}^\mu (\mathbf{w}^\top \boldsymbol{\xi}_{\text{in}}^\mu)}{\|\mathbf{w}\|_2} \,. \tag{A2.26}$$

We are only interested in the performance regime where all patterns are correctly recalled (i.e., correctly classified). This means that we only compare positive margins, since a negative margin indicates that there is one or more patterns that no longer can be correctly recalled. The results are plotted in Fig. A2.2, and demonstrate that the margin for the MHN quickly drops with increasing load, while the iterative learning rule achieves a margin close to the theoretical optimum derived by Gardner (1988). Moreover, as the maximum storage capacity M_{max} of each learning rule can be found at the intersection between the margin curve and the line $\gamma = 0$, the capacity of the online rule can be estimated to $\sim 0.7N^2$, which is more than an order of magnitude higher than that of the MHN, which is $\sim 0.05N^2$.

A2.4 Generalized pseudoinverse rule

When the network is linear and underdetermined, meaning $M < N$, we can make sure that all patterns are attractors by modeling each neuron as a *least-squares SVM* (Suykens & Vandewalle, 1999) instead of a conventional SVM, so that the weights satisfy

$$\min_{\mathbf{w}} \|\mathbf{w}_i\|_2 \quad \text{s.t.} \quad \mathbf{w}_i^\top \boldsymbol{\xi}^\mu = \xi_i^\mu, \quad \forall \mu, i \,. \tag{A2.27}$$

This is a minimum-norm interpolation problem, and yields the solution

$$\mathbf{s}^{(t+1)} = \text{sgn}\left[\mathbf{X}\mathbf{K}^{\dagger}\mathbf{X}^{\top}\mathbf{s}^{(t)}\right] = \text{sgn}\left[\mathbf{X}\mathbf{X}^{\dagger}\mathbf{s}^{(t)}\right] \tag{A2.28}$$

where $\mathbf{K} = \mathbf{X}^{\top}\mathbf{X}$ is the *kernel matrix* and $\mathbf{K}^{\dagger}$ its Moore-Penrose pseudoinverse, and where we have used the property $\mathbf{K}^{\dagger}\mathbf{X}^{\top} = (\mathbf{X}^{\top}\mathbf{X})^{\dagger}\mathbf{X}^{\top} = \mathbf{X}^{\dagger}$. This is the *pseudoinverse learning rule* (Personnaz et al., 1986).

The derivation can be extended to MHNs by performing interpolation on the feature map $\boldsymbol{\phi}(\boldsymbol{\xi}^{\mu})$. Assuming that the problem is still underdetermined, so that $M < N_{\phi}$, we aim to find the weights

$$\min_{\mathbf{w}} \|\mathbf{w}_i\|_2 \quad \text{s. t.} \quad \mathbf{w}_i^{\top}\boldsymbol{\phi}(\boldsymbol{\xi}^{\mu}) = \xi_i^{\mu}, \quad \forall \mu, i \tag{A2.29}$$

which, analogously to the linear case, produces the optimal state update

$$\mathbf{s}^{(t+1)} = \text{sgn}\left[\mathbf{X}\mathbf{K}^{\dagger}K(\mathbf{X}, \mathbf{s}^{(t)})\right] \tag{A2.30}$$

where $\mathbf{K} = K(\mathbf{X}, \mathbf{X}) = \boldsymbol{\phi}(\mathbf{X})^{\top}\boldsymbol{\phi}(\mathbf{X})$. This can, again, be simplified to

$$\mathbf{s}^{(t+1)} = \text{sgn}\left[\mathbf{X}\boldsymbol{\phi}(\mathbf{X})^{\dagger}\boldsymbol{\phi}(\mathbf{s}^{(t)})\right] \tag{A2.31}$$

where we can identify the weight matrix $\mathbf{W} = \mathbf{X}\boldsymbol{\phi}(\mathbf{X})^{\dagger}$. This is the *generalized pseudoinverse learning rule*. Note that, if the feature-expanded patterns $\{\boldsymbol{\phi}(\boldsymbol{\xi}^{\mu})\}_{\mu=1}^{M}$ are linearly independent, the kernel matrix is invertible and we have $\mathbf{K}^{\dagger} = \mathbf{K}^{-1}$.

A2.5 The kernel memory network for continuous patterns

A2.5.1 Minimum norm interpolation and attractor basin size

Proof of Property 4 (Robust auto-associative memory with continuous patterns)**.** In the most general variant of this setting, each neuron i is modeled as a linear regressor with a neuron-specific feature map $\boldsymbol{\phi}_i$ and a state s_i which is updated according to

$$s_i^{(t+1)} = \mathbf{w}_i^{\top}\boldsymbol{\phi}_i(\mathbf{s}^{(t)}) \, . \tag{A2.32}$$

All patterns $\mathbf{X}$ are guaranteed to be fixed points of the dynamics by finding the weights that satisfy

$$\xi_i^{\mu} = \mathbf{w}_i^{\top}\boldsymbol{\phi}_i(\boldsymbol{\xi}^{\mu}), \ \forall \mu, i \, . \tag{A2.33}$$

In order for each pattern to also be an attractor, the weights must satisfy the additional constraint

$$\|\mathbf{J}_s\|_2\Big|_{\mathbf{s}^{(t)}=\boldsymbol{\xi}^{\mu}} < 1, \ \forall \mu \tag{A2.34}$$

where $\mathbf{J}_s$ is the Jacobian of the state update rule with respect to the input $\mathbf{s}^{(t)}$. The meaning of Eq. A2.34 is that the spectral norm of the Jacobian must be less than 1 when evaluated at each pattern. The reason for this is that the update rule, which computes $\mathbf{s}^{(t+1)}$ as a

function of $\mathbf{s}^{(t)}$ (either synchronously or asynchronously) is continuously differentiable with respect to $\mathbf{s}^{(t)}$ and therefore satisfies the mean value inequality, so that

$$\left\|\mathbf{s}_1^{(t+1)} - \mathbf{s}_2^{(t+1)}\right\|_2 \leq \hat{J}_s \left\|\mathbf{s}_1^{(t)} - \mathbf{s}_2^{(t)}\right\|_2 \tag{A2.35}$$

where $\hat{J}_s$ is an upper bound of the spectral norm, meaning

$$\|\mathbf{J}_s\|_2 \leq \hat{J}_s \ . \tag{A2.36}$$

If $\hat{J}_s < 1$ at a pattern $\boldsymbol{\xi}^\mu$, it is also possible to find a neighborhood around $\boldsymbol{\xi}^\mu$ where $\hat{J}_s < 1$ holds as well, due to the continuity of the state update rule. Given this, the Banach fixed point theorem ensures that the state update rule is a contractive map in a region surrounding $\boldsymbol{\xi}^\mu$ and, equivalently, that $\boldsymbol{\xi}^\mu$ is a stable attractor (Radhakrishnan et al., 2020). While the complete basin of attraction of $\boldsymbol{\xi}^\mu$ might be difficult to compute exactly, we can define a subset of the basin as the set of points $\mathcal{S}^\mu$ in the open neighborhood of $\boldsymbol{\xi}^\mu$ satisfying

$$\mathcal{S}^\mu = \{\mathbf{s}^{(t)} : \|\mathbf{J}_s\|_2 < 1\} \ . \tag{A2.37}$$

Given that the spectral norm of the Jacobian is upper-bounded by the Frobenius norm, that is

$$\|\mathbf{J}_s\|_2 \leq \|\mathbf{J}_s\|_F \ , \tag{A2.38}$$

we can obtain a lower bound of the extent of the basin of attraction with the set

$$\widehat{\mathcal{S}}^\mu = \{\mathbf{s}^{(t)} : \|\mathbf{J}_s\|_F < 1\} \ . \tag{A2.39}$$

We can write $\mathbf{J}_s$ as

$$\mathbf{J}_s = \overline{\mathbf{W}} \cdot \overline{\mathbf{J}}_\phi \tag{A2.40}$$

where

$$\overline{\mathbf{W}} = \begin{pmatrix} \mathbf{w}_1^\top & 0 & \cdots & 0 \\ 0 & \mathbf{w}_2^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{w}_N^\top \end{pmatrix} \ , \quad \overline{\mathbf{J}}_\phi = \begin{pmatrix} \mathbf{J}_{\phi_1} \\ \mathbf{J}_{\phi_2} \\ \vdots \\ \mathbf{J}_{\phi_N} \end{pmatrix} \tag{A2.41}$$

and where $\mathbf{J}_{\phi_i}$ is the Jacobian of $\boldsymbol{\phi}_i(\mathbf{s}^{(t)})$ with respect to $\mathbf{s}^{(t)}$. This gives us

$$\|\mathbf{J}_s\|_F = \|\overline{\mathbf{W}} \cdot \overline{\mathbf{J}}_\phi\|_F \leq \|\overline{\mathbf{W}}\|_F \cdot \|\overline{\mathbf{J}}_\phi\|_F \tag{A2.42}$$

where the last expression is given by the Cauchy-Schwartz inequality. Since $\|\overline{\mathbf{J}}_\phi\|_F$ depends only on the kernel, which is fixed, the right-hand side of Eq. A2.42 can only be minimized by finding a set of weights that minimize $\|\overline{\mathbf{W}}\|_F$. By first rewriting this norm as

$$\|\overline{\mathbf{W}}\|_F^2 = \sum_i^N \|\mathbf{w}_i\|_2^2 \tag{A2.43}$$

we see that its minimum is obtained by minimizing $\|\mathbf{w}_i\|_2$, $\forall i$. Combining this requirement with Eq. A2.33 is equivalent to performing a minimum norm interpolation, that is

$$\min_{\mathbf{w}_i} \|\mathbf{w}_i\|_2 \quad \text{s.t.} \quad \xi_i^\mu = \mathbf{w}_i^\top \boldsymbol{\phi}_i(\boldsymbol{\xi}^\mu), \quad \forall \mu, i \; . \tag{A2.44}$$

If we now assume, as in the binary case, that all neurons use the same feature map, so that $\boldsymbol{\phi}_i = \boldsymbol{\phi}$, $\forall i$, the solution can be compactly written as in Eq. 2.22. This maximizes a lower bound of the attractor basin size, as defined by $\mathcal{S}^\mu$, for each pattern $\boldsymbol{\xi}^\mu$. □

A2.5.2 Normally distributed patterns

Proof of Property 5.1 (EPK network at zero temperature)**.** Using the notation $\Delta = \|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2$, we have that

$$\lim_{\beta\to\infty} \left(\frac{\Delta}{r}\right)^\beta = \begin{cases} 0 \; , & \Delta < r \\ 1 \; , & \Delta = r \\ \infty \; , & \Delta > r \end{cases} \tag{A2.45}$$

from which it follows that

$$\lim_{\beta\to\infty} \exp\left[-\left(\frac{\Delta}{r}\right)^\beta\right] = \begin{cases} 1 \; , & \Delta < r \\ e^{-1} \; , & \Delta = r \\ 0 \; , & \Delta > r \end{cases} \tag{A2.46}$$

which is equivalent to $\Theta(r - \Delta)$ with $\Theta(0) = e^{-1}$. We combine this with the assumption that the patterns are unique and that $\min_{\mu,\nu\neq\mu}\|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2 > r$ and obtain

$$\lim_{\beta\to\infty} K_{\text{EPK}}(\boldsymbol{\xi}^\mu, \boldsymbol{\xi}^\nu) = \Theta(r - \|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2) = \begin{cases} 1 \; , & \mu = \nu \\ 0 \; , & \mu \neq \nu \end{cases} \tag{A2.47}$$

which can be written compactly as

$$\lim_{\beta\to\infty} \mathbf{K}_{\text{EPK}} = \mathbf{I}_M \; . \tag{A2.48}$$

It directly follows that

$$\lim_{\beta\to\infty} \mathbf{K}_{\text{EPK}}^{-1} = \mathbf{I}_M^{-1} = \mathbf{I}_M \tag{A2.49}$$

and, therefore,

$$\lim_{\beta\to\infty} \mathbf{X}\mathbf{K}_{\text{EPK}}^{-1} K_{\text{EPK}}(\mathbf{X}, \mathbf{s}^{(t)}) = \mathbf{X}\,\Theta(r^2 - \|\mathbf{X} - \mathbf{s}^{(t)}\|_2^2) \; . \tag{A2.50}$$

□

Proof of Property 6 (Robustness to white noise)**.** With $\mathbf{s}^{(0)} = \boldsymbol{\xi}^\mu + \boldsymbol{\epsilon}$, we have

$$\|\boldsymbol{\xi}^\mu - \mathbf{s}^{(0)}\|_2^2 = \|\boldsymbol{\epsilon}\|_2^2 = \|\sigma\boldsymbol{\epsilon}_0\|_2^2 = \sigma^2\|\boldsymbol{\epsilon}_0\|_2^2 \tag{A2.51}$$

where $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, from which it follows that $\|\boldsymbol{\epsilon}_0\|_2^2$ is a random variable with a $\chi^2(N)$ distribution. According to the central limit theorem, we also have

$$\lim_{N\to\infty} \frac{\|\boldsymbol{\epsilon}_0\|_2^2 - N}{\sqrt{2N}} \sim \mathcal{N}(0, 1) \tag{A2.52}$$

where we have used the fact that each term in $\|\boldsymbol{\epsilon}_0\|_2^2$ is $\chi^2(1)$-distributed, and has mean 1 and variance 2. We will therefore make the approximation $\|\boldsymbol{\epsilon}_0\|_2^2 \sim \mathcal{N}(N, 2N)$ for large N. This gives us

$$\sigma^2\|\boldsymbol{\epsilon}_0\|_2^2 \sim \mathcal{N}(\sigma^2 N, 2\sigma^4 N) \ . \tag{A2.53}$$

The original pattern $\boldsymbol{\xi}^\mu$ will only be recovered if $r^2 - \|\boldsymbol{\xi}^\mu - \mathbf{s}^{(0)}\|_2^2 \geq 0$, which is satisfied in at least 50% of trials if $r^2 \geq \sigma^2 N$. The maximum variance with which this type of recovery still holds is thus

$$\sigma^2_{\text{max}} = r^2/N \ . \tag{A2.54}$$

□

Proof of Property 7 (Exponential storage capacity)**.** In the limit $\beta \to \infty$, the boundary of the basin of attraction surrounding each pattern is sharp. In this setting, we are guaranteed that each pattern can be recalled without errors as long as $\min_{i,j\neq i}\|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\|_2 > 2r$. We will therefore estimate the storage capacity by calculating the number of patterns, on average, that can be loaded into the network before at least two attractor basins overlap and the condition above is violated (see Fig. A2.3).

We begin by observing that for two random patterns $\boldsymbol{\xi}^i, \boldsymbol{\xi}^j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$, we have

$$\frac{1}{2}\|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\|_2^2 \sim \chi^2(N) \tag{A2.55}$$

which, using the central limit theorem as in Eq. A2.52, can be approximated as $\mathcal{N}(N, 2N)$ for large N, thereby yielding

$$\|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\|_2^2 \sim \mathcal{N}(2N, 8N) \ . \tag{A2.56}$$

We now assume that the squared Euclidean distance between each pair of patterns $\boldsymbol{\xi}^i, \boldsymbol{\xi}^j$ in a set of M given patterns $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^M$ is an independent sample of a random variable, denoted Δ^2, which is distributed as in Eq. A2.56. This is, of course, an approximation which neglects that the pairwise distances between any set of points are inter-dependent. Nonetheless, for relatively large N and M, the approximation accurately describes the empirical distance distribution.

Relying on this assumption, the process of drawing M random patterns becomes equivalent to drawing $M(M-1)/2$ unique pairwise distances Δ^2 from the distance distribution. The probability of drawing a sample $\Delta^2 \leq 4r^2$ can be calculated using the cumulative distribution function for the standard normal distribution, given by $\Phi(x) = \frac{1}{2}\,\text{erfc}(-x)$, according to

$$\mathbb{P}(\Delta^2 \leq 4r^2) = \frac{1}{2}\,\text{erfc}\left(\frac{N - 2r^2}{2\sqrt{N}}\right) \ . \tag{A2.57}$$

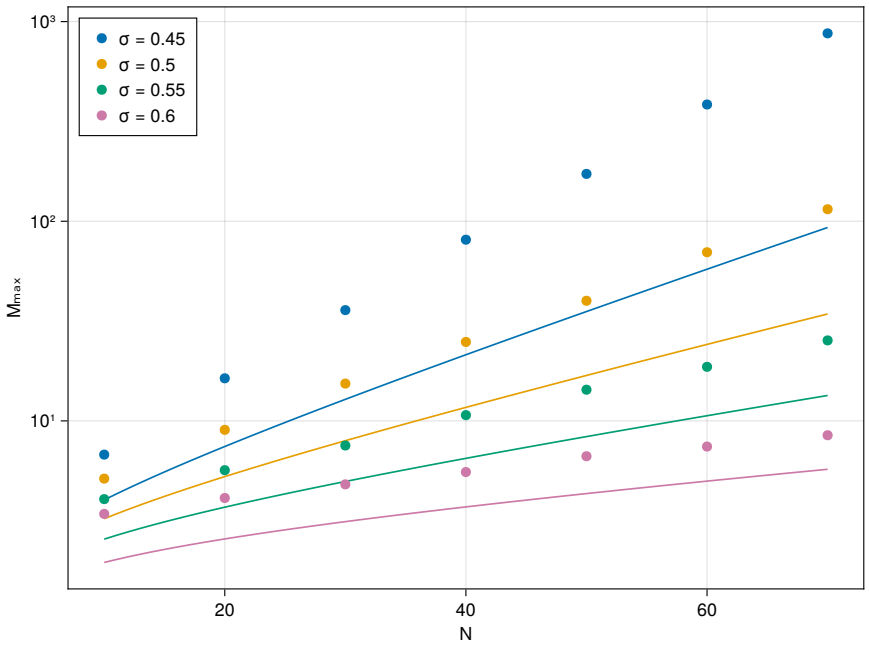


Figure A2.3: Plot of the storage capacity of the EPK network at $\beta \to \infty$ with normally distributed patterns. Dots represent means ($\pm$ s.e.m.) over 1000 simulations, in which the capacity is determined by the number of patterns sampled until one of the pairwise distances is smaller than $2r$. The standard error is too small to be visible. Lines correspond to the bound in Eq. A2.63. Note that the plot is log-linear, so the linear increase indicates that $M_{\max}$ scales exponentially in N.

The average number of samples of Δ^2 one needs to draw before a sample satisfies $\Delta^2 \leq 4r^2$ is given by $\mathbb{P}(\Delta^2 \leq 4r^2)^{-1}$. This determines the maximum number of patterns that the network, on average, can store, according to

$$\frac{M_{\max}(M_{\max}-1)}{2} = \frac{1}{\mathbb{P}(\Delta^2 \leq 4r^2)} \,. \tag{A2.58}$$

We combine this expression with the approximation $M_{\max}(M_{\max}-1) \approx M_{\max}^2$ (which holds for large $M_{\max}$) and Eq. A2.57, and obtain

$$M_{\max} = 2\,\mathrm{erfc}\left(\frac{N-2r^2}{2\sqrt{N}}\right)^{-1/2} \,. \tag{A2.59}$$

We now parameterize the radius r in terms of the largest tolerable noise amplitude, according to Eq. A2.54. This gives us

$$M_{\max} = 2\,\mathrm{erfc}\left(\frac{\sqrt{N}(1-2\sigma_{\max}^2)}{2}\right)^{-1/2} \,. \tag{A2.60}$$

Given that the erfc function can be well approximated using the asymptotic expansion

$$\mathrm{erfc}(x) \approx \frac{e^{-x^2}}{\sqrt{\pi}x} \sum_{n=0}^{\infty} (-1)^n \frac{(2n-1)!!}{(2x^2)^n} \tag{A2.61}$$

for large arguments x, we can obtain a tight lower bound of erfc^{-1} as long as N is large and $\sigma^2_{\max} \lesssim 1/2$ with the inverse zeroth order expansion, thereby obtaining

$$\mathrm{erfc}(x)^{-1} \geq \sqrt{\pi} x e^{x^2} \ . \tag{A2.62}$$

We insert Eq. A2.62 in A2.60 and finally obtain

$$M_{\max} \geq \sqrt{2\sqrt{\pi N}(1-2\sigma^2_{\max})}\exp\left[\frac{N(1-2\sigma^2_{\max})^2}{8}\right] \ . \tag{A2.63}$$

□

A2.5.3 Patterns on the hypersphere

Property 8 (Storage capacity: patterns on the hypersphere)**.** *At $\beta \to \infty$, the average maximum number of patterns that the EPK network can store and recall without errors is lower-bounded by*

$$M_{\max} \geq \sqrt{\sqrt{8\pi N}(1-2r^2)}\exp\left[\frac{N(1-2r^2)^2}{4}\right] \tag{A2.64}$$

when each pattern is randomly drawn from $\mathbb{S}^{N-1}$.

Proof. We begin by observing that for two random patterns $\boldsymbol{\xi}^i, \boldsymbol{\xi}^j \in \mathbb{S}^{N-1}$, we have

$$\|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2^2 = 2(1 - \boldsymbol{\xi}^{\mu\top}\boldsymbol{\xi}^\nu) \ . \tag{A2.65}$$

The probability distribution for the inner product $\omega = \boldsymbol{\xi}^{\mu\top}\boldsymbol{\xi}^\nu$ has been derived by Tony Cai & Jiang (2012), and is

$$\omega \sim \frac{1}{\sqrt{\pi}}\frac{\Gamma(\frac{N}{2})}{\Gamma(\frac{N-1}{2})}(1-\omega^2)^{\frac{N-3}{2}} \tag{A2.66}$$

which, for large N, can be approximated as

$$\omega \sim \mathcal{N}(0, \frac{1}{N}) \ . \tag{A2.67}$$

We use Eq. A2.67 in A2.65 and obtain

$$\Delta^2 \sim \mathcal{N}(2, \frac{4}{N}) \ . \tag{A2.68}$$

The probability of placing a pair of random points on $\mathbb{S}^{N-1}$ with $\Delta^2 \leq 4r^2$ is thus

$$\mathbb{P}(\Delta^2 \leq 4r^2) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\sqrt{N}(1-2r^2)}{\sqrt{2}}\right) \ . \tag{A2.69}$$

The average number of samples of Δ^2 one needs to draw before a sample satisfies $\Delta^2 \leq 4r^2$ is given by $\mathbb{P}(\Delta^2 \leq 4r^2)^{-1}$, and the maximum number of patterns that the network, on

average, can store, is therefore

$$M_{\max} = 2\,\mathrm{erfc}\left(\frac{\sqrt{N}(1-2r^2)}{\sqrt{2}}\right)^{-1/2} . \tag{A2.70}$$

We use the lower bound in Eq. A2.62 in A2.70 and finally obtain

$$M_{\max} \geq \sqrt{\sqrt{8\pi N}(1-2r^2)}\exp\left[\frac{N(1-2r^2)^2}{4}\right] . \tag{A2.71}$$

□

A2.5.4 Patterns on the hypercube

Property 9.1 (Robustness to flipped bits)**.** *Assume that we are given a set of unique patterns* $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^M \in \{\pm 1\}^N$ *with* $\min_{\mu,\nu\neq\mu}\|\boldsymbol{\xi}^\mu - \boldsymbol{\xi}^\nu\|_2 > 2r$*, and that the EPK network is initialized in a distorted pattern* $\mathbf{s}^{(0)} = \boldsymbol{\xi}^\mu \odot \boldsymbol{\epsilon}$*, where* $\boldsymbol{\epsilon} \in \{\pm 1\}^N$*, with* $\mathbb{P}(\epsilon_i = -1) = \rho, \forall i$*. Then, at* $\beta \to \infty$*, the maximum bit-wise error probability* $\rho_{\max}$ *with which* $\boldsymbol{\xi}^\mu$ *can be recovered in at least 50% of trials is*

$$\rho_{\max} = r^2/4N . \tag{A2.72}$$

Proof. With $\mathbf{s}^{(0)} = \boldsymbol{\xi}^\mu \odot \boldsymbol{\epsilon}$, we have

$$\|\boldsymbol{\xi}^\mu - \mathbf{s}^{(0)}\|_2^2 = \|2\boldsymbol{\epsilon}_{\mathcal{B}}\|_2^2 = 4\|\boldsymbol{\epsilon}_{\mathcal{B}}\|_2^2 \tag{A2.73}$$

where $\boldsymbol{\epsilon}_{\mathcal{B}} \in \{0,1\}^N$, with each entry being a random variable distributed as $(\boldsymbol{\epsilon}_{\mathcal{B}})_i \sim$ Bernoulli(ρ). This implies that $\|\boldsymbol{\epsilon}_{\mathcal{B}}\|_2^2 \sim$ Binomial(N,ρ), which can be approximated as $\mathcal{N}(\rho N, \rho(1-\rho)N)$ for large N. This gives

$$4\|\boldsymbol{\epsilon}_{\mathcal{B}}\|_2^2 \sim \mathcal{N}(4\rho N, 16\rho(1-\rho)N) . \tag{A2.74}$$

Again, the original pattern $\boldsymbol{\xi}^\mu$ will only be recovered if $r^2 - \|\boldsymbol{\xi}^\mu - \mathbf{s}^{(0)}\|_2^2 \geq 0$, which is satisfied in at least 50% of trials if $r^2 \geq 4\rho N$. The maximum bit-wise error probability with which this type of recovery still holds is thus

$$\rho_{\max} = r^2/4N . \tag{A2.75}$$

□

Remark. In Eq. A2.54, σ roughly quantifies the maximum noise fluctuation around a pattern that is tolerable with a given radius r, while still being able to recover the pattern in a majority of trials. In the case of Eq. A2.75, ρ instead quantifies the maximum tolerable bit-wise error probability.

Property 9.2 (Storage capacity: patterns on the hypercube)**.** *At* $\beta \to \infty$*, the average maximum number of bipolar patterns with sparseness* f *that the EPK network can store and*

recall without errors is lower-bounded by

$$M_{\max} \geq 2\left(\frac{\pi N}{2\tilde{f}(1-\tilde{f})}\right)^{1/4}\left(\tilde{f}-4\rho_{\max}\right)^{1/2}\exp\left[\frac{N(\tilde{f}-4\rho_{\max})^2}{4\tilde{f}(1-\tilde{f})}\right] \tag{A2.76}$$

where $\tilde{f}=2f(1-f)$ *and* $\rho_{\max}$ *is the maximum bit-wise error probability tolerated by the network.*

Proof. This proof is, again, a slightly modified variant of the proof of Property 7. First, we observe that for two random patterns $\boldsymbol{\xi}^\mu, \boldsymbol{\xi}^\nu \in \{\pm 1\}^N$ with sparseness f, so that $\mathbb{P}(x_i^{\mu,\nu}=1)=f$, it is true that

$$\frac{1}{4}\|\boldsymbol{\xi}^\mu-\boldsymbol{\xi}^\nu\|_2^2 \sim \text{Binomial}(N,\tilde{f}) \tag{A2.77}$$

where $\tilde{f}=2f(1-f)$ denotes the probability that $\boldsymbol{\xi}^\mu$ and $\boldsymbol{\xi}^\nu$ differ at any given entry. For large N, we can again approximate Binomial$(N,\tilde{f})$ with $\mathcal{N}(\tilde{f}N,\tilde{f}(1-\tilde{f})N)$, which gives us

$$\|\boldsymbol{\xi}^\mu-\boldsymbol{\xi}^\nu\|_2^2 \sim \mathcal{N}(4\tilde{f}N, 16\tilde{f}(1-\tilde{f})N)\ . \tag{A2.78}$$

We use this to compute the upper bound of the probability of drawing a distance Δ^2 which satisfies $\Delta^2 \leq 4r^2$, as in Eq. A2.57. The result is

$$\mathbb{P}(\Delta^2 \leq 4r^2) = \frac{1}{2}\,\text{erfc}\left(\frac{\tilde{f}N-r^2}{\sqrt{2\tilde{f}(1-\tilde{f})N}}\right)\ . \tag{A2.79}$$

Following the same derivations used to produce Eqs. A2.58 and A2.59, we arrive at

$$M_{\max} = 2\,\text{erfc}\left(\frac{\tilde{f}N-r^2}{\sqrt{2\tilde{f}(1-\tilde{f})N}}\right)^{-1/2}\ . \tag{A2.80}$$

As before, we parameterize the radius r in terms of the maximum tolerable bit-wise error probability according to Eq. A2.75 and yield

$$M_{\max} = 2\,\text{erfc}\left(\frac{\sqrt{N}(\tilde{f}-4\rho_{\max})}{\sqrt{2\tilde{f}(1-\tilde{f})}}\right)^{-1/2}\ . \tag{A2.81}$$

We finally replace erfc with the lower bound in Eq. A2.62. This is valid for large N and when $\rho \lesssim \tilde{f}/4$. We obtain

$$M_{\max} \geq 2\left(\frac{\pi N}{2\tilde{f}(1-\tilde{f})}\right)^{1/4}\left(\tilde{f}-4\rho_{\max}\right)^{1/2}\exp\left[\frac{N(\tilde{f}-4\rho_{\max})^2}{4\tilde{f}(1-\tilde{f})}\right]\ . \tag{A2.82}$$

□

A2.6 Comparison to neuron models with active dendrites

To demonstrate how single neurons in kernel memory networks are generalizations of abstract neuron models with active dendrites, we begin by considering a neuron in the feature form (Eq. 2.6). We will here use the Heaviside activation function with threshold θ, denoted Θ_θ, instead of sgn, and assume that patterns and states are in $\{0,1\}^N$, where N is the number of inputs. This, however, does not change the fundamental properties of our model, as SVMs can be formulated for binary data with minor modifications. Assuming a polynomial feature map $\boldsymbol{\phi}$ of degree p, the feature vector will consist of all possible monomials of degree $\leq p$ composed of the states of its input neurons. Setting, for example, $p = 2$ gives us

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{x}) = (1, &\sqrt{2}x_1, \ldots, \sqrt{2}x_N, \\ &\sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_N, \sqrt{2}x_2x_3, \ldots, \sqrt{2}x_{N-1}x_N, \\ &x_1^2, \ldots, x_N^2)\ .\end{aligned} \tag{A2.83}$$

Case (i): By limiting the feature map to only include a subset of all terms, our model is reduced to the 2-degree *sigma-pi unit* (Rumelhart & McClelland, 1986, p. 73), which can be written as

$$s_{\text{out}} = \Theta_\theta \left[\sum_i w_i \prod_{j \in \mathcal{C}_i} s_{\text{in},j} \right] \tag{A2.84}$$

where w_i is the weight of cluster i, which consists of a product of all inputs $s_{\text{in},j}$ whose indices j are contained in the set $\mathcal{C}_i$. From a neurophysiological perspective, each product represents the cross-talk between a set of synapses. By including such multiplicative interactions, synapses can both gate and amplify each other, to generate supra-linear input currents.

Case (ii): If we now further constrain this model to include only a subset of the cross-terms x_ix_j, $i \neq j$, and parameterize each cross-term weight as $w_{ij} = w_iw_j$, our neuron model is reduced to the *clusteron* (Mel, 1991), which can be written as

$$\begin{aligned} s_{\text{out}} &= \Theta_\theta \left[\sum_i^N \sum_{j \in \mathcal{C}_i} w_i w_j s_{\text{in},i} s_{\text{in},j} \right] \\ &= \Theta_\theta \left[\sum_i^N w_i s_{\text{in},i} \left(\sum_{j \in \mathcal{C}_i} w_j s_{\text{in},j} \right) \right] \end{aligned} \tag{A2.85}$$

where $\mathcal{C}_i$ now is the set describing all inputs j that input i should be paired with.

Case (iii): We now consider a neuron in the kernel form (Eq. 2.7) with an inner-product kernel $K(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i^\top \mathbf{x}_j)$. By setting $\xi^\mu_{\text{out}} = 1$, $\forall \mu$, and assuming binary inputs, our model is reduced to

$$s_{\text{out}} = \Theta_\theta \left[\sum_\mu^M \alpha^\mu k \left(\sum_i^N \xi^\mu_{\text{in},i} s_{\text{in},i} \right) \right] \tag{A2.86}$$

which is equivalent to the pyramidal cell as a two-layer neural network, as defined by Poirazi et al. (2003). According to the original interpretation, this neuron model is comprised of M subunits, which can represent, for example, separate parts of a dendritic tree. All subunits receive the inputs and produce separate outputs which are all summed in the soma. Each subunit μ is characterized by the input weights $\boldsymbol{\xi}_{\text{in}}^{\mu}$ (which serves as a mask), the output weight α^{μ} (which determines how strongly the subunit influences the response at the soma), and the activation function k.

Chapter 3

Optimal memory consolidation and pruning

This chapter is based on the following manuscript:

> "Optimal memory consolidation and compression with multiplicative synaptic plasticity and pruning"
> **Georgios Iatropoulos**, Johanni Brea, Wulfram Gerstner
> *In preparation*.

Abstract. During learning of a new task, cortical circuits exhibit brief synaptic growth, followed by a longer process of sleep-based synaptic pruning that preserves a sparse cortical connectivity. It remains unclear, however, what computational purpose pruning serves in long-term memory, and how to incorporate this into existing mathematical models of synaptic plasticity. Here, we propose a normative account of memory consolidation and pruning by deriving a synaptic learning rule that stores memories with maximal noise-tolerance and minimal connection density in a recurrent neural network. The model reproduces several features of learning from the wake-sleep cycle, such as structured memory replay, multiplicative hetero- and homosynaptic plasticity, synaptic cross-talk, as well as simultaneous plasticity expression in multiple sub-cellular components. Finally, the model predicts that intrinsic synaptic noise scales sublinearly with synaptic strength. This is confirmed by a meta-analysis of multiple published datasets on synaptic volatility.

Author contributions. GI created the model, performed the simulations, and analyzed the data. JB assisted in the data analysis and theoretical derivations. GI, JB, and WG wrote the article.

Acknowledgements. The authors would like to thank Prof. Haruo Kasai, Prof. Noam Ziv, Prof. Armen Stepanyants, Prof. Kimberly Fenn, and Dr. Rohan Gala for sharing their experimental data. This study was supported by funding from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology, to the Blue Brain Project, a research center of the École Polytechnique Fédérale de Lausanne (EPFL).

3.1 Introduction

Following a decades-long history of brain imaging studies and cognitive testing in healthy humans and patients with brain lesions, it is today generally accepted that long-term memories are stored in a distributed network of neurons primarily located in temporal cortex (Squire et al., 2004, 2015; Tonegawa et al., 2015; Roy et al., 2022). Anatomical studies of this brain region have demonstrated a high degree of local recurrent connectivity, both among pyramidal cells and inhibitory neurons (Thomson & Lamy, 2007; Harris & Shepherd, 2015). As a result of these findings, the attractor network framework has become a popular choice for modeling long-term memory (Hopfield, 1982; Khona & Fiete, 2022). The fundamental idea of this approach is to represent local cortical circuitry by a recurrent neural network, in which each memory corresponds to a distinct pattern of neural activity that acts as an attractor of the network's dynamics.

The process of imprinting a memory is modeled with a synaptic learning rule that configures the connections so as to transform activity patterns into stable attractors. At optimal configuration, the network's storage and noise robustness is maximized. This state has been extensively characterized in theoretical studies (Gardner, 1988; Köhler & Widmaier, 1991; Brunel et al., 2004; Chapeton et al., 2012; Brunel, 2016; Zhang et al., 2019) and has been proposed as an organizing principle for modeling cortical connectivity (Chapeton et al., 2012; Brunel, 2016; Zhang et al., 2019). However, the question of how a biological synaptic learning rule could induce such optimal synaptic configuration in cortex remains unanswered.

Past work on synaptic plasticity modeling has predominantly been phenomenological and based on the dependence of long-term potentiation (LTP) and depression (LTD) on, for example, cellular calcium concentration (Shouval et al., 2002; Graupner & Brunel, 2012), membrane voltage (Clopath et al., 2010), or spike timing (Morrison et al., 2008; Markram et al., 2011). While there exist plasticity models derived from assumptions of optimal storage, these are either problematic to implement biologically (Personnaz et al., 1986; Anlauf & Biehl, 1989) or require unrealistic network configurations, such as maximal connection density (Tsodyks & Feigel'man, 1988; Amari, 1989). The latter point is noteworthy; although the large number of axonal-dendritic appositions observed in cortical circuits suggests a high potential connectivity (Kalisman et al., 2005), electrophysiology has demonstrated that functional connectivity is, in fact, very sparse (Thomson & Lamy, 2007; Lefort et al., 2009). Moreover, cortical connections are dynamic and change in an experience-dependent way over hours and days (Trachtenberg et al., 2002; Holtmaat et al., 2005; Le Bé & Markram, 2006). Shortly following the learning of a new task, recruited neurons exhibit a rapid growth of new dendritic spines (Xu et al., 2009; Chen et al., 2015). Over the course of subsequent days, only a subset of these are selected for maturation, while the rest retract. This process of consolidating new memory traces and pruning excess connectivity has been found to occur primarily during sleep (Chen et al., 2015; Li et al., 2017; Zhou et al., 2020).

Over the span of a lifetime, cortical connectivity sparsens (Petanjek et al., 2011) and transitions from being comprised of mainly small and weak spines (filopodia) in infancy, to large and mature spines in adulthood (Grutzendler et al., 2002; Zuo et al., 2005a). Moreover, adult animals that have been reared in enriched environments end up with a higher density

of cortical connections compared to stimulus-deprived ones (Globus et al., 1973; Turner & Greenough, 1985).

These dynamics are typically neglected in attractor network models, which often assume a fixed connectivity. Although activity-dependent cortical rewiring has been modeled phenomenologically (Butz & van Ooyen, 2013; Zheng et al., 2013; Fauth et al., 2015; Deger et al., 2018; Gallinaro et al., 2022), a mathematically principled way to incorporate structural changes in task-driven plasticity models is still lacking.

In a series of recent theoretical studies, it has been shown that sparse structured connectivity can be induced in optimal attractor networks by imposing an appropriate weight scaling (Chapeton et al., 2012; Brunel, 2016). This, however, has proven challenging to implement in synaptic learning rules, as it typically requires additive weight regularization (Sacramento et al., 2015), weight thresholding (Chechik et al., 1998; Scholl et al., 2021) or gradient thresholding (Alemi et al., 2015). Such constraints imply that the synapse model would need to be fine-tuned to each specific storage problem, and also stands in contrast to experimental data demonstrating that homeostatic plasticity is multiplicative (Turrigiano, 2008).

Here, we reconcile many of the discrepancies between plasticity models and experimental data by taking a normative approach to long-term memory modeling. We derive a synaptic learning rule that solves the optimization problem of storing memories with maximal noise tolerance. Crucially, by representing the synaptic strength as a product of multiple internal components, the learning rule implicitly finds sparse forms of storage, thus eliminating any need for manual weight tuning or thresholding. Additionally, features such as prioritized replay of new memories, multiplicative hetero- and homosynaptic plasticity, and synaptic cross-talk naturally emerge from the derivation.

By evaluating optimal attractor networks across a range of weight sparsities, we demonstrate that there exists a trade-off between the noise robustness of memories and their efficiency in terms of synaptic resource use, as predicted from fundamental information theory. We find that an optimal level of pruning is achieved when each synapse contains only a small number of plasticity expression sites ($\sim$2-3 internal components), at which point the amount of retrievable information stored per synapse is maximized. Our model also predicts that intrinsic noise fluctuations should scale sublinearly with synaptic weight, which we corroborate in a new analysis of multiple published datasets on synaptic fluctuations. We finally offer predictions regarding the synaptic and cognitive properties of sleep-based memory consolidation that can guide future experimental studies.

3.2 Results

3.2.1 Memory consolidation as sparse optimization

To focus our analysis on the conceptual and computational aspects of long-term memory consolidation, we consider the simplest representation of a local circuit of cortical pyramidal cells, namely a recurrent network of N excitatory binary neurons in discrete time (Fig. 3.1a).

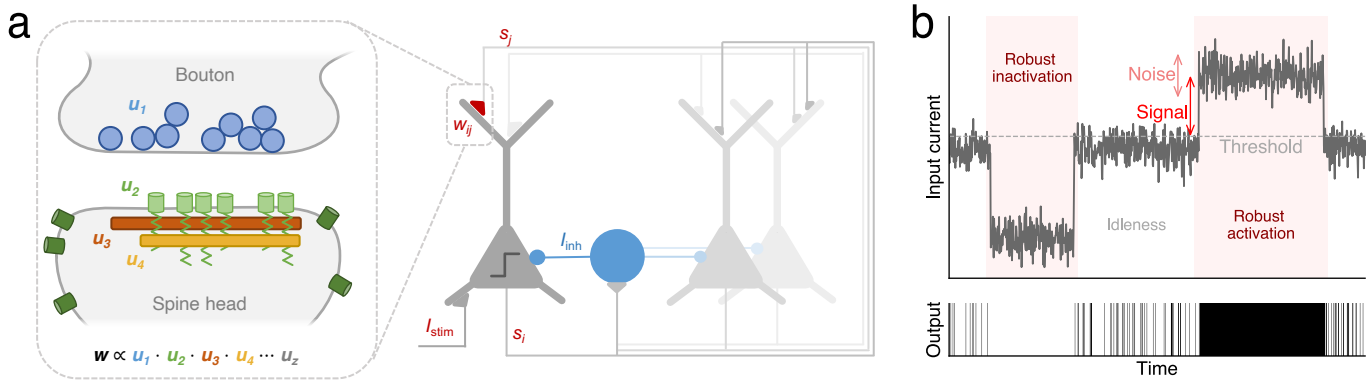


Figure 3.1: General model schematics. **(a)** Diagram of the circuit together with the synapse model (in box). The circuit is constituted of binary, excitatory neurons (gray) that are recurrently connected with non-negative connection weights, and receive inhibitory input from a single, common source (blue). Each recurrent weight (see box) is a product of multiple factors that represent the efficacy of sub-synaptic components (e.g., receptor concentration and scaffolding protein content). Note that no mathematical distinction is made between basal or apical inputs, and we only include them in the illustration as an interpretation of external and recurrent currents, respectively. **(b)** Illustration of input dynamics during idleness and noise-robust recall (pink areas) in a single neuron. Consolidation maximizes the signal-to-noise ratio of the current, where the signal is the smallest current deflection during recall (i.e., only the second deflection).

At each point in time t, each neuron $i = 1, \ldots, N$ is characterized by an output state $s_i(t)$, which signifies if the neuron is active ($s_i = 1$) or inactive ($s_i = 0$). Biologically, these states can be seen as representing brief intervals of elevated or suppressed firing (Cossart et al., 2003). A neuron assumes the active state only if the sum of its input currents reaches a positive value. The total input current, in turn, is comprised of four terms, according to

$$I_i(t) = \sum_{j=1}^{N} w_{ij} s_j(t-1) + I_{\text{stim},i}(t) - I_{\text{inh},i}^{(\text{slow})}(t) - I_{\text{inh}}^{(\text{fast})}(t) \tag{3.1}$$

where the first term represents the excitatory synaptic input from all other neurons in the network, with $w_{ij} \geq 0$ denoting the connection strength from neuron j to i. This can be seen as a representation of cortical top-down input, as anatomical data has demonstrated that recurrent connections carrying associative information among pyramidal cells primarily project to dendritic spines in the apical dendritic tree (Larkum, 2013). During training, each neuron receives a second excitatory current $I_{\text{stim},i}$, which represents a stimulus-driven, bottom-up input originating from preceding areas in the cortical processing stream. Finally, each neuron receives the inhibitory currents $I_{\text{inh},i}^{(\text{slow})}$ and $I_{\text{inh}}^{(\text{fast})}$, which regulate the balance between excitatory and inhibitory inputs on slow and fast time scales, respectively, and ensure that the total output activity in the network is stable over time (see Methods).

In our mathematical analysis of the storage properties of the network, we concentrate solely on the recurrent connections w_{ij}. This is motivated by experimental work suggesting that recall of long-term memory primarily relies on top-down projections among pyramidal cells. As such, both I_{stim} and I_{inh} serve only as auxiliary parameters, for the purpose of training

and stabilizing the network, respectively.

Over the course of one day of simulated learning, the network is trained to store M memories. Each memory corresponds to a random pattern of active and inactive neurons, where the desired activity of neuron i in pattern $\mu = 1, \ldots, M$ is described by the binary variable ξ_i^μ. The probability of a neuron being active in a pattern is given by the activity level $0 < f \leq 0.5$.

We assume that the goal of memory consolidation is to tune the excitatory connections of each neuron so as to maximize the noise-robustness with which stored memories are recalled. The robustness of a single neuron i at the moment of recall can be quantified as the deflection of its input current from the activation threshold. Over a set of multiple patterns, we define the neural robustness as the smallest deflection across the whole set (Fig. 3.1b), and denote this $\Delta I_{\text{min},i}$ (see Methods). To make this metric independent of the parameterization of our neuron model, we transform it from a current to a distance in the space of neural activity, by normalizing with a scaled sum of the input weights. The result is a generalized error margin

$$K_{q,i} = \frac{\Delta I_{\text{min},i}}{(\sum_j w_{ij}^q)^{1/q}} \tag{3.2}$$

where q is a positive scaling factor that determines which type of distance metric that is used to measure the margin. Index i will be omitted to simplify the notation.

We now define our goal in mathematical terms as finding the weights that store all memories as stable attractors in a way that maximizes K_q. The characteristics of such a solution critically depends on q, which acts as a regularizer of the optimization. For example, maximizing $K_{q=2}$ is equivalent to maximizing the signal-to-noise ratio (SNR) of the input current (see Methods). Although this solution produces optimal noise robustness (Kepler & Abbott, 1988; Krauth et al., 1988) and can be found with conventional machine learning techniques, it is highly non-sparse, since the solution exhibits a large number of small weights (Amit et al., 1989). This is undesirable for three reasons. First, a dense connectivity would imply a prohibitively high cost on metabolic energy, given that synapse maintenance is a primary source of energy consumption in the brain (Harris et al., 2012). Second, without any pruning, all available connections are used at once, which prevents any recycling of synaptic resources for continual learning of memories across separate training sessions. Finally, a dense connectivity between excitatory neurons directly disagrees with anatomical data (Thomson & Lamy, 2007; Lefort et al., 2009; see Introduction).

For $q < 2$, the normalization factor in K_q becomes more influenced by small weights at the expense of large ones. This implicitly forces the optimal connectivity to sparsen, in order for K_q to be maximized.

3.2.2 Learning with complex synapses and memory replay

How should a learning rule be constructed to maximize K_q in a way that is consistent with empirical observations of synaptic dynamics? To answer this, we first note that empirical synaptic strength, as measured in the postsynaptic potential or current, is an aggregate

quantity that is determined by the interaction of several protein complexes that combine to form the internal structure of a synapse (Nishiyama & Yasuda, 2015). During induction of long-term potentiation or depression, structural and chemical changes cascade throughout this molecular interaction network, causing the concentration and configuration of each component to be altered over the course of seconds to minutes. This ultimately results in an increase or decrease in the combined functional strength of a synapse.

We model this internal synaptic structure by expressing each weight w_{ij} as the product of z internal components u_{ijk}, where $k = 1, \ldots, z$, so that $w_{ij} \propto u_{ij1} \cdot u_{ij2} \cdots u_{ijz}$ (Fig. 3.1c). Each variable u can be seen as the relative concentration (or efficacy) of a collection of one or more subcellular building-blocks that are necessary to form a functional connection. In addition, all components are assumed to be linked to each other in a signaling cascade of the type proposed by Benna & Fusi (2016). To simplify the analysis, we further assume that the dynamics of the internal signaling evolves over a timescale that is much shorter than the behavioral timescale that governs stimulus encounters and learning in the entire neural circuit. The internal components can therefore be assumed to be in equilibrium, so that $u_{ij1} = u_{ij2} = \ldots = u_{ij}$. This yields the final synapse model

$$w_{ij} \propto u_{ij}^{z} \,, \tag{3.3}$$

where the proportionality constant will be set to one.

The maximization of the robustness K_q can now be implemented by modifying the internal synaptic factors u instead of the entire weight w, by letting the network carry out the following three-step consolidation algorithm:

(0) *Few-shot learning*: Prior to consolidation, we assume that each stimulus pattern already has been stored as a stable attractor, albeit with sub-optimal robustness. This can, for example, have been the result of one- or few-shot learning, during which the network has been subjected to each pattern during a period of high learning rate (this step is outside the scope of this study and will not be further elaborated).

(i) *Replay and tagging*: Each memory pattern is presented to the network as a brief cue, so that it is successfully recalled. Over the course of a single replay cycle, each neuron i separately identifies the pattern $\mu_i^\star$ that generates the smallest current deflection. Weights that receive presynaptic input during that pattern are tagged, with an LTD-tag if the deflection is hyperpolarizing or with an LTP-tag if the deflection is depolarizing.

(ii) *Weight update*: At the end of a replay cycle, tagged weights have their internal components updated according to

$$\Delta u_{ij} \propto u_{ij}^{z-1} \tag{3.4}$$

where the update is positive if the weights have been tagged for LTP, and negative for LTD.

(iii) *Weight normalization*: All synaptic components in neuron i are divided by a factor

proportional to $(\sum_j u_{ij}^2)^{1/2}$.

Steps (i)-(iii) are repeated until convergence has been reached. Note, however, that with a slow enough learning rate, the algorithm is stable and can be repeated indefinitely, without risking that weights grow unrealistically large. The end results, for any chosen $z = 1, 2, 3, \ldots$, is a maximization of K_q with $q = 2/z$ for each individual neuron (see Appendix A3.1). In other words, applying the consolidation algorithm with a high number of synaptic components is equivalent to maximizing K_q with a small q, and leads to increased sparsification.

It is important to highlight the fact that the above algorithm is not phenomenological, but entirely derived from normative assumptions. This is equally true for the reparameterization of the weights, which originates from a of well-established machine learning technique for implicitly biasing gradient-based optimization to find sparse solutions (Hoff, 2017; Amid & Warmuth, 2020) (see Methods). An important property of this approach is that it ensures that the weight normalization in step (iii) always is multiplicative, even as the learning rule prunes connections.

In continuous time, the dynamics of the complete weights w_{ij} throughout the consolidation process can be described compactly with the differential equation

$$\frac{\mathrm{d}w_{ij}}{\mathrm{d}t} \propto \underbrace{\left(1 - \frac{\sum_j w_{ij}^{2/z}}{\text{const.}}\right) w_{ij}}_{\text{heterosynaptic}} \quad \pm \quad \underbrace{\eta \xi_j^{\mu_i^\star} w_{ij}^{2(1-\frac{1}{z})}}_{\text{homosynaptic}} \tag{3.5}$$

where η is a learning rate and $\xi_j^{\mu_i^\star}$ a binary tagging variable that is active only for those weights that have been flagged for consolidation during memory replay. This formulation further demonstrates that the process of weight consolidation comprises two distinct processes: a homosynaptic part that is active only when a weight has been tagged in memory replay, and a heterosynaptic process that always is active and functions as a homeostatic mechanism that prevents inputs from growing too large. The hoterosynaptic term reproduces and generalizes previously studied homeostatic plasticity models (Toyoizumi et al., 2014) (see Appendix A3.2).

3.2.3 Multiplicative synapses produce optimal storage efficiency

We first characterized the connectivity and storage properties of attractor networks trained to maximize K_q for different values of q. We numerically optimized a recurrent network of $N = 1000$ neurons to store M binary patterns with activity level f. The storage load $\alpha = M/N$ ranged from roughly 8 % to 70 % of the maximal load α_c, while the activity level f was set to 50%, 10%, 5%, 3% or 1 %. The weight sparsity of the solution was varied by choosing $z = 1, 2, 3$ or 4, which corresponds to $q = 2, 1, \frac{2}{3}$ and $\frac{1}{2}$.

We quantified storage robustness by computing the average SNR across all neurons (Fig. 3.2a), and by testing the network's ability to successfully retrieve patterns after being presented with distorted cues (Suppl. Figs. A3.1 and A3.2). Distortions were randomly

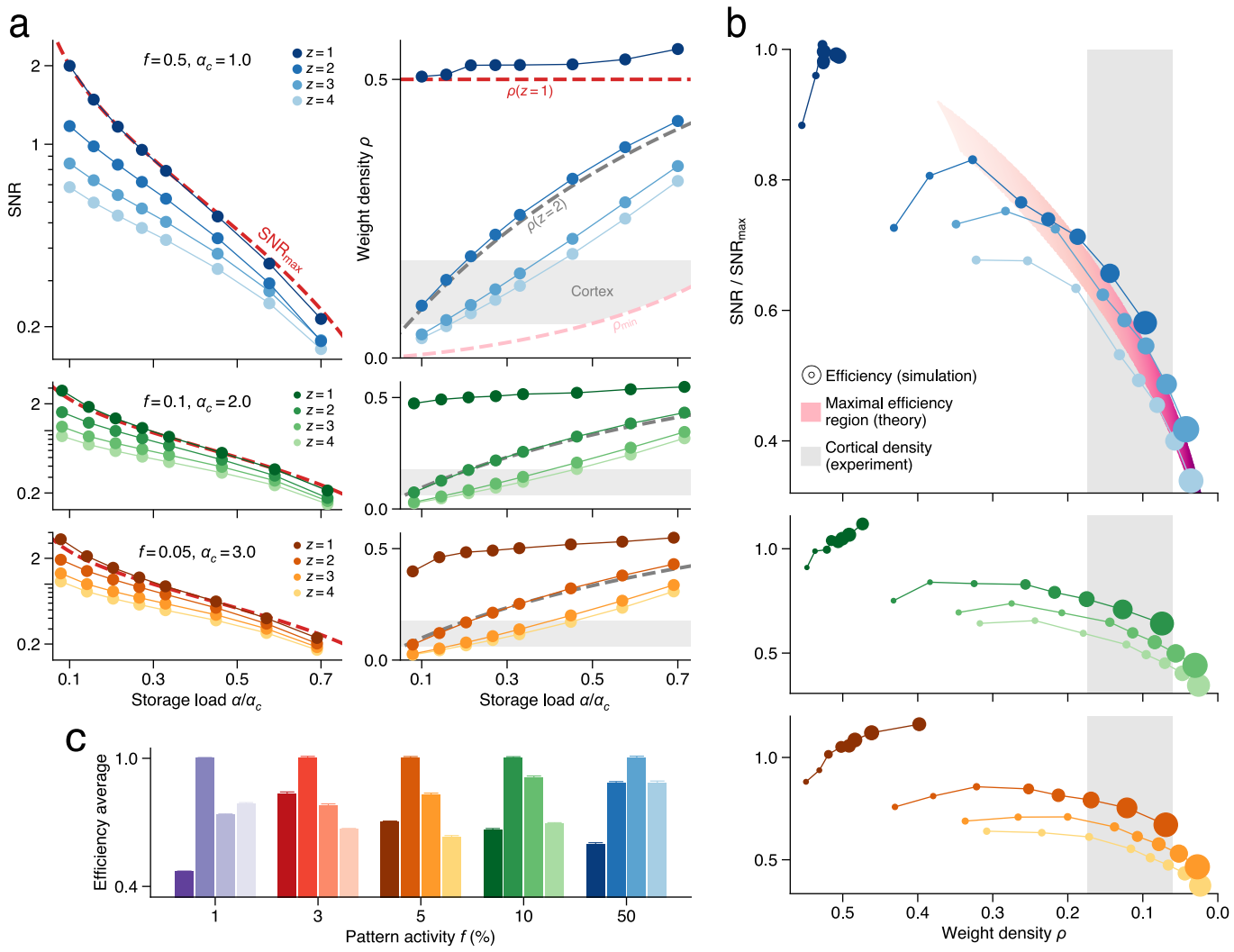


Figure 3.2: Attractor networks at dense and sparse optimality. **(a)** Average SNR (left) and weight density ρ (right) as a function of storage load for pattern activities $f = 0.5$ (top), $f = 0.1$ (middle), and $f = 0.05$ (bottom). Circles indicate mean over at least 8 independent simulations (SEM is smaller than markers and omitted). Dashed lines represent theoretical solutions. The gray area marks the mean $\pm$ SEM for ρ as estimated from 124 datasets on cortical connection probability in mice, rats, cats, and ferrets (Zhang et al., 2019). **(b)** All networks organized in a two-dimensional space according to robustness and sparsity. The pink region represents the optimal trade-off between the two quantities, as estimated by theory. **(c)** The efficiency, averaged over all storage loads and distortion levels (mean $\pm$ SD). This is highest at $z = 2$ for all $f < 0.5$.

introduced in each pattern by independently flipping bits in such as way so that the new, corrupted pattern retained the same activity level as the original pattern (see Methods). We found that networks with $z = 1$ had the highest SNR and, consequently, could tolerate the highest level of noise before memory recall collapsed. This result agreed well with the theoretically predicted maximum SNR (Gardner, 1988) and was expected given that an optimization with $z = 1$ directly maximizes the SNR. However, the most robust solution also had the highest connection density ρ (Fig. 3.2a). As z was increased, the solution became increasingly sparse, at the expense of the SNR and noise tolerance, which also decreased. Thus, optimal storage with sparse connectivity resulted in a decrease in the size of the attractor basins, and a deteriorated ability to successfully recall memories from noisy cues. Notably, however, only networks with $z \geq 2$ exhibited connection densities comparable to those measured in cortex (Fig. 3.2a, gray area), which suggests that memory robustness alone is an insufficient principle to model cortical long-term memory.

In order to quantify the trade-off between robustness and synaptic resource use, we computed how much retrievable information each network was able to store per functional connection as

$$Q = -\frac{\hat{\alpha}}{\rho}\left[f \log_2(f) + (1-f)\log_2(1-f)\right] \tag{3.6}$$

where $\hat{\alpha}$ is the storage load that can be retrieved during testing. We refer to this quantity as the *efficiency*: a network with higher Q is capable of recalling more information using fewer connections, which, in turn, indicates that its storage is more compressed and efficient. In the noise-free setting (results not shown), a higher z implicitly forces the network to become sparser and thereby more efficient. While networks with $z \leq 2$ cannot exceed 2 bits/synapse, which is obtained only when storage is saturated, this limit can, in fact, be surpassed with $z > 2$. The consequence, however, of each connection carrying more information is that every erroneous bit in the input has a more disruptive effect on error-correction and pattern retrieval, thus causing a deteriorated robustness.

To estimate the efficiency in a noisy setting, we tested how many of the stored pattern each network could successfully retrieve after being provided with distorted cues. We computed the efficiency averaged across a range of distortion levels. The results are presented for each network, at each load, in a two-dimensional space (Fig. 3.2b) according to the network robustness ($\mathrm{SNR}/\mathrm{SNR}_{\max}$) and connection density. In this representation, one can see that networks with $z = 1$ consistently produce highly robust, yet highly dense, configurations, which cause the efficiently to be relatively low. As z increases, networks find progressively sparse, but less robust solutions to the consolidation problem. However, the combination of robustness and sparsity that is theoretically predicted to produce maximal robustness is only reached with moderately pruned networks (i.e., $z = 2, 3$), suggesting that these networks are more efficient. To verify this, we computed the average efficiency across all storage loads and distortion levels (Fig. 3.2c). For all levels of pattern activity f, the efficiency grand average was consistently highest for networks with $z = 2$ or 3. The same result was obtained in tests with isotropic noise (Suppl. Fig. A3.3). In other words, intermediately pruned networks were, on average, able to maintain a higher amount of recallable information per synapse compared to both denser and sparser networks, across the same range of storage

loads and pattern activity levels.

These results are best explained within the framework of information theory. The process of cuing an attractor network with a distorted memory and synchronously updating it once can be likened to the transmission of a binary message through a noisy channel (Fig. 3.2e). The receiver, in this case, is the same network in the next time step, which sees the previous neural outputs and attempts to decode and error-correct them in order to retrieve the original message. The noisy-channel theorem states that any such system should exhibit a trade-off between its ability to correct errors and the compression of information in the message. This is indeed what we find, as the network with the highest number of connections also is the most robust.

From a biological perspective, the case $z = 2$ is particularly noteworthy. In Eq. 3.5, this value yields a purely multiplicative learning rule, with a homeostatic component that preserves a constant average connection weight. In a network with stable neural activity, this homeostatic rule is precisely equivalent to a regulation of the average input current (see Appendix A3.2). Our results demonstrate that a satisfactory degree of synaptic pruning can be achieved with a multiplicative plasticity model that incorporates biologically plausible homeostatic scaling.

3.2.4 Simultaneous consolidation and pruning with multiplicative synapses in sleep

In order to implement the consolidation algorithm in section 3.2.2 in a circuit with both excitatory and inhibitory realistic connections, it remains necessary to determine how the inhibitory current $I_{\text{inh}}^{(\text{slow})}$ can be optimized in a biologically plausible way. Deriving a learning rule for $I_{\text{inh}}^{(\text{slow})}$ directly from the maximization of K_q is problematic, as the result implies anti-Hebbian inhibitory plasticity, which is incompatible with experimental data (Hennequin et al., 2017).

In the specific case $z = 2$, this issue can be resolved by rewriting the combination of homeostatic and inhibitory plasticity in a way that allows for optimization of excitatory weights with a fixed level of inhibition (see Appendix A3.7). The resulting learning algorithm is identical to the one stated in section 3.2.2, except for the following modification of the last two steps:

(ii) & (iii) *Weight update with* $z = 2$: **All** weight components are updated according to

$$\Delta u_{ij} \propto (\eta_{\text{het}} + \eta_{\text{hom}} \xi_j^{\mu_i^*}) u_{ij} \tag{3.7}$$

where η_{het} is a low, heterosynaptic learning rate and η_{hom} is an added (homosynaptic) learning rate that is active only if a weight is tagged. The change is positive if weights are tagged for LTP, and negative for LTD.

It is worth reiterating that this learning rule is not phenomenological, but directly emerges from gradient-based maximization. There are three important features that should be

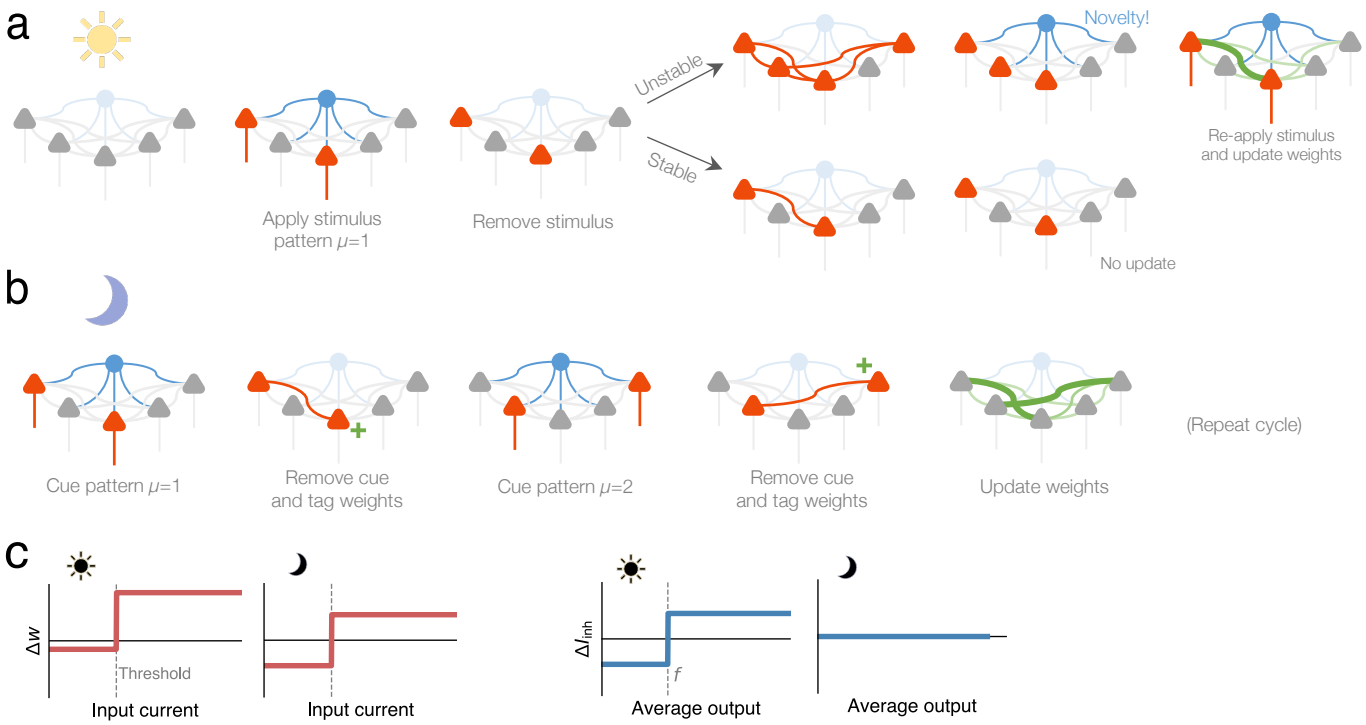


Figure 3.3: Schematic of consolidation over a day with multiplicative plasticity. **(a)** In wakefulness, few-shot learning occurs by clamping the network with an external stimulus, determining if the stimulus is novel, and updating the recurrent weights. **(b)** In sleep, replay tags the most novel pattern for LTP/LTD, and weights are updated at the end of the replay cycle. **(c)** In wakefulness, LTP outweighs LTD, while, in sleep, LTP and LTD is balanced (red curves). For inhibition, updates are carried out at every step to maintain desired network activity during wakefulness. During sleep, inhibition is fixed (blue curves).

highlighted. First, this reformulation can only be done with the weight reparameterization using $z = 2$, as it critically relies on multiplicative weight updates.

Second, this reformulation requires neither explicit inhibitory plasticity nor explicit weight scaling. The magnitude of the weights is, instead, regulated implicitly by the fixed $I_{\text{inh}}^{(\text{slow})}$. This is particularly practical for modeling sleep-based memory consolidation and pruning. During sleep, cortical neurons alternate between dramatically different states of activity, which are believed to be caused by a rapid replay of memories occurring as part of the consolidation process. Modeling inhibitory plasticity under these conditions would normally be problematic, since a consistent read-out of neural activity would be hampered by the sudden shifts between qualitatively different network states. With the model in Eq. 3.7, this complication is avoided altogether.

Third, Eq. 3.7 predicts that plasticity during consolidation is expressed in all connections, and with the same sign; the weight change is, however, stronger is connections that were active during tagging than in those that were silent. This can be interpreted as a form of plasticity diffusion or cross-talk, whereby LTP- or LTD-triggering molecules spread from tagged connections, through the dendrites, and induce attenuated forms of heterosynaptic

LTP or LTD in neighboring connections. It is also consistent with experimental evidence showing that calcium spikes in the dendritic arbors of pyramidal cells are significantly amplified and spread-out during sleep (Li et al., 2017).

To demonstrate how the learning rules in Eqs. 3.5 and 3.7 can be incorporated in a single, self-consistent model of memory formation and consolidation over the course of a day, we simulated how a network of $N = 1000$ neurons with $z = 2$ can learn to optimally store $M = 300$ patterns in two separate phases. In the first phase, representing wakefulness (Fig. 3.3a), the network performed self-supervised few-shot learning, while actively being regulated by homeostatic scaling, inhibitory plasticity, and fast inhibitory feedback. Each pattern μ was presented to the network in the form of a strong excitatory input current I_{stim} that depolarized neurons that needed to be active ($\xi_i^\mu = 1$), while the stabilizing effect of the fast inhibitory feedback current $I_{\text{inh}}^{(\text{fast})}$ immediately hyperpolarized neurons that needed to be silent ($\xi_i^\mu = 0$). In the following time step, $I_{\text{inh}}^{(\text{fast})}$ either dropped to zero, indicating that the pattern was a stable attractor, or remained non-zero, which indicated that the pattern had not yet been correctly stored. The novelty of a pattern could in this way be read out directly from $I_{\text{inh}}^{(\text{fast})}$. Weights were updated homo- and heterosynaptically if the novelty signal was triggered; otherwise, only homeostatic scaling was performed. Furthermore, only one u-factor per weight was allowed to change during wakefulness, while the other one was kept fixed. Tonic inhibition, represented by $I_{\text{inh}}^{(\text{slow})}$, was also updated with a Hebbian, inhibitory plasticity rule so that the average input current was properly balanced (see Methods). Patterns were presented in a randomized order until the novelty signal stopped being triggered, at which point all patterns had been correctly encoded by the network.

In the second phase, representing sleep (Fig. 3.3b), the network performed memory consolidation and pruning. Homeostatic scaling was inactivated, and the tonic inhibitory current $I_{\text{inh}}^{(\text{slow})}$ was kept fixed (Fig. 3.3c). Memory replay was carried out by cuing and recalling every stored pattern in the network. In each replay cycle, every neuron i individually tagged the input that was least robust, i.e., most novel. At the end of the cycle, weights were updated according to Eq. 3.7, with $\xi_j^{\mu_i^\star}$ being the tagged pattern. Both u-factors were now allowed to change.

Figure 3.4 demonstrates two simulation examples, with low ($M = 200$) and high ($M = 350$) storage load. During wakefulness, few-shot learning rapidly encoded all patterns as stable attractors, albeit in a suboptimal manner. Only 20 to 40 presentations per pattern were needed until novelty stopped being triggered and all patterns could be recalled without errors (Fig. 3.4a), but the SNR was on average low. During sleep, however, consolidation quickly improved the SNR by more than an order of magnitude (Fig. 3.4b), while synaptic connections were pruned by decaying to zero at an exponential rate, resulting in a substantial increase in network efficiency. The effect of pruning was also detectable in the slight dominance of LTD over LTP during sleep-based plasticity (Fig. 3.4b, inset). The connection density finally converged to values close to the theoretical predictions.

The processes of consolidation and pruning were not entirely concurrent, but could largely be separated into two qualitatively distinct phases (Fig. 3.4b, light and dark blue backgrounds). In the first epochs of sleep, minor weight adjustments quickly produced a near-maximal SNR,

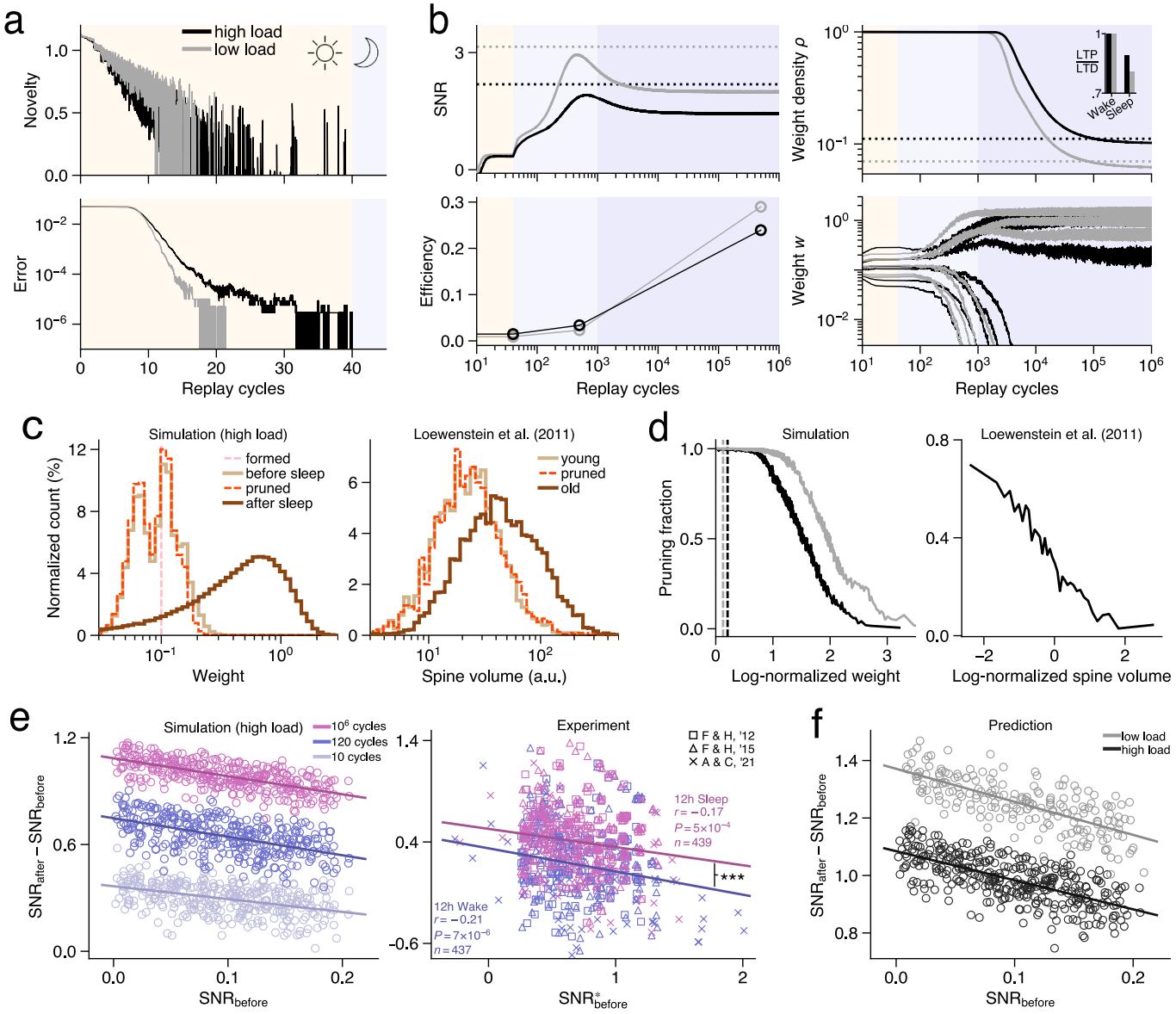


Figure 3.4: Simulated consolidation over a single day with multiplicative plasticity. **(a)** Pattern novelty and recall error for high load ($\alpha = 0.35$) and low load ($\alpha = 0.20$), during wakeful learning (yellow background). **(b)** Plots of SNR, efficiency, connection density, and individual weights, during sleep-based consolidation (blue background; the two phases of consolidation are indicated by shade). Dashed and dotted lines are theoretical results for maximum SNR (for any q), and minimum ρ (for $q = 1$), respectively. Inset: ratio of LTP to LTD. **(c)** Left: Weight distribution at formation (epochs = 0), before sleep, and after sleep (survived and pruned), for the high load simulation. Right: Volume distribution of new, pruned, and old dendritic spines in experimental data. **(d)** Pruning fraction in simulation (left) and experimental data (right) as a function of weight (left) and spine volume (right). **(e)** SNR-change after, compared to before, consolidation, in simulations (left; each circle is a pattern) and in humans (left; each points is a subject). Behavioral data has been slightly jittered for clarity. **(f)** Simulated SNR-change at the end of sleep for different loads.

without much pruning occurring. The effect of pruning was only visible at a later stage, after unimportant weights had been sufficiently depressed, or removed, and remaining weights had been further potentiated. At this point, the weight configuration started approaching a sparse solution and the SNR dropped from the maximum.

The distributions of pre-sleep connections and pruned connections closely overlapped each other (Fig. 3.4c, left), while the small number of connections that survived sleep were, on average, stronger. We compared this with the experimental data from Loewenstein et al. (2011, 2015) and found similar results (Fig. 3.4c, right; more details in Methods). The distribution of dendritic spine volume for young spines (age $\leq$ sampling interval Δt) closely matched that for pruned spines, while old spines (age $> \Delta t$) generally were larger.

An analysis of the fate of individual connections (Fig. 3.4d, left) revealed that the probability of pruning gradually decreased as a function of connection strength, which, again, agreed with the experimental data (Fig. 3.4d, right). Connections that at the beginning of sleep had a strength close to initialization were completely pruned, while most of those that had been potentiated survived. Notably, we found that the shape of the pruning curve depended on the storage load, so that consolidation of fewer patterns caused a higher fraction of strong weights to be pruned. This adjustment was automatically imposed by the learning algorithm, without any need for external tuning. The additional pruning was also reflected in the LTP to LTD imbalance, which was larger when consolidating fewer patterns.

3.2.5 Preferential consolidation of weakly encoded memories in sleep

Behavioral studies on sleep-based consolidation in humans have demonstrated that memories with weak initial encoding are strengthened to a higher degree, and thus benefit more, from sleep (Schapiro et al., 2018; Denis et al., 2020). To evaluate if this effect was reproduced by our consolidation model, we compared the SNR improvement of each memory over the course of sleep, relative to the initial value. We found a significant negative correlation (Fig. 3.4e, left), indicating that memories that were weakly encoded before sleep had been strengthened more after sleep. This is caused by a ceiling effect: as the consolidation algorithm pushes the SNR of each memory close to the maximal limit, memories that start with a low SNR will inevitably exhibit a larger improvement than those starting with a high SNR.

To quantitatively compare these results with behavioral data, we re-analyzed three large, published datasets on sleep-based consolidation of declarative memory (Fenn & Hambrick, 2012, 2015; Ashton & Cairney, 2021). In each study, humans were tasked with learning 40 word-pair associations, and their recall performance was tested closely before and after a 12 h-interval of sleep or wakefulness. We applied the framework of signal detection theory and modeled the memory trace strength in each subject as a latent continuous variable that, at encoding time, is perturbed by normally distributed noise (Mickes et al., 2009). At test time, the model posits that only traces stronger than a subject-specific threshold can be recalled. We estimated the trace SNR in each subject (denoted SNR*) as the distance between the recall threshold and the inferred average trace strength (see Methods). This

allowed us to compare the SNR*-improvement in each subject after sleep to the SNR* before sleep. In agreement with our simulations, we found a significant negative correlation (Fig. 3.4e, right), again indicating that a weaker memory encoding prior to sleep was linked to a larger improvement in encoding after sleep.

Although we found the same effect in the wakefulness condition (Fig. 3.4e, right), the improvement in SNR* was systematically smaller compared to that produced by sleep, as indicated by the significant downward shift of the regression curve (two-tailed t-test, $P = 2 \times 10^{-6}$, $n = 876$; see Methods). This effect could be reproduced by our network by subjecting it to fewer replay cycles during consolidation (Fig. 3.4e, left). These results suggest that a consolidation process similar to that occurring during sleep also could takes place in wakefulness, albeit with reduced efficacy or over a shorter duration.

In addition, our model predicts that a similar upward or downward shift in SNR-improvement can be produced after full sleep-based consolidation by reducing or increasing, respectively, the amount of memories that are consolidated (Fig. 3.4f).

3.2.6 Intrinsic synaptic noise scales sublinearly with weight

The learning rule presented above crucially relies on parameterizing each connection weight w_{ij} as a product of multiple factors u_{ijk}, which represent the efficacies or concentrations of subsynaptic constituents. We therefore aimed to investigate if such a model can be tested with experimental measurements of synaptic dynamics. The past two decades of imaging cortical dendrites has demonstrated that the strength of a synapse is strongly correlated with its size (Holler et al., 2021), which, in turn, is highly prone to intrinsic noise; a phenomenon referred to as synaptic volatility (Mongillo et al., 2017; Ziv & Brenner, 2018). How would the addition of intrinsic noise in our model of the synaptic ultrastructure manifest in measurements of synaptic volatility? To answer this, we first note that intrinsic noise, as the name implies, is independent of activity-related, homosynaptic plasticity, and therefore present even when all glutamatergic transmission has been silenced. We assume that this noise reflects the combined sum of multiple internal noise sources caused by thermal fluctuations, such as spontaneous chemical reactions, conformational changes, as well as protein degradation and turnover. We therefore model this with a white noise term ϵ_{syn}, which is added to each subsynaptic component u_{ijk}. The result, under conditions of blocked excitatory synaptic transmission, is that each weight fluctuates according to the stochastic process

$$\frac{\mathrm{d}w_{ij}}{\mathrm{d}t} \propto \left(1 - \frac{\sum_j w_{ij}^{2/z}}{\text{const.}}\right) w_{ij} \; + \; w_{ij}^{1-\frac{1}{z}} \sigma_{\text{syn}} \epsilon_{\text{syn}} \; , \tag{3.8}$$

where σ_{syn} is a positive parameter that determines the amplitude of the noise fluctuations, and ϵ_{syn} is biased Gaussian noise. Hence, our model predicts that the size of intrinsic synaptic noise, whether potentiating or depressing, should scale as $\mathcal{O}(w^{1-\frac{1}{z}})$. For the maximally robust network, when $z = 1$, noise is purely additive and uncorrelated to weight strength, while for sparser networks, when $z > 1$, noise always scales sublinearly with w. It is only in the limit $z \to \infty$ that the noise term becomes proportional to the weight.

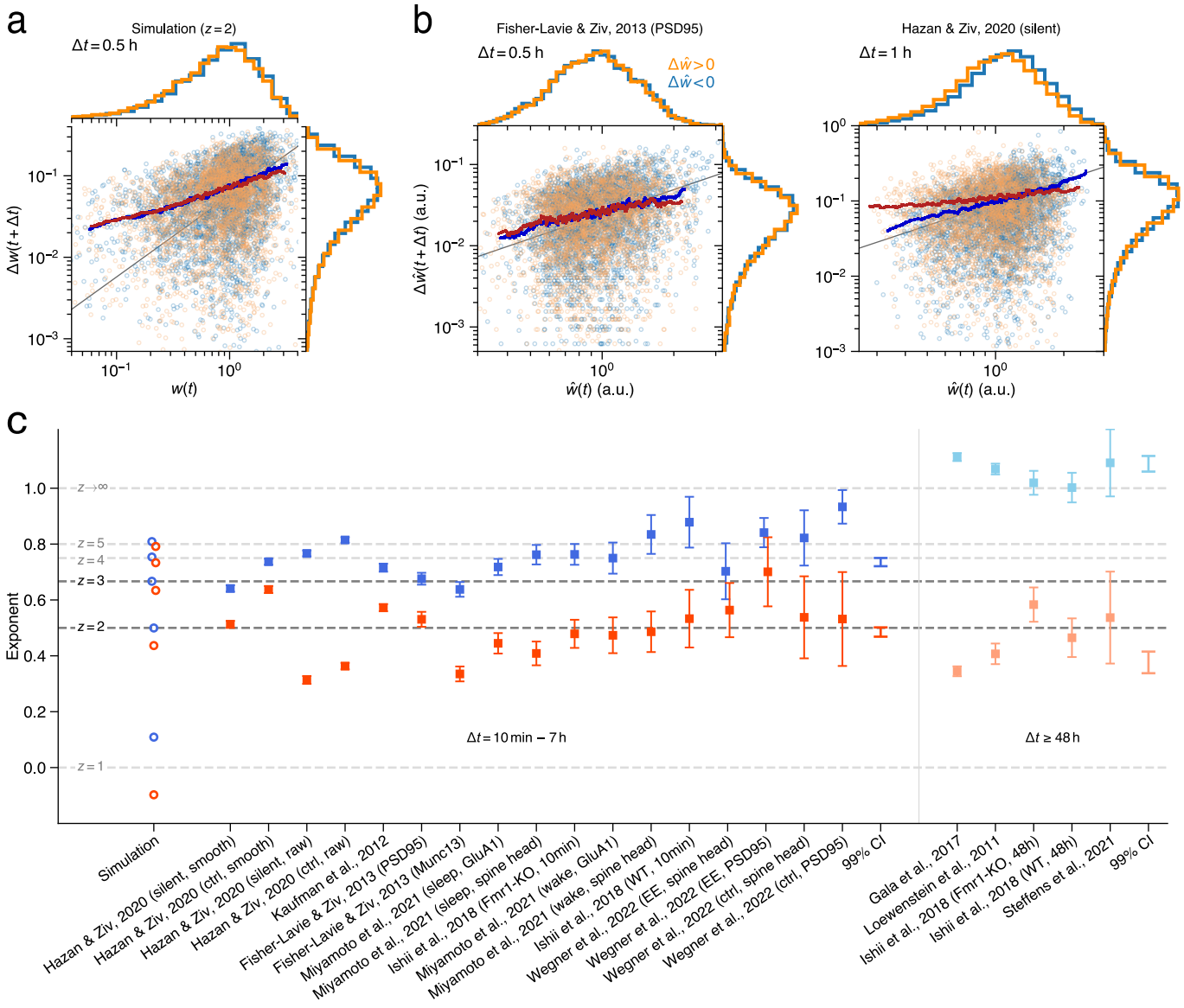


Figure 3.5: Scaling of synaptic fluctuations. **(a)** Simulated synaptic volatility of 1000 synapses governed by the stochastic process in Eq. 3.8 with $z = 2$ (see also Suppl. Fig. A3.4). **(b)** Two example datasets on synaptic strength change plotted as function of initial strength (circles are individual synapses). A moving average produces straight lines (dark red/blue for potentiation/depression), indicating a power-law relation. The exponent was estimated by bootstrapped linear regression (see Methods). **(c)** Estimated power-law exponents in simulated and experimental synaptic fluctuations (mean ± SEM). Experiments are grouped into short ($\Delta t = 10\,\text{min} - 7\,\text{h}$) and long sampling intervals ($\Delta t \geq 48\,\text{h}$), and summarized with a weighted 99% confidence interval. Labels contain a publication reference and a brief methodological descriptor; complete details are provided in supplementary Tables A3.4 and A3.5.

To compare this model with experimental data, we re-analyzed 22 published datasets of synaptic strength measurements from 9 separate studies (Loewenstein et al., 2011; Kaufman et al., 2012; Fisher-Lavie & Ziv, 2013; Gala et al., 2017; Ishii et al., 2018; Hazan & Ziv, 2020; Miyamoto et al., 2021; Steffens et al., 2021; Wegner et al., 2022). These publications span more than a decade of research and employ different measurement techniques, such as fluorescence microscopy, and super-resolution nanoscopy, both in cultured neurons and *in vivo* (see Suppl. Tab. A3.4 and A3.5 for details). Common to all studies, however, is that they contain measurements of synaptic strength, size, or a proxy of the two, from a large population of synapses (typically $\gtrsim 10^3$) that have been individually tracked over extended periods of time (typically ranging from 24 h to almost 30 d).

In each dataset, we first paired all synaptic strength changes $\Delta\hat{w}(t + \Delta t)$ between two consecutive measurements (at times t and $t + \Delta t$) with the initial strength $\hat{w}(t)$. We then separated the data into potentiation ($\Delta\hat{w} > 0$) and depression ($\Delta\hat{w} < 0$) and analyzed the two cases separately, given that the dependence on initial strength can differ qualitatively between LTP and LTD (Bi & Poo, 1998). Lastly, in order to reduce the effect of noise and outliers, we calculated the average absolute change $\langle|\Delta\hat{w}|\rangle$ as a function of initial strength using a moving average, and we plotted the results in logarithmic scale. The analysis revealed that, for both potentiation and depression, $\langle|\Delta\hat{w}|\rangle$ had a linear dependence on $\hat{w}$ in logarithmic space, indicative of a power-law relation in the original data (i.e., $\langle|\Delta\hat{w}|\rangle \propto \hat{w}^x$) (Fig. 3.5b). The exponent of the power-law, which is equivalent to the slope of the line in logarithmic space, was obtained with bootstrapped linear regression (see Methods).

In order to validate the analysis above, we also generated synthetic data of synaptic strength fluctuations by numerically simulating 1,000 independent realisations of Eq. 3.8 (Fig. 3.5a, see also Suppl. Fig. A3.4). This was analyzed in the same way as the experimental data and agreed well with theoretical predictions. The results are summarized in Figure 3.5c (circles).

For the datasets with high sampling frequency (i.e., short observational time intervals $\Delta t = 10\,\text{min}$ to 7 h) and large samples sizes, synaptic strength fluctuations displayed a sublinear scaling exponent, with a value of 0.49 ± 0.02 (99 % weighted confidence interval) for synaptic potentiation and 0.74 ± 0.01 for synaptic depression. These estimates were remarkably reliable and close to the range predicted by our synaptic noise model in Eq. 3.8 with $z = 2$, 3, and 4. It should be noted, however, that our model assumes that activity-dependent synaptic transmission is either negligible or entirely blocked, in order to make simple and precise predictions; the inclusion of extrinsic synaptic noise would make our model considerably more complicated. As such, the theoretical results are only approximately applicable to the experimental measurements, which, in almost all cases, are perturbed by the presence of extrinsic synaptic noise. The data by Hazan & Ziv (2020) is a notable exception, as this was acquired during a complete block of glutamatergic transmission. In this case, the noise scaling coincides almost exactly with the theoretical lines for $z = 2$ and 3, as we obtain 0.51 ± 0.01 for potentiation and 0.64 ± 0.01 for depression (mean $\pm$ SEM over 100 bootstrapped samples).

Datasets with smaller sample sizes or low sampling frequencies generally showed a higher estimated scaling exponent for synaptic depression, together with a larger error margin. This

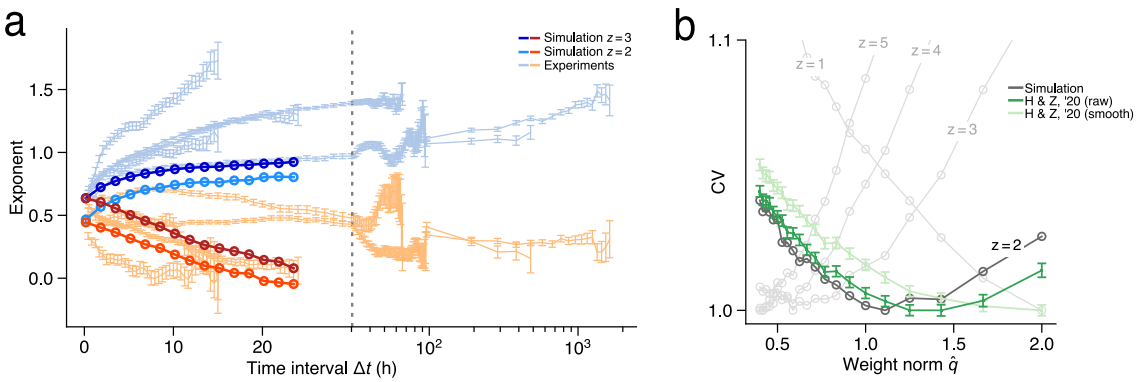


Figure 3.6: Simulations with $z = 2$ and 3 reproduce synaptic fluctuation statistics over time. **(a)** The estimated power-law exponent as a function of the sampling interval Δt in simulations and experimental data from Loewenstein et al. (2011), Kaufman et al. (2012), Fisher-Lavie & Ziv (2013), Gala et al. (2017), and Hazan & Ziv (2020) (mean ± SEM). **(b)** CV over 24 h (mean ± SEM) for different synaptic weight norms $(\sum \hat{w}^{\hat{q}})^{1/\hat{q}}$.

was particularly evident in the datasets with very long windows of time between observations ($\Delta t \geq 48$ h). While the exponent for synaptic potentiation decreased to 0.38 ± 0.04, it was generally higher than one (1.09 ± 0.03) for synaptic depression, similarly to previously reported results from an analysis of this type (Morrison et al., 2007).

The fact that synaptic depression consistently was found to have a larger scaling exponent than potentiation indicates that internal synaptic noise can be described by a stationary stochastic process, which predominantly potentiates weak synapses and depresses strong ones, thereby forcing synapses towards the mean of the strength distribution. This is consistent with past experimental literature showing that synaptic strength distributions are unimodal and stable over time.

In order to further analyze the effect of sampling frequency on synaptic noise scaling, we artificially increased the time window between measurements by sub-sampling the data, and we plotted the estimated scaling exponent as a function of the new Δt (Fig. 3.6a). For synaptic depression, the results within studies mirrored those across studies, as the scaling exponent tended to increase with larger Δt, and appeared to converge to values $\gtrsim 1$. For potentiation, the exponent decreased as a function of Δt to values ranging between 0 and 0.5. These trends were reproduced in the simulated data, even with different z values. These results demonstrate that estimates of synaptic noise scaling can be uninformative if the time between measurements is too long. Given that noise is state-dependent and non-linear, the total size of the perturbation accumulated by a synapse over long time intervals will only correspond to a time-averaged weight change. This may occlude the relation between instantaneous fluctuations and the weight, which can only be observed over short time windows.

3.2.7 Estimating homeostatic scaling from intrinsic synaptic noise

Our model of intrinsic synaptic dynamics does not only make predictions about the scaling of synaptic noise, but also implies qualitatively different forms of synaptic homeostasis. For a given z, the homeostatic term in Eq. 3.8 ensures that the $\frac{2}{z}$-th moment of all incoming weights (i.e. $\langle w^{2/z} \rangle$) is stable and close to the prescribed constant. For $z = 2$, this means that the term multiplicatively regulates the average weight, while for $z = 1$ it approximately regulates the variance. Hence, in the absence of external noise, our model predicts that the distribution of a neuron's incoming excitatory weights should exhibit a fixed $\frac{2}{z}$-th moment, while other moments are allowed to vary. To test this, we returned to the data collected during synaptic blocking in the study by Hazan & Ziv (2020). At each measurement, we calculated the norm $\|\hat{w}\|_{\hat{q}} = (\sum \hat{w}^{\hat{q}})^{1/\hat{q}}$ separately for each putative neuron, with different values of $\hat{q}$. We then computed the coefficient of variance (CV) for each norm over 24 h of experimental time, in order to quantify how much different moments of the $\hat{w}$-distribution varied over time. The same analysis was applied to the synthetic data generated from the simulations of Eq. 3.8.

Our simulated results confirmed theoretical predictions (Fig. 3.6b), as we found that the weight norm with smallest CV coincided with the value used in the homeostatic scaling. For example, in simulations with $z = 2$, the weight norm that fluctuated least over time was $\|w\|_1$. The experimental results were found to closely match the simulation curve with $z = 2$. This is consistent with the results in the previous section and lends further support to the bipartite synapse model. It also suggests that synaptic homeostatic plasticity regulates the average strength of incoming synapses, even in the absence of any synaptic input current.

3.3 Discussion

We have derived a family of synaptic learning rules that maximize the noise robustness of attractor memories in recurrent neural networks subject to varying degrees of synaptic pruning. We propose these learning rules as a general mathematical model of optimal memory consolidation under synaptic resource constraints. In our definition of optimal consolidation we assume the following scenario: A recurrent network has first been subjected to brief but intense sensory-driven stimuli whose neural activity patterns have been imprinted as stable attractors. However, the encoding is weak and far from optimal, and the purpose of consolidation is therefore to tune connections so as to maximally strengthen the encoding. Strength, or noise robustness, is in this context defined as an error margin measured with a chosen distance metric. In order to maximize the margin in a biologically realistic manner, our second fundamental assumption states that each synaptic weight can be parameterized as a product of factors, each quantifying a partial efficacy or potency of the synapse. With this, we derive a class of learning rules that maximizes memory robustness is a way that naturally incorporates memory replay, homo- and heterosynaptic plasticity, multiplicative homeostatic scaling, and rapid, automatic pruning, without thresholding or explicit synaptic regularization.

A long-standing problem in past theoretical work on synaptic plasticity has been the seeming

contradiction between experimental studies reporting that cortical connections are sparse (Thomson & Lamy, 2007; Lefort et al., 2009) and rapidly form and retract (Le Bé & Markram, 2006), while others find that individual synaptic changes are multiplicative (Turrigiano, 2008; Loewenstein et al., 2011); in the statistics literature, multiplicative learning rules are typically associated with smooth changes and dense solutions. Our synaptic plasticity model reconciles these observations, and displays both properties as part of a combined process of consolidation and pruning.

3.3.1 The robustness-redundancy trade-off

Using the derived synaptic learning rule, we have characterized the optimal storage in attractor networks under varying degrees of sparsity. It is important to note that this analysis differs in many ways from previous work on pruning in attractor networks, which typically has focused on either quenched (Gardner, 1989; Bouten et al., 1990) or annealed removal (Bouten et al., 1990; Chapeton et al., 2015) of a pre-defined number of connections. In order for this type of pruning to be implemented, the network would need to know the appropriate number of weights to remove, prior to the learning of a task; from the standpoint of biological plausibility, this is problematic. The amount of pruning would also need to be carefully chosen, as any *a priori* constraint placed on the connection density that deviates from the unconstrained optimal solution would cause the storage capacity to deteriorate (Bouten et al., 1990; Chapeton et al., 2015).

In contrast, we implement the pruning mechanism as a regularized optimization, without specifying the connection density explicitly. This allows the amount of pruning to be automatically adapted to each set of patterns. Notably, the pre-defined degree of regularization (determined by z) has no impact on the storage capacity, which always equals that of the unconstrained network. As long as all patterns are linearly separable, they can be stored using our learning rule with any positive z.

Regularization does, however, have a detrimental effect on robustness, as our results show that sparse connectivity leads to low SNR. An explanation using information theory and the noisy channel theorem is proposed in the main text.

Under the assumption that cortex can be modeled as an attractor network, our results demonstrate that neither storage robustness nor synaptic sparsity alone is sufficient to form an organizing principle for cortical circuits, as has previously been suggested (Krieg & Triesch, 2014; Brunel, 2016). Maximizing the former leads to unrealistically dense connectivity, while optimizing the latter causes noise-tolerance to be so low that memories are practically irretrievable. We argue, instead, that an optimal compromise is achieved at moderate levels of pruning, with $z = 2$ or 3. This maximizes the average storage efficiency, as measured by the amount of retrievable information per synapse.

3.3.2 Replay and sleep

In the special case $z = 2$, we argue that our synaptic learning is particularly suitable as a model of sleep-based memory consolidation. The learning rule can, in this case, be implemented using only memory replay, multiplicative weight updates, and synaptic cross-talk, with tonic inhibition and silenced external input. These conditions match remarkably well with the neurophysiological signs of sleep in hippocampus and neocortex. During slow-wave sleep (SWS), external input tends to be attenuated in favor of endogenously generated neural activity, which is highly structured and displays cyclic periods of elevated spiking followed by silence (Klinzing et al., 2019). The periods of heightened activity are synchronized across hippocampus and cortex, and likely reflect a sequential reinstatement of past experiences (Ji & Wilson, 2007; Schreiner et al., 2021). On the cognitive level, SWS has been shown to be crucial for declarative memory consolidation (Gais & Born, 2004).

In contrast, the function of rapid eye movement (REM) sleep remains more contested. Although this sleep stage also displays replay (Louie & Wilson, 2001) and silencing of external inputs (Aime et al., 2022), is is better characterized by elevated levels of synaptic plasticity, wide-spread dendritic calcium activity, and pruning (Li et al., 2017; Zhou et al., 2020). This is believed to support a synaptic stabilization and optimization of memory traces that have been distributed and consolidated on a systems level during preceding SWS.

In our model of sleep, we do not distinguish between different sleep-stages, but instead aim to demonstrate how consolidation and pruning can be accomplished with a minimal number of auxiliary plasticity mechanisms. Nonetheless, it is interesting to note that our synaptic learning rule inherently produces an optimization trajectory that exhibits two phases. It first approaches a dense solution, which achieves a near-maximal robustness, without any noticeable pruning. Later, upon further optimization, the solution sparsifies, at the expense of robustness, which displays a small drop. These two phases are reminiscent of the complementary functions of SWS and REM-sleep, and suggests that the biphasic process of sleep-based consolidation could be explained within a single modeling framework.

While our model of memory replay follows directly from the mathematical derivation of optimal learning, it has a simple, intuitive interpretation. When a pattern that is completely new and unrelated to any previous memory is cued to the network, it generally elicits a weak neural response, which means that the current deflection from the threshold, on average, is small and often in the wrong direction. The pattern can therefore not be correctly recalled and has an error margin close to zero. The purpose of replay in our model is for each neuron to update the synapses of the memory with smallest error margin, that is, the memory that is perceived as most novel. Memories with a strong encoding, either because they are old or resemble previously seen patterns, will never undergo synaptic change. This form of replay is a parsimonious consolidation mechanism that concurrently strengthens the trace of all memories by only updating the weights of the most labile pattern. Similar prioritization in replay and consolidation has recently been shown to occur in hippocampus during wakefulness (Schapiro et al., 2018) and sleep (Denis et al., 2021).

Our consolidation model offers an alternative hypothesis to two previously proposed theories

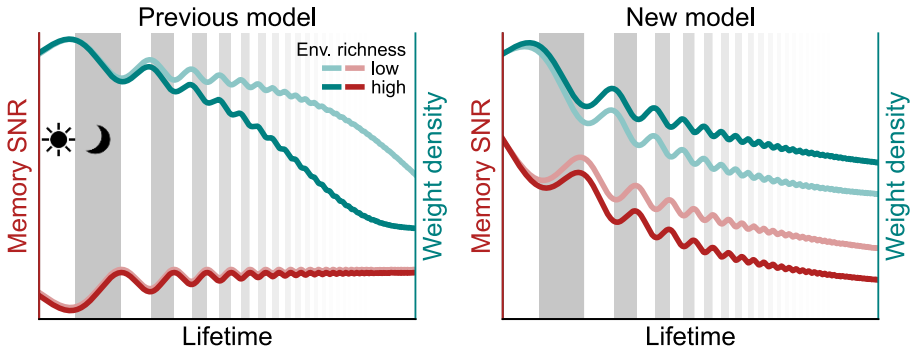


Figure 3.7: Schematic of predicted consolidation across lifetime. In previous models (left), robustness is assumed to be capped. Networks storing many patterns (high env. richness) will end up with *sparser* connectivity than those storing few patterns (low env. richness). In the new model (right), robustness is maximized at each bout of consolidation. Networks with high storage will end up with *denser* connectivity than those with low storage.

of sleep-based memory consolidation. The first, which we refer to as the *unlearning theory* (Crick & Mitchison, 1983; Hopfield et al., 1983), argues that the function of dream sleep is to replay and unlearn spurious attractors, in order to indirectly increase the robustness of desired memories. We, instead, demonstrate that the same goal can be accomplished by replaying only information that already has been seen, without any need to identify spurious memories or to invoke reversed plasticity.

The second hypothesis, termed the *over-fitted brain* (Hoel, 2021), argues that learning in wakefulness causes the brain to over-fit sensory data, and that the purpose of dream sleep is to improve generalization by replaying and relearning noisy variants of stored information. While noise injection can be used to improve robustness in attractor networks (Rubin et al., 2017), it is mathematically equivalent to directly maximizing the error margin, as done in our work (see Appendix A3.1). An added benefit of our approach is that it is metabolically more sparing. Even though it necessitates a structured replay-and-update procedure, it updates only one pattern per replay cycle, and no noise is needed.

3.3.3 Implications for life-long learning

Over the course of an animal's development, memories are gradually accumulated and incorporated into the brain through an interlaced sequence of coarse learning in wakefulness and consolidation in sleep (see Fig. 3.7 for illustration). In this setting, our consolidation model predicts that both memory robustness and connectivity would follow a decreasing trend over long periods of time, given that storage of a larger amount of memories late in life requires more synapses and implies a smaller average SNR compared to early life, even under optimal learning conditions (Fig. 3.7, right). By extension, our model predicts that an animal reared under stimulus deprivation, during which it presumably forms fewer memories, should exhibit a lower connection density late in life, compared to an animal reared under control conditions or stimulus enrichment. This, indeed, agrees with experimental results from the early literature on structural plasticity (Globus et al., 1973; Turner & Greenough,

1985).

Importantly, our model stands in contrast to a previously proposed model of life-long learning (Chapeton et al., 2012), which is based on the assumption that memory robustness has a fixed upper limit, and that the consolidation process instead maximizes the amount of memories stored with this robustness. While this model also produces a decreasing connectivity trend across life, it predicts that the connection density should end up higher in animals that have formed less memories than controls (Fig. 3.7, left).

3.3.4 Biological interpretation of the synapse model

Our consolidation model crucially relies on the representation of each synapse as a product of multiple subcellular components linked in a signaling cascade. This suggests that the structural complexity of a synapse could serve a computational and metabolic purpose, by implicitly biasing cortical connectivity to be sparse, thus lowering energy consumption and freeing unneeded synaptic resources for future learning. These results are consistent with and complementary to previous theoretical work on modular synaptic ultrastructure (Lisman, 2017) and synaptic consolidation, which has demonstrated that synapses containing chemical cascades of fast and slow components can vastly improve memory lifetime (Benna & Fusi, 2016) as well as the energy efficiency of plasticity (Li & van Rossum, 2020).

The result of our meta-analysis, indicating that synaptic strength fluctuations scale roughly as $\mathcal{O}(\sqrt{w})$ over short timescales, not only supports our synapse model with two or three internal expression sites, but it also confirms previous predictions from computational studies of biophysically detailed, compartmental synapses (Shouval, 2005; Triesch et al., 2018). It is important to note that our scaling analysis is concerned only with structural, long-term synaptic changes, which should not be confused with the inter-spike variability caused by short-term plasticity, as this also displays $\mathcal{O}(\sqrt{w})$ scaling (Loebel et al., 2013).

The particular dynamics of our bipartite synapse model, as implemented in wakefulness and sleep, is consistent with the tagging-and-capture hypothesis (Redondo & Morris, 2011), since the plastic u-factor can be interpreted as a tag, while the second u-factor represents a much slower plasticity process. During wakefulness, only the tag is allowed to change, and the slow factor remains fixed. This enables the network to quickly memorize patterns without extensive rewiring. During sleep, when consolidation is assumed to take place, both factors change, including the slow one. This allows the network to converge to an optimally pruned weight configuration. This suggests that tagging-and-capture in a multiplicative synapse can have an additional function, by shifting a network from quick and shallow learning in wakefulness, to slow but optimal consolidation in sleep.

While we assume in the main text that each u-factor represents a chemical component inside the post-synaptic neuron or dendritic spine, it is also possible to interpret each factor as part of an entire pre- and post-synaptic structure. In this case, our finding that $z = 2$ or 3 factors per connection is optimal fits particularly well with the binomial synapse model, which is commonly used in the experimental literature and estimates the average connection strength as a product of three factors, namely the release probability, the number of synapses per

connection, and the quantal size.

3.3.5 Predictions and future work

Our model makes two general predictions about the dynamics of consolidation on the neural, and behavioral level. First, on the behavioral level, we predict that the memory items that are weakly encoded prior to sleep should display a larger imrovement in the SNR after sleep, which should translate to a higher success rate in recall tests on the population level. While we partly confirm this with three large, published datasets, these cover only a part of the range of initial encodings. Furthermore, we predict that the decreasing trend observed in these datasets should be shifted down if subjects are required to memorize more information, and vice versa.

On the neural level, our consolidation model predicts that the balance between LTD and LTP shifts across wakefulness and sleep. In wakefulness, we suggest that the brain primarily performs few-shot learning of low-activity patterns. This is known to require LTP to be stronger than LTD in order to compensate for the higher prevalence of LTD events (Tsodyks & Feigel'man, 1988), which is in agreement with what tends to be reported in experimental data (O'Connor et al., 2005). In sleep, on the other hand, our model predicts that only the weakest pattern in each replay cycle should produce synaptic updates. Consequently, LTP and LTD should occur equally often, and the amplitudes of LTP and LTD should therefore be comparable. Although this prediction has not yet been tested directly, it fits with the experimental finding that the concentration of acetylcholine, a modulator of synaptic plasticity, is significantly lowered in SWS-mediated consolidation of declarative memory (Gais & Born, 2004).

3.4 Methods

3.4.1 Network model

We model a local cortical circuit of pyramidal cells as a recurrent network of N binary neurons. At time t, the output state $s_i(t)$ of each neuron $i = 1, \ldots, N$ is given by

$$s_i(t) = \Theta\left[I_{\text{tot},i}(t)\right] \tag{3.9}$$

where Θ is the Heaviside function and $I_{\text{tot},i}$ is the total input current, which is calculated as

$$I_{\text{tot},i}(t) = I_{\text{exc},i}(t) + I_{\text{stim},i}(t) - I^{(\text{slow})}_{\text{inh},i}(t) - I^{(\text{fast})}_{\text{inh}}(t) \tag{3.10}$$

where all terms, except for $I^{(\text{fast})}_{\text{inh}}$, are non-negative. The first term is the excitatory input, which is determined by the recurrent connectivity and the previous state of the network according to

$$I_{\text{exc},i}(t) = \sum_{j=1}^{N} w_{ij} s_j(t-1) \tag{3.11}$$

where $w_{ij} \geq 0$ denotes the connection strength from neuron j to i. Self-connections are not allowed, meaning $w_{ii} = 0$.

The second current term, $I^{(\text{slow})}_{\text{inh},i}$, is a constant (tonic) inhibitory current, which, effectively, acts as a threshold. This is neuron-specific and changes slowly, on a time-scale comparable to that of the excitatory weights (see plasticity rules below). In contrast, the additional inhibitory term $I^{(\text{fast})}_{\text{inh}}$ is global and fast-changing. In each time step, it stabilizes network activity by allowing only the fN neurons with largest input currents to be active.

In our mathematical analysis below, we use I_{inh} as a shorthand for $I^{(\text{slow})}_{\text{inh}}$, unless stated otherwise.

3.4.2 Memory patterns

Each memory pattern consists of a random binary vector ξ^μ_i, where $i = 1, \ldots, N$ indexes the neuron, and $\mu = 1, \ldots, M$ the memory item. For simulations in section 3.2.3, each element ξ^μ_i is independently assigned one with probability $0 < f < 0.5$ and zero with probability $1 - f$. The parameter f is the average fraction of active neurons in each pattern, and is therefore referred to as the pattern activity level.

For wake-sleep simulations, each pattern contains *exactly* fN ones and $(1-f)N$ zeros, although the location of ones and zeros is randomized.

For both types of simulations, the mean and variance of the activity across patterns, for a neuron i, is given by

$$\mathbb{E}_\mu[\xi^\mu_i] = fM\,, \quad \mathbb{V}_\mu[\xi^\mu_i] = f(1-f)M\,. \tag{3.12}$$

However, for the activity across neurons, within a pattern μ, we have

$$\mathbb{E}_i[\xi_i^\mu] = fN\,, \quad \mathbb{V}_i[\xi_i^\mu] = \begin{cases} f(1-f)N\,, & \text{for efficiency simulations} \\ 0\,, & \text{for sleep simulations} \end{cases} \tag{3.13}$$

This slight difference in pattern types makes it possible to perform recall simulations with $I_{\text{inh}}^{(\text{fast})}$, which allows only fN neurons to be active at a time.

3.4.3 SNR and error margin

When the network has static inhibition and is in a state of idle background activity, we assume that every neuron randomly activates at each time t with a probability $\tilde{f}$, where typically $0 \leq \tilde{f} \leq f$. Under these conditions, we use the central limit theorem (since $N \gg 1$) to estimate the excitatory input current to a neuron as normally distributed with mean

$$\mathbb{E}_t[I_{\text{exc},i}(t)] = \tilde{f} \sum_j^N w_{ij} \tag{3.14}$$

and variance

$$\mathbb{V}_t[I_{\text{exc},i}(t)] = \tilde{f}(1-\tilde{f}) \sum_j^N w_{ij}^2\,. \tag{3.15}$$

A detailed derivation can be found in the Appendix A3.3. At the moment a pattern is recalled and the network enters a stable attractor, each neuron is either silenced or activated. The response is determined by a current deflection from the threshold, given by

$$\Delta I_i^\mu = \left| \sum_j^N w_{ij}\xi_j^\mu - I_{\text{inh},i} \right| \tag{3.16}$$

which generally exceeds the level of background noise. The level of noise robustness with which neuron i contributes to the recall process can be quantified with the SNR, where ΔI_i^μ is the signal (see Fig. 3.1c). Each pattern is, in this way, characterized by an independent SNR with respect to each neuron i. To simplify the evaluation, we estimate robustness across *all patterns* by computing the smallest SNR that neuron i has during recall. The signal is now comprised of the smallest current deflection, meaning

$$\text{Signal}_i = \Delta I_{\text{min},i} = \min_\mu \Delta I_i^\mu\,, \tag{3.17}$$

while the noise is the largest fluctuation in the input current. Given that background activity during recall typically is lower than idle activity, a worst-case scenario is given by Eq. 3.15 with $\tilde{f} = f$, so that

$$\text{Noise}_i = \sqrt{f(1-f)\sum_j^N w_{ij}^2}\,, \tag{3.18}$$

which finally yields

$$\mathrm{SNR}_i = \frac{\Delta I_{\min,i}}{\sqrt{f(1-f)\sum_j^N w_{ij}^2}} \,. \tag{3.19}$$

In our definition of optimal robustness under sparsifying constraints, we generalize the notion of SNR and use the error margin

$$K_q = \frac{\Delta I_{\min}}{(\sum_j^N w_{ij}^q)^{\frac{1}{q}}} \tag{3.20}$$

where subscript i has been omitted (see Appendix A3.1 for details). Here, one can directly see that a maximization of K_2 is equivalent to a maximization of SNR.

3.4.4 Theoretical solutions

Details regarding the calculations of theoretical solutions can be found in Appendices A3.4, A3.5, and A3.6.

3.4.5 Synapse model

For a fixed q, we can maximize K_q by parameterizing each weight w_{ij} as

$$w_{ij} = \prod_k^z u_{ijk} \tag{3.21}$$

where $z = 2/q$ (Hoff, 2017) and instead maximize

$$K^{(u)} = \frac{\Delta I_{\min}}{(\sum_j^N \sum_k^z u_{ijk}^2)^{1/2}} \,. \tag{3.22}$$

Since it can be shown that the solution is characterized by $u_{ij1}^* = u_{ij2}^* = \ldots = u_{ijz}^*$, we simplify the optimization by assuming this structure *a priori*. In practice, we therefore use the parameterization

$$w_{ij} = u_{ij}^z \tag{3.23}$$

and we maximize

$$K^{(u)} = \frac{\Delta I_{\min}}{(\sum_j^N u_{ij}^2)^{1/2}} \,. \tag{3.24}$$

Note that this optimization function is independent of q.

3.4.6 General learning rule

We maximize $K^{(u)}$ using projected gradient ascent. The result is outlined in Algorithm 1, which, in the limit of small learning rates $\eta \to 0$, is equivalent to the procedure described in the main text. In the case $z = 1$ and without sign-constraints, it is reduced to the iterative method introduced by Krauth & Mezard (1987).

Algorithm 1 Sparse Optimal Perceptron

for $t = 0, 1, 2, \ldots$ **do**

$\mu_i^\star \leftarrow \arg\min_\mu \, (2\xi_i^\mu - 1)\left(\sum_j w_{ij}\xi_j^\mu - I_{\text{inh},i}\right)$ ▷ tagging weakest pattern

$\hat{u}_{ij} \leftarrow \left[u_{ij} + \eta(2\xi_i^{\mu_i^\star} - 1)\xi_j^{\mu_i^\star} u_{ij}^{z-1}\right]_+$ ▷ sign-constrained Hebbian update

$u_{ij} \leftarrow \hat{u}_{ij}/(\sum_j \hat{u}_{ij}^2)^{1/2}$ ▷ weight normalization

$w_{ij} \leftarrow u_{ij}^z$

$I_{\text{inh},i} \leftarrow I_{\text{inh},i} - \eta_{\text{inh}}(2\xi_i^{\mu_i^\star} - 1)$ ▷ inhibition update

end for

In the limit of small learning rates, we can also describe the dynamics of u_{ij} in continuous time with the gradient flow equation

$$\frac{\mathrm{d}u_{ij}}{\mathrm{d}t} \propto \left(1 - \frac{\sum_j u_{ij}^2}{\text{const.}}\right) u_{ij} + \eta(2\xi_i^{\mu_i^\star} - 1)\xi_j^{\mu_i^\star} u_{ij}^{z-1} \,. \tag{3.25}$$

With a change of variables back to w_{ij}, we recover Eq. 3.5. For more details on this derivation and its relation to previously published homeostatic learning rules, see Appendix A3.2.

3.4.7 Numerical optimization and evaluation

The results in Figure 3.2 were produced by training networks using Algorithm 1 with a fixed learning rate. During training, the performance of the network was evaluated with the average SNR, the error, and the weight density. The average SNR was computed as

$$\langle \text{SNR} \rangle = \frac{1}{N} \sum_i \text{SNR}_i \tag{3.26}$$

while the error was defined as the average fraction of incorrect bits after one state update, meaning

$$E = \frac{1}{2NM} \sum_i^N \sum_\mu^M 1 - (2s_i^\mu - 1)(2\xi_i^\mu - 1) \tag{3.27}$$

where

$$s_i^\mu = \Theta\left[\sum_j^N w_{ij}\xi_j^\mu - I_{\text{inh},i}\right] \,. \tag{3.28}$$

The weight density was computed as

$$\rho = \frac{1}{N^2} \sum_{i,j}^N \Theta[w_{ij} - w_0] \tag{3.29}$$

where w_0 is a threshold used to determine if a weight has been removed. Due to the finite size of the weight updates and the limits of machine precision, some weights converged to values close to, but not exactly, zero, such as $\sim 10^{-18}$. The threshold was therefore set to

$w_0 = 10^{-10}$.

The optimization was considered to have converged once three conditions were satisfied: (i) $E = 0$, (ii) $\langle \text{SNR} \rangle$ changed by less than 10^{-4} over 10^4 epochs, and (iii) ρ changed by less than $2 \cdot 10^{-4}$ over 10^4 epochs.

After optimization, noise tolerance was evaluated by initiating the network in a distorted version of each pattern, updating the network 50 times, and evaluating if the network's final state was close to the original pattern. The criterion for closeness was that the error of the final state was $E < 0.1f$. This test was performed 20 times per pattern, and the recall ratio (RR) of the network was computed as the average fraction of patterns that could be retrieved across all trials. We defined the maximal noise tolerance as the noise level at which RR fell below 50%.

For a given noise level $0 \leq p \leq 2f$, we generated distorted patterns by flipping each bit with probability

$$p_{\text{flip}} = \begin{cases} p/2(1-f) & \text{for } 0 \to 1 \\ p/2f & \text{for } 1 \to 0 \end{cases} . \tag{3.30}$$

This ensured that the average activity level was kept at f for all distorted patterns (see Suppl. Fig. A3.5). After updating, the network was deemed close to the original pattern if the error of the final state was $E < 0.1f$.

The fraction of retrievable patterns for each noise-level was computed as

$$\hat{\alpha}(\alpha, p) = \alpha \cdot \text{RR}(\alpha, p) \tag{3.31}$$

where α is the storage load

$$\alpha = M/N \, . \tag{3.32}$$

The largest possible (critical) storage load was denoted α_c. We computed the efficiency Q as the number of bits per synapse that could be retrieved at a given storage load and noise-level, meaning

$$Q(\alpha, p) = -\frac{\hat{\alpha}}{\rho} \left[f \log_2(f) + (1-f) \log_2(1-f) \right] \tag{3.33}$$

while the grand average was computed over all storage loads and noise-levels

$$\text{Efficiency average} = \frac{1}{\alpha_{\max} p_{\max}} \int_{\alpha_{\min}}^{\alpha_{\max}} \int_{0}^{p_{\max}} Q(\alpha, p) \, \mathrm{d}\alpha \, \mathrm{d}p \tag{3.34}$$

where the integrals were computed numerically using the trapezoidal rule.

3.4.8 Simulating wakefulness and sleep

Our simulation of network dynamics in wakefulness has some similarity to previous models of familiarity detection and learning (Sohal & Hasselmo, 2000; Alemi et al., 2015). However,

this aspect of our work is not our main contribution, but rather a simple prototype used to demonstrate that our general learning rule is consistent with neurophysiological conditions seen in both wakefulness and sleep.

In wakefulness, the network was simulated with fast and slow inhibition. The network was clamped to a pattern μ by providing a strong stimulus current

$$I_{\text{stim},i} = w_{\text{stim}} \xi_i^\mu \tag{3.35}$$

and performing a single update. Here, w_{stim} is a positive, global weight that determines the strength of the bottom-up input (see Fig. 3.1a). Next, the stimulus current was released and a second update was performed. The novelty indicator of the pattern was determined according to

$$\xi^\star = \begin{cases} 1 & \text{if } I_{\text{inh}}^{(\text{fast})} \neq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{3.36}$$

Finally, the network was updated a third time with I_{stim} active so that $s_i = \xi_i^\mu$, and the weights we updated according to

$$\begin{aligned} \Delta u_{ij1} &= \eta_{\text{het}} \left(1 - \frac{\sum_j w_{ij}}{\text{const.}}\right) u_{ij1} + \eta_{\text{hom}} \xi^\star (s_i - f) s_j u_{ij2} \\ \Delta u_{ij2} &= 0 \\ w_{ij} &= u_{ij1} \cdot u_{ij2} \end{aligned} \tag{3.37}$$

while inhibition was updated according to

$$\Delta I_{\text{inh},i} = \eta_{\text{inh}} \operatorname{sgn}\left[I_{\text{inh},i}^\star - I_{\text{inh},i}\right] \tag{3.38}$$

where η_{het}, η_{hom}, and η_{inh} are the heterosynaptic, homosynaptic, and inhibitory learning rates. The optimal inhibition at each time is denoted $I_{\text{inh},i}^\star$ and is calculated as

$$I_{\text{inh},i}^\star = \mathbb{E}_{\text{exc}} + \sqrt{2fN\mathbb{V}_{\text{exc}}}\, \text{erfc}^{-1}(2f) \tag{3.39}$$

where $\mathbb{E}_{\text{exc}}$ and $\mathbb{V}_{\text{exc}}$ is the expectation and variance given in Eqs. 3.14 and 3.15. In this weight parameterization, we kept the two u-factors separate and allowed only one factor to be plastic while the other remained fixed. Note, however, that these simulations can be performed with both factors plastic and equal, in which case $w_{ij} = u^2$.

After wakeful learning, the two u-factors were equilibrated by setting $u_{ij1} = u_{ij2} = u_{ij} = \sqrt{w_{ij}}$. In sleep, the network was initialized in each pattern μ by an external cue and updated once. Since each pattern had been encoded as an attractor, the network remained in the state $s_i = \xi_i^\mu$. The input current deflection was now read out in each neuron i according to

$$\Delta I_i = \left| \sum_j w_{ij} s_j - I_{\text{inh},i} \right| \tag{3.40}$$

and the weakest pattern was tagged according to

$$\xi^\star = \arg\min_{\xi} \ \Delta I_i \ . \tag{3.41}$$

At the end of each pattern replay cycle, the weights were updated according to

$$\Delta u_{ij} = (2\xi_i^\star - 1)(\eta_{\text{het}} + \eta_{\text{hom}}\xi_j^\star)u_{ij} \ . \tag{3.42}$$

Note that the combination of Eqs. 3.40-3.42 is equivalent to Algorithm 1, but implemented in a self-supervised fashion (see Appendix A3.7).

The ratio of LTP-to-LTD was calculated as the number of homosynaptic LTP events relative LTD events, weighted by the fraction of homosynaptic learning rate in LTP relative LTD, so that

$$\frac{\text{LTP}}{\text{LTD}} = \frac{n_{\text{hom}}^{(\text{LTP})}}{n_{\text{hom}}^{(\text{LTD})}} \frac{\eta_{\text{hom}}^{(\text{LTP})}}{\eta_{\text{hom}}^{(\text{LTD})}} \ . \tag{3.43}$$

In wakeful learning, this equals precisely 1.

3.4.9 Simulating synaptic volatility

In simulations of internal synaptic noise, we assumed, as in wakefulness, that one u-factor is plastic and changes with a fast time constant τ_{fast}, while remaining factors are slower and characterized by the time constant τ_{slow}. Each weight was therefore parameterized as

$$w = u_{\text{fast}} \cdot u_{\text{slow}}^{z-1} \tag{3.44}$$

where the fast factor was governed by the stochastic process

$$\tau_{\text{fast}} \frac{\mathrm{d}u_{\text{fast}}}{\mathrm{d}t} = \eta_{\text{het}} \left(1 - \frac{\sum w^{\frac{2}{z}}}{\text{const.}}\right) u_{\text{fast}} \ + \ \eta_\epsilon \left(k\epsilon_1 u_{\text{slow}}^{z-1} + (1-k)\epsilon_0\right) \tag{3.45}$$

and the slow factor by

$$\tau_{\text{slow}} \frac{\mathrm{d}u_{\text{slow}}}{\mathrm{d}t} = \eta_{\text{het}} \left(1 - \frac{\sum w^{\frac{2}{z}}}{\text{const.}}\right) u_{\text{slow}} \ + \ (u_{\text{fast}} - u_{\text{slow}}) \ . \tag{3.46}$$

As before, η_{het} denotes the heterosynaptic learning rate, while η_ϵ scales the amplitude of the noise fluctuations injected by the two biased gaussian noise terms ϵ_1 and ϵ_0, where the first term models activity-dependent noise and the second term internal noise. The relative strength of the two noise sources is set with $0 \leq k \leq 1$. The results in Fig. 3.6b were produced with the values

$$\tau_{\text{fast}} = 10\,\text{min} \ , \quad \tau_{\text{slow}} = 5\,\text{h} \tag{3.47}$$

and

$$k = \begin{cases} 0 & \text{for silent model} \\ 0.5 & \text{for control model} \end{cases} \ . \tag{3.48}$$

However, the outcome is qualitatively the same for time constants much closer to each other, or even with $\tau_{\text{slow}} < \tau_{\text{fast}}$. In the latter case, all u-factors are approximately identical, and the parameterization reduces to $w = u^z$.

The sampling time was set to $T_{\text{sample}} = \tau_{\text{fast}}$ and the total length of the simulation was $T_{\text{sim}} = \tau_{\text{fast}} \cdot 10^3$, which roughly equals 7 d.

3.4.10 Experimental data: Connectivity

The experimental data on connection probability among cortical excitatory cells was taken from a publicly available compilation of 124 datasets that were included in a meta-analysis published in Zhang et al. (2019). Each study in the dataset was assigned a weight β_i according to the number of evaluated connections n_{conn}, so that

$$\beta_i = \frac{n_{\text{conn}}^{(i)}}{\sum_i^{n_{\text{sets}}} n_{\text{conn}}^{(i)}} \,. \tag{3.49}$$

The weighted mean (wM) and standard error of the mean (wSEM) of the connection probability P_{conn} was then estimated using

$$\text{wM} = \frac{1}{n_{\text{sets}}} \sum_i^{n_{\text{sets}}} \beta_i P_{\text{conn}}^{(i)} \tag{3.50}$$

$$\text{wSEM} = \frac{1}{n_{\text{sets}} - 1} \sqrt{\sum_i^{n_{\text{sets}}} \beta_i \left(P_{\text{conn}}^{(i)} - \text{wM} \right)^2} \,. \tag{3.51}$$

3.4.11 Experimental data: Synaptic volatility

We compiled 23 datasets containing synaptic measurements from 9 previously published studies. In general, each datapoint consisted of a measured proxy of synaptic strength ($\hat{w}$) together with a change in strength ($\Delta\hat{w}$) following a time interval Δt. Each dataset was divided into LTD ($\Delta\hat{w} < 0$) and LTP ($\Delta\hat{w} > 0$) events. The average change $\langle\Delta\hat{w}\rangle$ was estimated as a function of initial strength by filtering the all datapoints in $(\Delta\hat{w}, \hat{w})$-space using a moving average with window size $n_{\text{syn}}/20$, where n_{syn} is the sample size.

The scaling exponent was estimated by fitting a line to the estimated mean change $\langle\Delta\hat{w}\rangle$ in logarithmic space. The mean and standard error of the exponent was estimated by repeating the averaging and line-fitting with bootstrapping. All datasets were boostrapped 1,000 times, except the sets from Kaufman et al. (2012), Fisher-Lavie & Ziv (2013), and Hazan & Ziv (2020), which were boostrapped 100 times due to their exceptionally large sample size.

To summarize all exponent estimates, each one was first assigned a weight according its inverse variance (squared standard error), meaning

$$\beta_i = \frac{\text{SEM}^{-2}}{\sum_i^{n_{\text{sets}}} \text{SEM}^{-2}} \,. \tag{3.52}$$

The weighted mean was then calculated with Eq. 3.50 and the weighted standard error with

$$\mathrm{wSEM} = \sqrt{\frac{1}{\sum_i^{n_{\mathrm{sets}}} \mathrm{SEM}^{-2}}} \,. \tag{3.53}$$

The 99% confidence interval was finally estimated as $[\mathrm{wM} \pm 2.58 \times \mathrm{wSEM}]$.

3.4.12 Experimental data: CV of synapse norm

In order to perform this analysis, we utilized the dendritic spine measurements acquired by Hazan & Ziv (2020) with blocked glutamatergic transmission and under normal conditions. Each measurement site in the original data was assumed to represent a separate neuron (N. Ziv, personal communication). For each neuron i, and at each time point t, we calculated the q-norm of the reported spine intensities, that is

$$\|\hat{w}\|_q^{(i,t)} = \left(\sum \hat{w}^q\right)^{\frac{1}{q}} \tag{3.54}$$

where the notation on the right-hand side has been simplified for readability. A single CV-value was obtained for each neuron and each norm according to

$$\mathrm{CV}_q^{(i)} = \frac{\sqrt{\hat{\mathbb{V}}_t\left[\|\hat{w}\|_q^{(i,t)}\right]}}{\hat{\mathbb{E}}_t\left[\|\hat{w}\|_q^{(i,t)}\right]} \,. \tag{3.55}$$

The mean and standard error of the CV was finally computed as

$$\text{Mean } \mathrm{CV}_q = \hat{\mathbb{E}}_i[\mathrm{CV}_q^{(i)}] \,, \quad \text{SEM } \mathrm{CV}_q = \sqrt{\hat{\mathbb{V}}_i[\mathrm{CV}_q^{(i)}]} \,. \tag{3.56}$$

The simulated data was analyzed precisely as the experimental data, with the only caveat being that all simulated weights were treated as belonging to a single neuron. Bootstrapping was applied to both experimental and simulated data by separately re-sampling the measurements made in each neuron at each time point 1,000 times.

3.4.13 Experimental data: Synaptic pruning

To analyze the properties of synaptic pruning, we utilized the dataset on dendritic spines by Loewenstein et al. (2011, 2015). This data contains spine volume measurements over six sessions, with a sampling interval of $\Delta t = 4\,\mathrm{d}$ (see Table A3.5 for details). Spines where separated into three categories: (i) Spines that were observed for the first time somewhere between sessions 2 to 6 were defined as "young" spines. The age of the spines observed in session 1 can not be determined, and these were therefore left out. (ii) Spines that disappeared at any time between sessions 2 to 6 were defined as "pruned". (iii) Spines that had been observed in at least one directly preceding session were defined as "old".

To estimate the pruning fraction, we first log-normalized the data by calculating the Z-score

in logarithmic space, according to

$$Z(\log x) = \frac{\log x - \mathbb{E}[\log x]}{\sqrt{\mathbb{V}[\log x]}} \ . \tag{3.57}$$

We then binned all spine volumes from sessions 1 to 5, and computed the ratio of pruned spines relative the total number of spines in each bin. The bins were sized so that there was approximately an equal number of spines in each one. The spines in session 6 were left out, as it is unknown how many of these that were pruned.

The simulated pruning fraction was calculated analogously, by comparing the pruned connections to all connection weights before sleep.

3.4.14 Experimental data: Memory consolidation in sleep

In the behavioral data on memory tests with word-pair associations, we modeled recall performance according to signal detection theory. We assumed that the trace of each memory was encoded in a subject according to a subject-specific strength, combined with normally distributed noise, so that all traces in a subject were approximately normally distributed after the initial training session. Furthermore, the fraction of memories that could be recalled correctly at test time were assumed to be those whose trace exceeded a subject-specific threshold. We estimated the average memory SNR within a subject as the distance from the average trace strength to the threshold. This is given by

$$\mathrm{SNR}^* = \Phi^{-1}\left[P_{\mathrm{recall}} + \epsilon\right] \tag{3.58}$$

where Φ is the normal cumulative distribution function and ϵ is a small corrective term added to avoid divergence; it is calculated as

$$\epsilon = (1 - 2P_{\mathrm{recall}}) \cdot 10^{-16} \ . \tag{3.59}$$

We calculated the SNR^* preceding the wake/sleep interval (denoted $\mathrm{SNR}^*_{\mathrm{before}}$), and the difference

$$\Delta\mathrm{SNR}^* = \mathrm{SNR}^*_{\mathrm{after}} - \mathrm{SNR}^*_{\mathrm{before}} \tag{3.60}$$

following the interval. Datapoints further than three standard deviations from the mean were considered outliers and removed. This was done separately in each dataset.

Comparisons between wake and sleep were done by fitting all data with the linear model

$$\Delta\mathrm{SNR}^* = \beta_0 + \beta_1 X_{\mathrm{cond}} + \beta_2 \mathrm{SNR}^*_{\mathrm{before}} + \beta_3 X_{\mathrm{cond}} \mathrm{SNR}^*_{\mathrm{before}} \tag{3.61}$$

where the group condition was coded by X_{cond} according to

$$X_{\mathrm{cond}} = \begin{cases} 0 & \text{if wake} \\ 1 & \text{if sleep} \end{cases} \ . \tag{3.62}$$

Significant differences in the intercept and slope between wake and sleep was determined by a two-tailed t-test of β_1 and β_3.

Appendix

A3.1 Derivation of consolidation algorithm

We consider a single binary neuron i with inputs from N other neurons j, and we define the general problem of optimal consolidation as finding the weights w_{ij} that maximize the error margin K_q. To express this in mathematical terms, we will represent all incoming weights to neuron i with the column vector $\boldsymbol{w}$ (subscript i is omitted for clarity). Likewise, the state of the network will be represented by the vector $\boldsymbol{s}(t)$ and each pattern by $\boldsymbol{\xi}^\mu$ for $\mu = 1, \ldots, M$. We can now define our optimization problem as

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \; K_q \tag{A3.1}$$

where the error margin can be expressed as

$$K_q = \frac{\Delta I_{\text{min}}}{\|\boldsymbol{w}\|_q} \tag{A3.2}$$

and the minimal deflection of the input current can be written

$$\begin{aligned} \Delta I_{\text{min}} &= \min_{\mu} \; \left(2\xi_i^\mu - 1\right)\left(\boldsymbol{w}^\top \boldsymbol{\xi}^\mu - I_{\text{inh}}\right) \\ &= \left(2\xi_i^\star - 1\right)\left(\boldsymbol{w}^\top \boldsymbol{\xi}^\star - I_{\text{inh}}\right) \; . \end{aligned} \tag{A3.3}$$

Assuming that all patterns already are encoded as attractors in the network, the last equation is equivalent to

$$\Delta I_{\text{min}} = \left|\boldsymbol{w}^\top \boldsymbol{\xi}^\star - I_{\text{inh}}\right| \; . \tag{A3.4}$$

The problem in Eq. A3.1 can be solved in three different ways:

Method (i): The first approach is to fix the norm $\|\boldsymbol{w}\|_q$, define a minimal deflection amplitude κ that all patterns must satisfy, and thereafter train the weights to maximize the number of stored patterns M. This variant of the optimization can be written

$$\begin{aligned} \boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \; M \quad \text{s.t.} \quad & \Delta I_{\text{min}} \geq \kappa \\ & \|\boldsymbol{w}\|_q \text{ and } \kappa \text{ const.} \end{aligned} \tag{A3.5}$$

This is how optimal memory storage originally was defined in the statistical physics literature (Gardner, 1988) (see section A3.4 and Suppl. Fig. A3.6), and it forms the basis of recent work on optimal long-term memory models (Chapeton et al., 2012; Brunel, 2016), where the inequality constraint is satisfied by directly including κ in the learning rule (Alemi et al., 2015) or by training on patterns with a fixed amount of noise (Rubin et al., 2017). We argue, however, that the assumption that cortical circuits have a fixed robustness and only learn to maximize the number of memories is problematic from a ethological perspective. It implies that cortical circuitry does not adapt to environmental cognitive pressures, but instead passively incorporates information when it is encountered without allowing for further improvement in the encoding.

Method (ii): The second way to formulate the optimization is to fix the minimal deflection amplitude κ and the number of patterns M, and instead minimize the norm $\|\boldsymbol{w}\|_q$. This can be written as

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_q \quad \text{s.t.} \quad \begin{aligned} &\Delta I_{\text{min}} \geq \kappa \,, \\ &M \text{ and } \kappa \text{ const.} \end{aligned} \tag{A3.6}$$

This formulation is also known as the max-margin classifier (or linear support vector machine) in the machine learning literature (Cortes & Vapnik, 1995). This is more suitable as a model of cognition and learning than Eq. A3.5, as it produces optimal storage by maximizing the robustness for any fixed number of patterns $M \leq M_c$, instead of the other way around. However, this model still poses a problem in terms of neurobiological realism, in that the weight norm needs to be adapted to each specific set of patterns. This is incompatible with the notion of homeostatic synaptic plasticity, which regulates neural input by preserving the overall strength of synaptic connections over time (Turrigiano, 2008).

Method (iii): The disadvantages of both previous models can be avoided by solving the optimization problem as follows: the weights are trained to maximize κ while the norm and the number of patterns are kept fixed. We express this as

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \kappa \quad \text{s.t.} \quad \begin{aligned} &\Delta I_{\text{min}} \geq \kappa \,, \\ &M \text{ and } \|\boldsymbol{w}\|_q \text{ const.} \end{aligned} \tag{A3.7}$$

Although this formulation is uncommon in the literature, it was initially treated by Krauth & Mezard (1987). The significance of this approach is that the network can be subjected to any homeostatic constraint on the weight norm, while being trained to maximize the robustness of any number of patterns $M \leq M_c$.

We apply the variable change

$$\boldsymbol{w} = \boldsymbol{u}^z \,, \quad z = 2/q \tag{A3.8}$$

where the exponent is applied element-wise, and obtain

$$\boldsymbol{u}^* = \arg\max_{\boldsymbol{u}} \kappa \quad \text{s.t.} \quad \begin{aligned} &\Delta I_{\text{min}} \geq \kappa \,, \\ &M \text{ and } \|\boldsymbol{u}\|_2 \text{ const.} \end{aligned} \tag{A3.9}$$

This problem can now be directly solved with projected gradient ascent (see Suppl. Fig. A3.7), which results in the iterative optimization described in Algorithm 1.

A3.2 Derivation of homeostatic scaling laws

An alternative method for solving Eq. A3.9 is to formulate the problem as a loss function with a penalty for the norm, as in

$$\mathcal{L} = H(\|\boldsymbol{u}\|_2^2) - \Delta I_{\text{min}} \tag{A3.10}$$

where H is a penalty function. Applying gradient descent to $\mathcal{L}$ results in a general learning rule that, in continuous time, is given by

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} \propto h(\|\boldsymbol{u}\|_2^2)\boldsymbol{u} + \eta(2\xi_i^\star - 1)\boldsymbol{\xi}^\star \odot \boldsymbol{u}^{z-1} \tag{A3.11}$$

where $h = H'$ and $\odot$ denotes element-wise multiplication. In the specific case $z = 2$, a variable change back to $\boldsymbol{w}$ results in

$$\frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} \propto h(\|\boldsymbol{w}\|_1)\boldsymbol{w} + \eta(2\xi_i^\star - 1)\boldsymbol{\xi}^\star \odot \boldsymbol{w}\ . \tag{A3.12}$$

Under the assumption that global activity in the network is stable over time, the norm $\|\boldsymbol{w}\|_1$ is directly proportional to the average excitatory input current, as shown in Eq. 3.14. With a slight abuse of notation, we can therefore rewrite the above equation as

$$\frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} \propto h(\langle I_\mathrm{exc}\rangle_t)\boldsymbol{w} + \eta(2\xi_i^\star - 1)\boldsymbol{\xi}^\star \odot \boldsymbol{w}\ . \tag{A3.13}$$

Case (i): If we choose the penalty function

$$H(\|\boldsymbol{u}\|_2^2) = \left(\text{const.} - \|\boldsymbol{u}\|_2^2\right)^2 \tag{A3.14}$$

we obtain the homeostatic factor

$$h(\langle I_\mathrm{exc}\rangle_t) = (\text{const.} - \langle I_\mathrm{exc}\rangle_t) \tag{A3.15}$$

which is equivalent to the homeostatic scaling rule introduced by Renart et al. (2003), albeit rewritten in terms of the excitatory input current instead of the input firing rate.

Case (ii): If we instead define the penalty as

$$H(\|\boldsymbol{u}\|_2^2) = \left(1 - \frac{\|\boldsymbol{u}\|_2^2}{\text{const.}}\right)^2 \tag{A3.16}$$

we retrieve the homeostatic factor

$$h(\langle I_\mathrm{exc}\rangle_t) = \left(1 - \frac{\langle I_\mathrm{exc}\rangle_t}{\text{const.}}\right) \tag{A3.17}$$

which is equivalent to the homeostatic rule introduced by Toyoizumi et al. (2014).

Case (iii): A third alternative for the penalty function is

$$H(x) = x\log(x) - x\ , \quad x = \frac{\|\boldsymbol{u}\|_2^2}{\text{const.}} \tag{A3.18}$$

which yields the homeostatic factor

$$h(\langle I_\mathrm{exc}\rangle_t) = \log\left(\frac{\text{const.}}{\langle I_\mathrm{exc}\rangle_t}\right)\ . \tag{A3.19}$$

This type of homeostatic scaling has, to the best of our knowledge, not been proposed previously in the literature.

It is important to note that even though all homeostatic rules regulate the average input current, they do so by monitoring different quantities. In case (i), the rule depends on the raw current deviation from the set-point, while, in case (ii), it depends on the percentage of the deviation. In the final case, the homeostatic rule depends only on the ratio of $\langle I_{\text{exc}} \rangle_t$ relative the set-point.

A3.3 Derivation of input current statistics

We consider the setting where the network is in a state of idle background activity. At every time-step, neurons have a probability f of becoming active, which means that each neural state $s_i(t)$ is a Bernoulli random variable with

$$\mathbb{E}_t[s_i(t)] = f \ , \quad \mathbb{V}_t[s_i(t)] = f(1-f) \ . \tag{A3.20}$$

The mean and variance of the excitatory input current is now given by

$$\begin{aligned} \mathbb{E}_t[I_{\text{exc},i}(t)] &= \mathbb{E}_t\left[\sum_j^N w_{ij} s_j(t-1)\right] = \sum_j^N w_{ij}\, \mathbb{E}_t[s_j(t-1)] \\ &= f \sum_j^N w_{ij} = f\|\boldsymbol{w}\|_1 \end{aligned} \tag{A3.21}$$

and

$$\begin{aligned} \mathbb{V}_t[I_{\text{exc},i}(t)] &= \mathbb{V}_t\left[\sum_j^N w_{ij} s_j(t-1)\right] = \sum_j^N w_{ij}^2\, \mathbb{V}_t[s_j(t-1)] \\ &= f(1-f) \sum_j^N w_{ij}^2 = f(1-f)\|\boldsymbol{w}\|_2^2 \end{aligned} \tag{A3.22}$$

respectively. This result depends only on the neural activity level across time, meaning $\mathbb{E}_t[s_i(t)]$ and $\mathbb{V}_t[s_i(t)]$, and not on the activity level within the network, that is $\mathbb{E}_i[s_i(t)]$ and $\mathbb{V}_i[s_i(t)]$.

It is also evident, from this analysis, that a solution to the storage problem in the mean-field limit $N \to \infty$ with a scaling $w \sim \mathcal{O}(1/\sqrt{N})$ corresponds to an L_2-regularization, whereas the scaling $w \sim \mathcal{O}(1/N)$ corresponds to an L_1-regularization.

A3.4 Theoretical solution for maximal SNR

As shown in Eq. 3.19, maximizing SNR is achieved by maximizing K_q with $q = 2$. This, in turn, can be done by solving Eq. A3.5. At optimality, the relationship between maximal

SNR and storage load α has been derived by Gardner (1988) and is given by

$$\alpha(m,\kappa) = \frac{1}{2}\left[\tfrac{1}{2}(1+m)\int_{\frac{vm-\kappa}{\sqrt{1-m^2}}}^{\infty} D(x)\left(\frac{\kappa - vm}{\sqrt{1-m^2}} + x\right)^2 \mathrm{d}x + \tfrac{1}{2}(1-m)\int_{\frac{-vm-\kappa}{\sqrt{1-m^2}}}^{\infty} D(x)\left(\frac{\kappa + vm}{\sqrt{1-m^2}} + x\right)^2 \mathrm{d}x\right]^{-1} \tag{A3.23}$$

where v is given by the solution to the equation

$$\tfrac{1}{2}(1+m)\int_{\frac{vm-\kappa}{\sqrt{1-m^2}}}^{\infty} D(x)\left(\frac{\kappa - vm}{\sqrt{1-m^2}} + x\right)\mathrm{d}x = \tfrac{1}{2}(1-m)\int_{\frac{-vm-\kappa}{\sqrt{1-m^2}}}^{\infty} D(x)\left(\frac{\kappa + vm}{\sqrt{1-m^2}} + x\right)\mathrm{d}x \tag{A3.24}$$

and where D is the standard normal distribution

$$D(x) = \frac{\exp(-\frac{1}{2}x^2)}{\sqrt{2\pi}}\,. \tag{A3.25}$$

The pattern magnetization m is related to the activity level f according to

$$f = \frac{1+m}{2}\,. \tag{A3.26}$$

while κ is linked to the SNR according to

$$\mathrm{SNR} = \frac{\kappa}{2\sqrt{f(1-f)}} \tag{A3.27}$$

since the solution is derived for $\|\boldsymbol{w}\|_2^2 = 1$. The maximal capacity α_c is obtained by solving Eq. A3.23 with $\kappa = 0$. Note that Eqs. A3.23 and A3.27 have both been adjusted with a factor $\frac{1}{2}$ to account for the fact that we allow only non-negative weights and use patterns with values 0 or 1, while the original solution was derived for unconstrained weights and patterns with ± 1.

Balanced patterns: In the specific case of balanced patterns ($f = 0.5$), Eq. A3.23 reduces to

$$\alpha(\kappa) = \frac{1}{2}\left[\int_{-\kappa}^{\infty} D(x)(\kappa + x)^2\,\mathrm{d}x\right]^{-1}\,. \tag{A3.28}$$

For synchronous state updates, the largest tolerable noise level can be written in terms of the smallest acceptable overlap m_{min} of the distorted pattern with the original pattern,

according to

$$p_{\max}(\kappa) = \frac{1 - m_{\min}(\kappa)}{4} \,. \tag{A3.29}$$

The smallest acceptable overlap has been derived by Kepler & Abbott (1988) and is determined by the solution to the equation

$$\begin{aligned} m_{\min}(\kappa) = 2 \int_{\kappa}^{\infty} D(x) \operatorname{erf}\left(\frac{x \cdot m_{\min}}{\sqrt{2(1-m_{\min}^2)}}\right) \mathrm{d}x \\ + \left[1 + \operatorname{erf}\left(\frac{\kappa}{\sqrt{2}}\right)\right] \operatorname{erf}\left(\frac{\kappa \cdot m_{\min}}{\sqrt{2(1-m_{\min}^2)}}\right) - 1 \,. \end{aligned} \tag{A3.30}$$

A3.5 Theoretical solution for maximal pruning

In order to obtain the solution for the lowest possible weight density $\rho_{\min}$, we first define the more general problem of finding the set of weights that maximize SNR under a fixed weight density ρ. We do this by adding an additional constraint to Eq. A3.5 with $q = 2$, thus yielding

$$\begin{aligned} \boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \; M \quad \text{s. t.} \quad & \Delta I_{\min} \geq \kappa \\ & \|\boldsymbol{w}\|_0 = \rho N \\ & \|\boldsymbol{w}\|_2 \,,\; \kappa \,,\; \text{and } \rho \text{ const.} \end{aligned} \tag{A3.31}$$

In the case of balanced patterns, this problem has been solved by Bouten et al. (1990). Analogously to Eq. A3.5, the solution is now described by the storage load as a function of κ and ρ according to

$$\alpha(\kappa, \rho) = \frac{2\rho + \frac{2}{\sqrt{\pi}} \operatorname{erfc}^{-1}(2\rho) \cdot \exp\left[-\operatorname{erfc}^{-1}(2\rho)^2\right]}{2 \int_{-\kappa}^{\infty} D(x)(\kappa + x)^2 \, \mathrm{d}x} \tag{A3.32}$$

where both α and ρ have been adjusted with a factor $\frac{1}{2}$ and 2 relative the original solution to, once again, adjust for sign-constrained weights. Here, we rely on a simple symmetry argument: The original solution always contains an equal number of positive and negative weights. Intuitively, one can therefore expect that a sign-constraint would cause precisely half of the weights to have the wrong sign and to be pruned in the new solution. This has, indeed, been proven in the case of saturation ($\alpha = \alpha_c$) by Yau (1992) and we conjecture that the same result applies for all α.

The smallest possible weight density $\rho_{\min}$ for each storage load is obtained by computing Eq. A3.32 with $\kappa = 0$. The result can be inserted in Eq. 3.6 to obtain the maximal efficiency

$$Q_{\max} = -\frac{\alpha}{\rho_{\min}} [f \log_2(f) + (1-f) \log_2(1-f)] \,. \tag{A3.33}$$

Note that the densest solution is equivalent to the unconstrained solution, since $\rho = 0.5$

reduces Eq. A3.32 to Eq. A3.28.

A3.6 Theoretical solution for $q=1$

The analytical solution to Eq. A3.5 with $q=1$ and arbitrary f was first derived by Brunel et al. (2004). Here, however, we will use the formulation derived by Zhang et al. (2019). The optimal weight density is given by

$$\rho = F_1(x) \tag{A3.34}$$

where x is obtained by finding the variables (x, v_-, v_+, σ) that solve the set of equations

$$\begin{cases} F_2(x) = \dfrac{\sqrt{2}}{\sigma} \\ F_3(x) = \dfrac{2K_1^2 N}{\sigma^2 (v_- + v_+)^2 f(1-f)} \\ \dfrac{f F_1(v_-) + (1-f) F_1(v_+)}{f F_2(v_-) + (1-f) F_2(v_+)} = \dfrac{-K_1^2 N}{\sqrt{2}\sigma x (v_- + v_+) f(1-f)} \\ f F_2(v_-) - (1-f) F_2(v_+) = 0 \\ v_- + v_+ > 0 \\ \sigma > 0 \end{cases} \tag{A3.35}$$

and where

$$\begin{cases} F_1(x) = \frac{1}{2}(1 + \mathrm{erf}(x)) \\ F_2(x) = \frac{1}{\sqrt{\pi}} e^{-x^2} + x(1 + \mathrm{erf}(x)) \\ F_3(x) = F_1(x) + x F_2(x) \end{cases} \tag{A3.36}$$

The error margin K_1 is linked to the storage load according to

$$\alpha = \frac{2K_1^2 N}{\sigma^2 (v_- + v_+)^2 f(1-f)} \frac{f F_3(v_-) + (1-f) F_3(v_+)}{(f F_1(v_-) + (1-f) F_1(v_+))^2} \,. \tag{A3.37}$$

A3.7 Learning rule for $q=1$ and fixed inhibition

In the specific case $q=1$ ($z=2$), our general learning rule, as described in discrete time in Algorithm 1, reduces to:

Algorithm 2 Sparse Optimal Perceptron ($q=1$)

for $t = 0, 1, 2, \ldots$ **do**
  $\xi^\star \leftarrow \arg\min_{\xi_\mu} (2\xi_i^\mu - 1)\left(\sum_j w_{ij}\xi_j^\mu - I_{\text{inh},i}\right)$ ▷ tagging weakest pattern
  $\hat{u}_{ij} \leftarrow u_{ij}\left(1 + \eta(2\xi_i^\star - 1)\xi_j^\star\right)$ ▷ Hebbian update
  $u_{ij} \leftarrow \hat{u}_{ij} / \sum_j \hat{u}_{ij}^2$ ▷ weight normalization
  $w_{ij} \leftarrow u_{ij}^2$
  $I_{\text{inh},i} \leftarrow I_{\text{inh},i} - \eta_{\text{inh}}(2\xi_i^\star - 1)$ ▷ inhibition update
end for

For very large networks ($N \to \infty$), the norm $\|\boldsymbol{w}\|_1$ can be implicitly constrained by fixing the inhibition $I_{\text{inh},i}$, since this value sets the scale of the mean excitatory input current (Brunel, 2016). However, if the activity level f is very small, only a tiny fraction of all weights change in each epoch, and a large number of epochs would consequently be required for $\|\boldsymbol{w}\|_1$ to converge to its appropriate value, if the network is initialized far from optimum.

The learning process can be sped up by implicitly changing the inhibition together with the weights. This can be achieved as follows: After applying the inhibitory change $\Delta I_{\text{inh},i}^{(t)} = -\eta_{\text{inh}}(2\xi_i^\star - 1)$ in the last line in Algorithm 2, we can shift the inhibition back to its original value by scaling both $I_{\text{inh},i}^{(t+1)}$ and $\boldsymbol{w}^{(t+1)}$ with the factor

$$1 + \eta_{\text{het}}(2\xi_i^\star - 1) = \frac{I_{\text{inh},i}^{(t)}}{I_{\text{inh},i}^{(t+1)}} = \frac{I_{\text{inh},i}^{(t)}}{I_{\text{inh},i}^{(t)} + \Delta I_{\text{inh},i}^{(t)}} \tag{A3.38}$$

where η_{het} is a small, positive constant. We combine this additional step with the Hebbian update in Algorithm 2 and remove the weight normalization as well as the inhibition update, and rename η to η_{hom}. The result is the sleep-based learning rule, expressed in algorithmic terms in as:

Algorithm 3 Sparse Optimal Perceptron ($q=1$ and I_{inh} const.)

for $t = 0, 1, 2, \ldots$ **do**
  $\xi^\star \leftarrow \arg\min_{\xi_\mu} (2\xi_i^\mu - 1)\left(\sum_j w_{ij}\xi_j^\mu - I_{\text{inh},i}\right)$ ▷ tagging weakest pattern
  $u_{ij} \leftarrow u_{ij} + u_{ij}\left(\eta_{\text{het}} + \eta_{\text{hom}}\xi_j^\star\right)(2\xi_i^\star - 1)$ ▷ Hebbian update with cross-talk
  $w_{ij} \leftarrow u_{ij}^2$
end for

This implicitly allows $I_{\text{inh},i}$ and, by extension, $\|\boldsymbol{w}\|_1$, to converge quickly to the optimal value, even if f is small. Note that the weight update in Algorithm 3 also should include the cross-term $\eta_{\text{het}}\eta_{\text{hom}}\xi_j^\star u_{ij}$, but we omit this due to the fact that $\eta_{\text{het}}\eta_{\text{hom}}$ is much smaller than both η_{het} and η_{hom} individually.

A3.8 Simulation parameters

Table A3.1: Simulation parameters used to produce the results in Fig. 3.2.

Parameter	$z = 1$	$z = 2$	$z = 3$	$z = 4$
η	10^{-4}	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$7 \cdot 10^{-3}$
η_{inh}	10^{-3}	$5 \cdot 10^{-3}$	$7 \cdot 10^{-3}$	$7 \cdot 10^{-3}$
$\|\boldsymbol{u}\|_2^2$	10	20	50	100

Table A3.2: Simulation parameters used to produce the results in Figs. 3.5 and 3.6.

Parameter	Value
η_{het}	10
η_ϵ	0.05
$\mathbb{E}[\epsilon_{0,1}]$	0.1^*
$\mathbb{V}[\epsilon_{0,1}]$	1^*
$\|\boldsymbol{w}\|_q^q$	1000
τ_{fast}	1
τ_{slow}	30
T_{sim}	999
T_{sample}	3
$\mathrm{d}t$	$5 \cdot 10^{-3}$

* The noise quantities $\epsilon_{0,1}$ are i.i.d. and drawn from a normal distribution.

Table A3.3: Simulation parameters used to produce the results in Fig. 3.4.

Parameter	Wake	Sleep[1]
η_{het}	1.582	$(1 + 24 \cdot [1 - \exp(-t/100)]) \cdot 10^{-3}$
η_{hom}	0.0316	$(2 + 48 \cdot [1 - \exp(-t/100)]) \cdot 10^{-3}$
η_{inh}	0.01	–
$\|\boldsymbol{w}\|_1$	100	–

[1] Learning rates were increased gradually during sleep with a time constant of 100 epochs, where the epoch is denoted with t.

A3.9 Synapse metadata

The following two tables contain details on the experimental data used to produce Figs. 3.5 and 3.6.

Table A3.4: Descriptions of the synapse datasets with short sampling intervals ($\Delta t = 10\,\text{min} - 7\,\text{h}$), ordered roughly by sample size.

Reference	Setting	Δt	Measure	Condition	Datapoints[1]	Weight (%)
Hazan & Ziv (2020)[2,4]	Rat Ctx culture	1 h	PSD95 FI	silent ctrl	45 600 (43 890) 39 677 (43 016)	13.6 (12.3) 13.1 (14.3)
Hazan & Ziv (2020)[3,4]	Rat Ctx culture	1 h	PSD95 FI	silent ctrl	44 498 (43 676) 39 631 (41 135)	10.6 (15.3) 15.8 (16.9)
Kaufman et al. (2012)	Rat Ctx culture	30 min	PSD95 FI	ctrl	25 847 (25 845)	25.0 (15.6)
Fisher-Lavie & Ziv (2013)	Mouse Ctx culture	25 min	PSD95 FI Munc13 FI	ctrl ctrl	9536 (10 347) 9545 (10 353)	5.8 (7.2) 5.8 (4.7)
Miyamoto et al. (2021)	Mouse MCtx L2/3 PC in vivo	7 h	GluA1 FI GluA1 FI SH FI SH FI	sleep wake sleep wake	1039 (1270) 346 (405) 1107 (1202) 371 (380)	3.1 (3.9) 1.0 (1.0) 2.3 (2.7) 0.8 (0.7)
Ishii et al. (2018)	Mouse VCtx L5 PC-ad in vivo	10 min	SH FI	WT Fmr1-KO	238 (237) 714 (719)	0.4 (0.4) 1.6 (2.4)
Wegner et al. (2022)	Mouse VCtx L5 PC-ad in vivo	30 min	PSD95 area PSD95 area SH area SH area	EE ctrl EE ctrl	169 (280) 105 (228) 237 (215) 161 (169)	0.3 (1.2) 0.1 (0.9) 0.4 (0.3) 0.2 (0.3)

Abbreviations: Ctx = cortex, ACtx = auditory cortex, BCtx = barrel cortex, MCtx = motor cortex, VCtx = visual cortex, PC = pyramidal cell, ad = apical dendrite, FI = fluorescence intensity, SH = spine head, ctrl = control, WT = wild-type, KO = knockout, EE = environmental enrichment.

[1] This is the total number of $(\hat{w}, \Delta\hat{w})$-pairs. It is determined both by the number of imaged synapses and the number of imaging sessions. Values outside parenthesis refer to potentiation ($\Delta\hat{w} > 0$) while those inside refer to depression ($\Delta\hat{w} < 0$).

[2] Smoothened data. See original publication for details.

[3] Raw (non-smoothened) data.

[4] Weights for smoothened and raw data have been halfed to avoid counting the same data twice.

Table A3.5: Descriptions of the synapse datasets with long sampling intervals ($\Delta t \geq 48\,\text{h}$), ordered roughly by sample size. Abbreviations and notation as in Table A3.4.

Reference	Setting	Δt	Measure	Condition	Datapoints	Weight (%)
Gala et al. (2017)[1]	Mouse BCtx L2/3/5 in vivo	96 h	Bouton FI	ctrl	12 829 (12 773)	72.0 (57.3)
Loewenstein et al. (2011)	Mouse ACtx L5 PC-ad in vivo	96 h	SH FI	ctrl	2459 (2552)	16.5 (31.3)
Ishii et al. (2018)	Mouse VCtx L5 PC-ad in vivo	48 h	SH FI	WT Fmr1-KO	350 (404) 417 (461)	4.7 (4.2) 6.0 (6.4)
Steffens et al. (2021)	Mouse MCtx L5 PC-ad in vivo	72-96 h	SH area	ctrl	168 (244)	0.8 (0.8)

[1] We included only measurements for which the bouton detection probability was $>90\%$.

A3.10 Supplementary figures

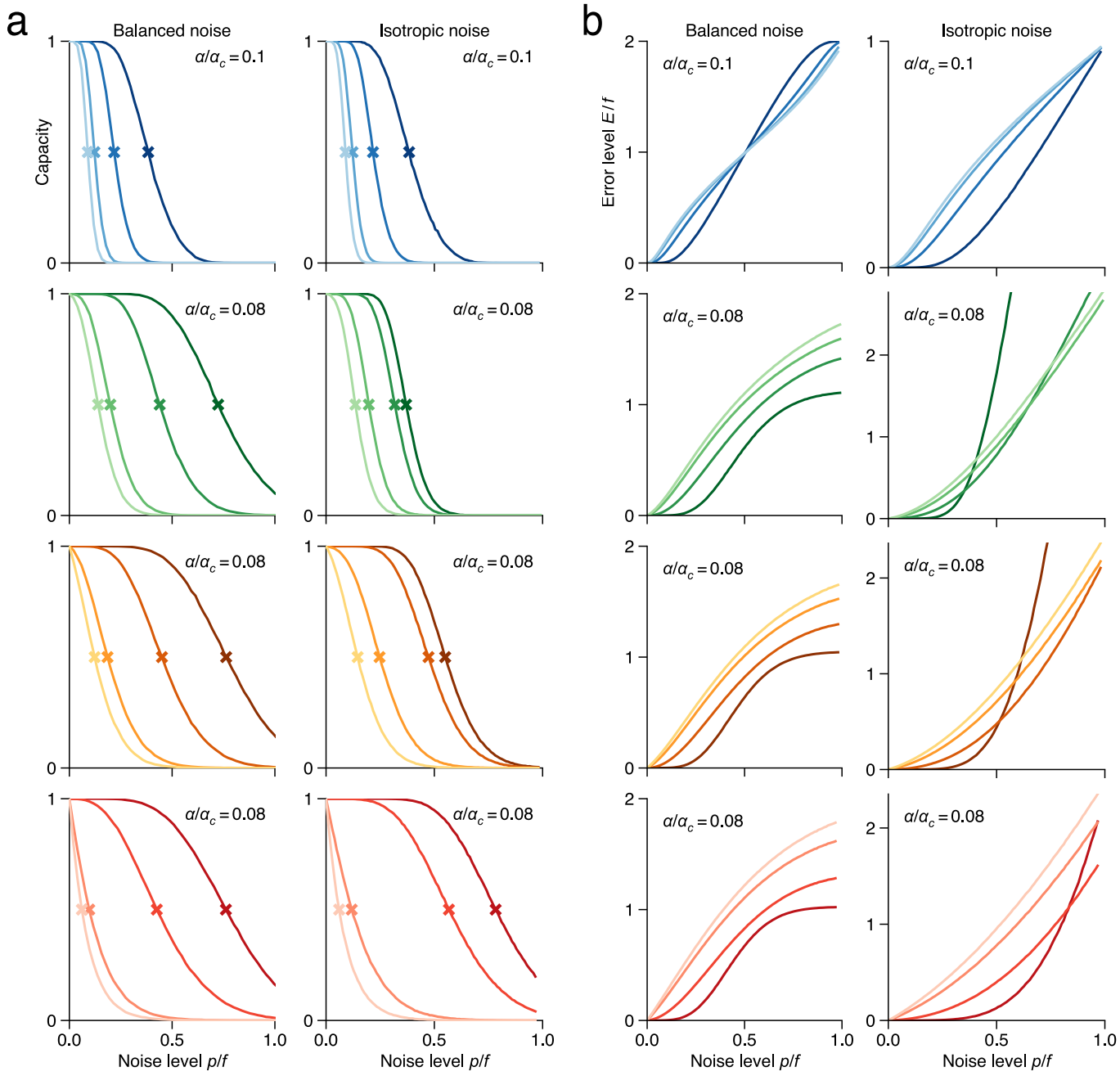


Figure A3.1: Recall testing in optimal attractor networks. **(a)** The fraction of memories that can be successfully retrieved, as a function of balanced and isotropic input noise, for pattern activity levels $f = 0.5$ (blues), $f = 0.1$ (greens), $f = 0.05$ (oranges), and $f = 0.03$ (reds). Crosses indicate the 50% capacity level. **(b)** Error rate after one synchronous update, as a function of input noise.

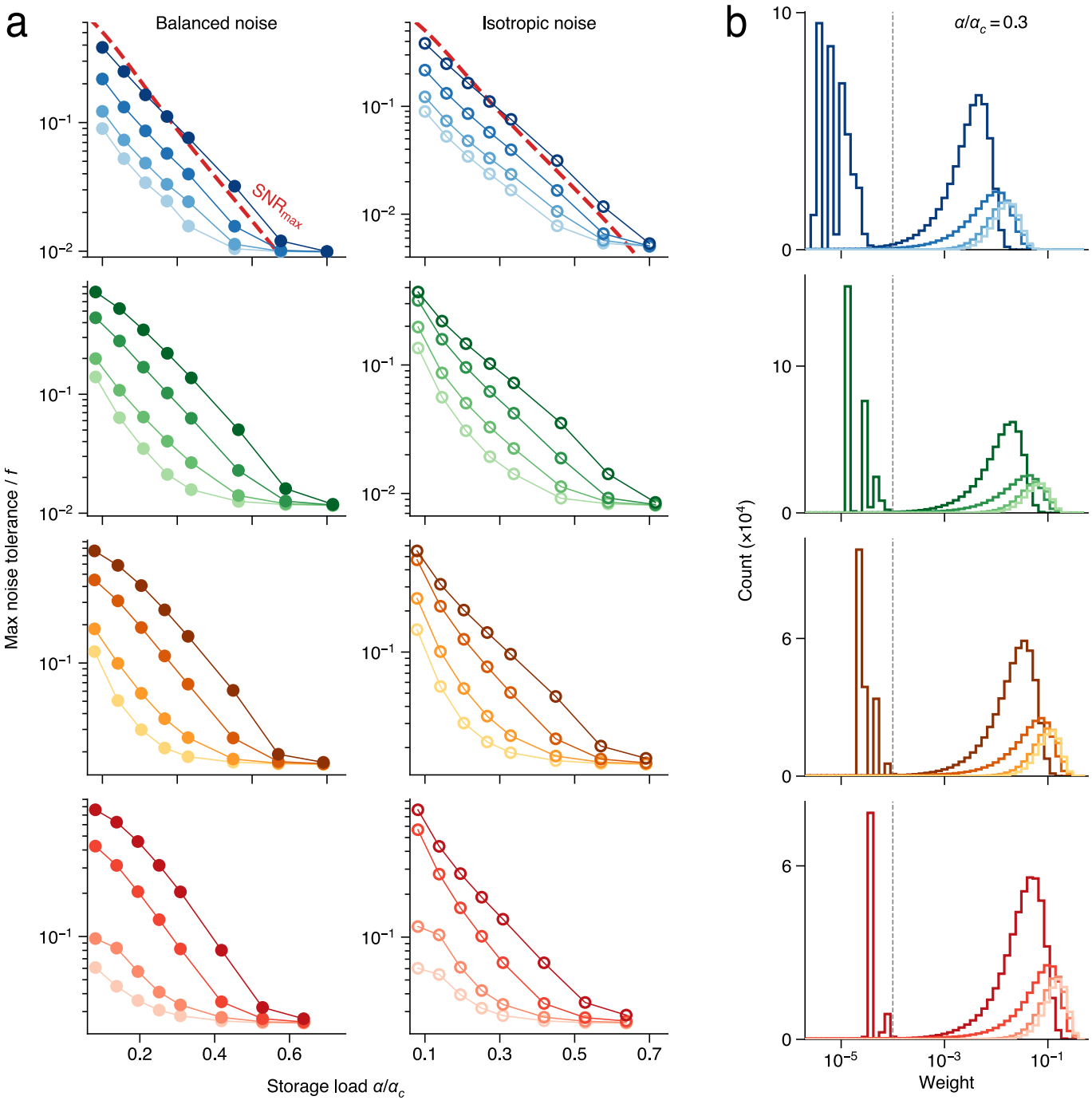


Figure A3.2: Additional properties of optimal attractor networks. **(a)** The highest tolerable noise level. **(b)** The distribution of weights. Although it is difficult to produce sparse solutions in the case $z = 1$ (darkest lines) due to additive weight updates and a fixed learning rate, we estimated the sparsity of the solution using the cutoff 10^{-4} (dashed vertical line).

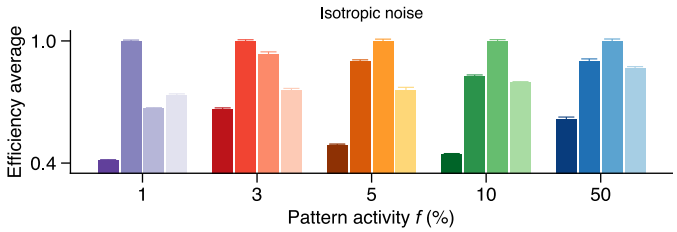


Figure A3.3: The efficiency, averaged over all storage loads and distortion levels with isotropic noise (mean $\pm$ SD). This is highest at either $z=2$ or 3.

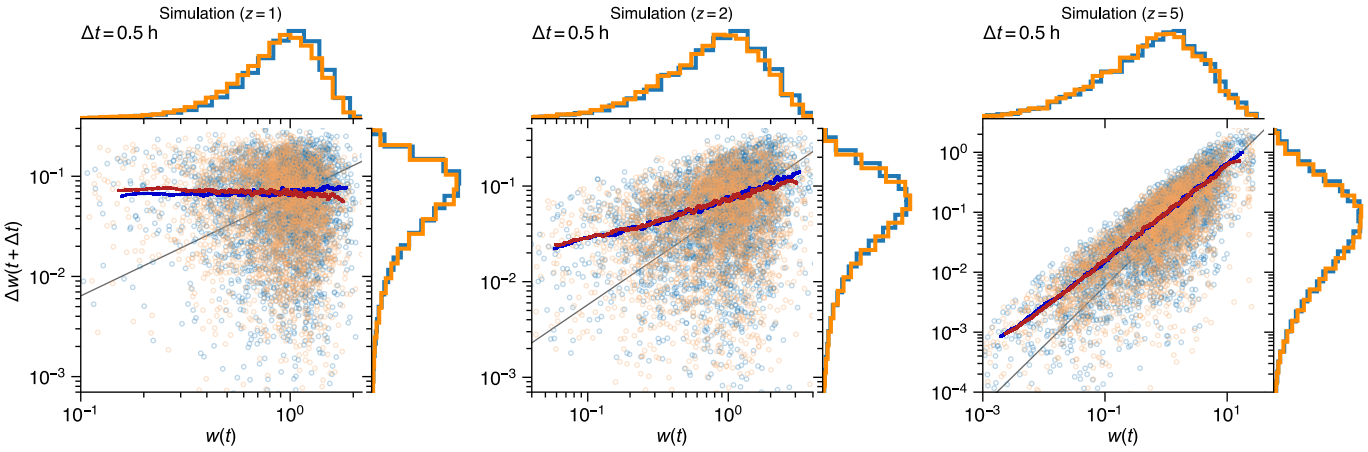


Figure A3.4: Simulated synaptic volatility. Each panel corresponds to a simulation of 1000 synapses governed by the stochastic process in Eq. 3.8 with different z (color legend can be found in Fig. 3.5a). The sampling time is scaled relative the characteristic time constant of the synapses to represent 30 min of biological time. The dark lines produced by the moving average follow power-laws, where the exponent (slope) increases with z (left to right panel).

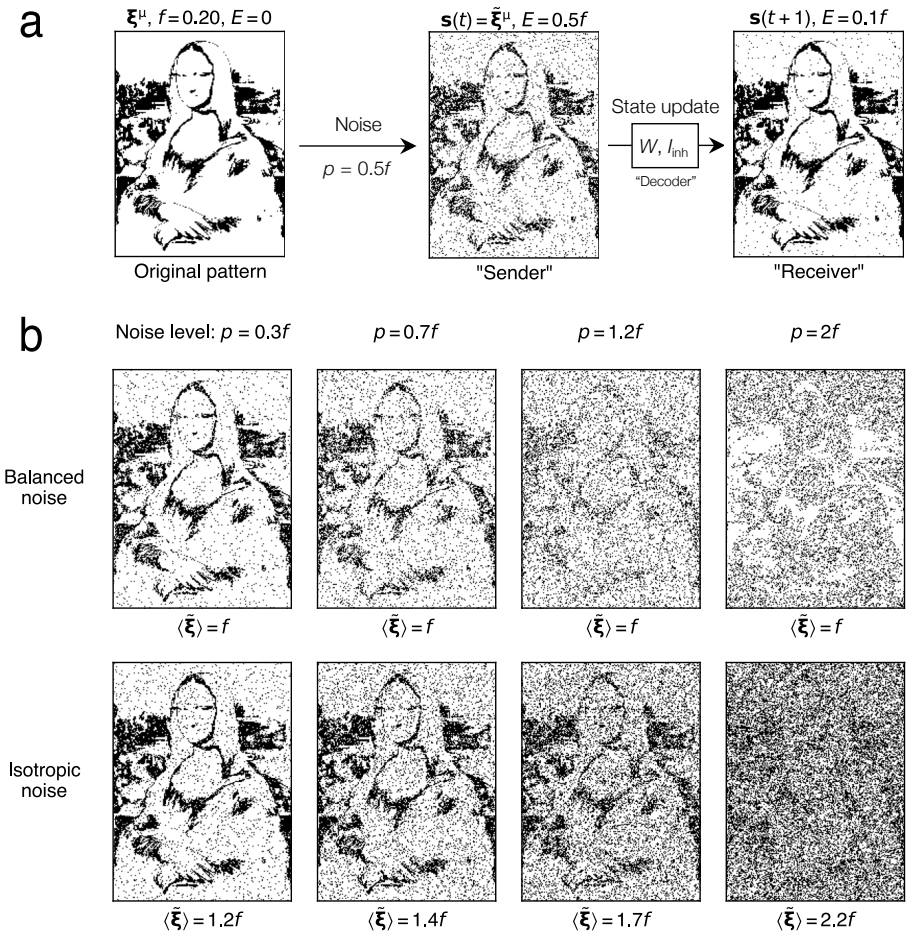


Figure A3.5: The effect of different noise models. **(a)** Illustration of the idea of how pattern retrieval from a noisy initialization can be seen a error-correction of a message communicated through a noisy channel. **(b)** Demonstration of the effect of balanced and isotropic noise on a binary pattern.

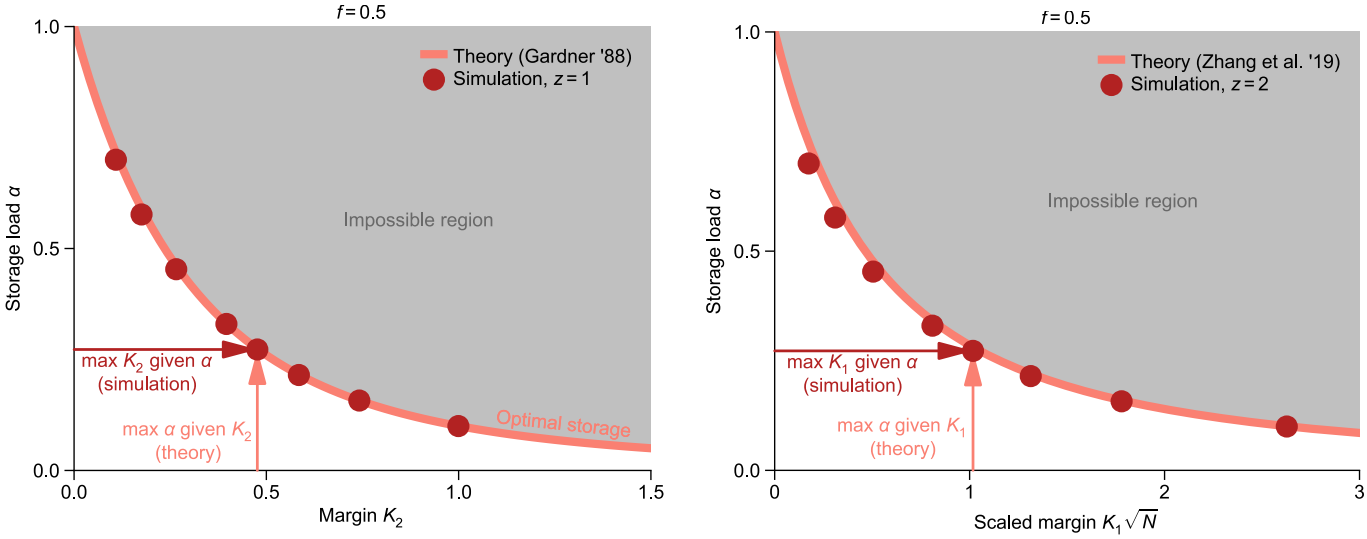


Figure A3.6: Theoretical and numerical storage optimization. In theoretical calculations of optimal storage, the load is maximized, with the margin considered fixed. In numerical optimization, the margin is instead maximized, with the load considered fixed. This is demonstrated both for K_2 (left) and K_1 (right).

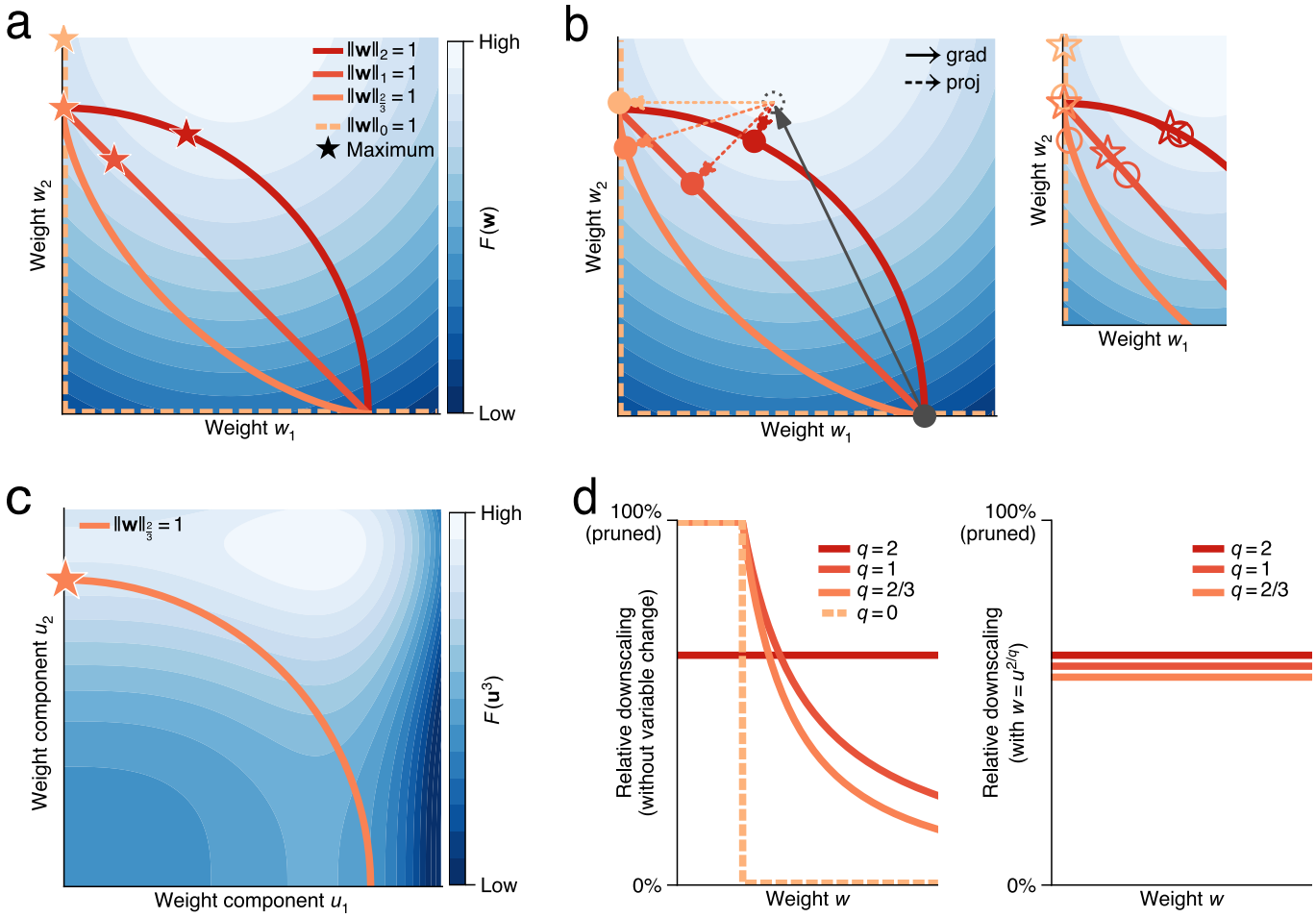


Figure A3.7: The effect of tripartite weights on the optimization landscape. **(a)** The landscape of the objective function $F(\mathbf{w}) = -1.5(w_1 - 0.55)^2 - (w_2 - 1.4)^2$ together with the minima, subject to certain weight constraints. **(b)** One step of projected gradient ascent. **(c)** The landscape of the objective function $F(\mathbf{u}^3) = -1.5(u_1^3 - 0.55)^2 - (u_2^3 - 1.4)^2$. **(d)** The downscaling imposed on weights during gradient descent with different forms of regularization, both without variable change (left) and with variable change to u (right).

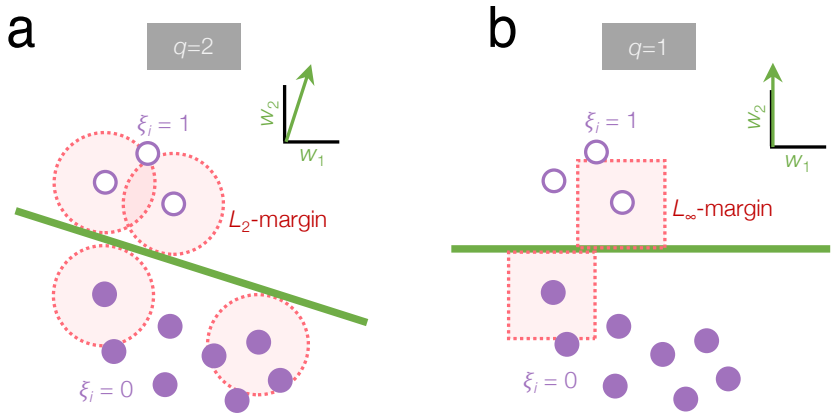


Figure A3.8: Geometrical explanation of $K_{q=1}$ maximization. **(a)** Maximization of K_2 corresponds to maximization of the L_2-margin. **(b)** Maximization of K_1 corresponds to a maximization of the L_∞-margin (Mangasarian, 1999; Rosset et al., 2003).

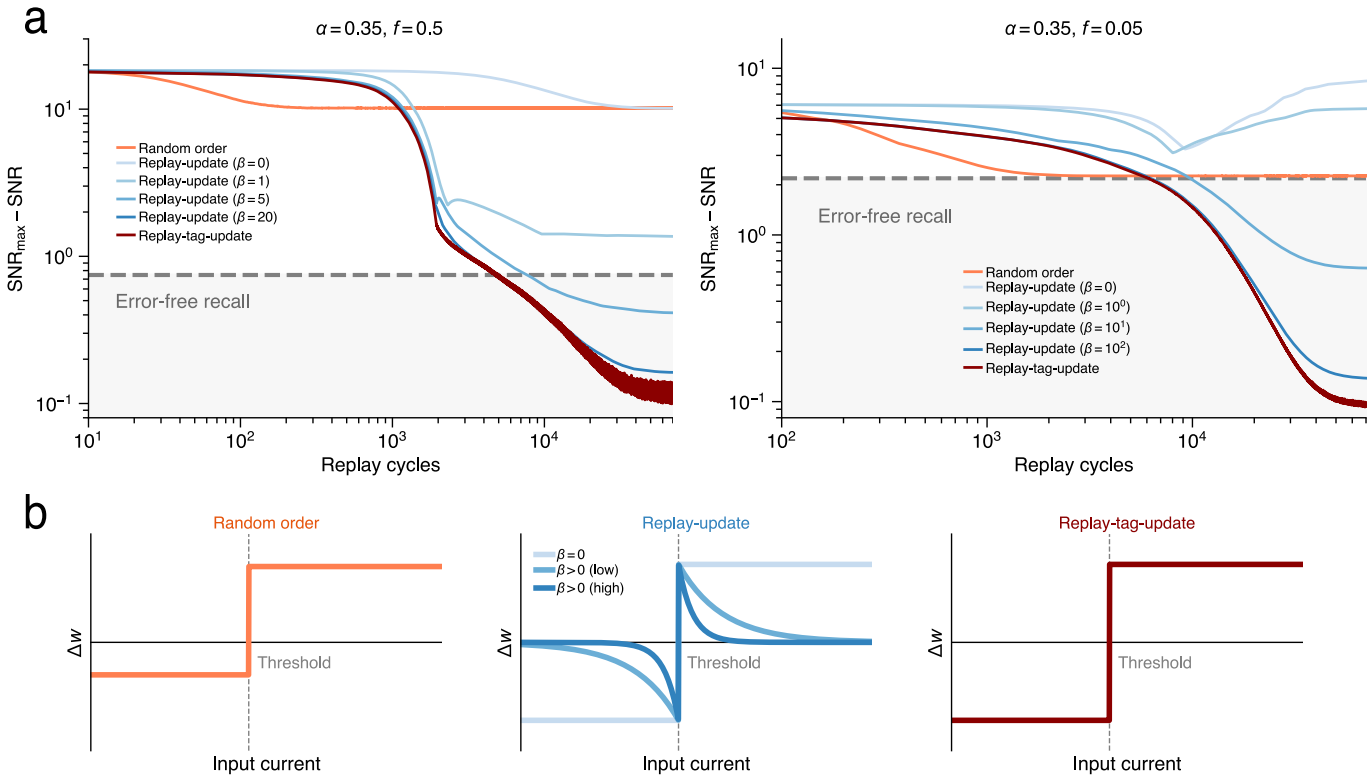


Figure A3.9: Control models for SNR maximization. **(a)** Comparison, in terms of closeness to $\mathrm{SNR}_{\max}$, between our consolidation model ($z = 1$, red line), a learning rule where patterns are updated in random order (orange line), and the normalized gradient descent algorithm for the exponential loss (Nacson et al., 2019), which does not require a tagging mechanism (blue lines). All algorithms use binary patterns, non-negative weights, and weight normalization. **(b)** The plasticity rules corresponding to each of the consolidation algorithms.

Chapter 4

Conclusion

Recurrent attractor neural networks are today the preferred mathematical tools for modeling local cortical circuits and, in particular, long-term memory. These models have been extensively studied for more than four decades, due to their intuitive accessibility and analytical simplicity. In this thesis, we have used the attractor network framework as a theoretical basis and starting point to answer a set of fundamental questions regarding learning and memory formation in cortex.

First, we sought to quantify and mathematically formalize the concept of engram consolidation in a neural network. Second, we asked how the process of synaptic pruning can be framed as a mathematical problem, and if it, in turn, can be linked to consolidation. Provided the definition of such a problem, our third question was whether or not a network can be trained to find a solution using a biologically plausible learning rule. Finally, supposing that the previous three questions can be answered, we asked if it is possible to use a learning rule for consolidation and pruning to resolve some of the discrepancies between current models of synaptic plasticity and the synaptic dynamics observed in recent experimental data.

The first question was treated in chapter 2, where we began by defining consolidation as the maximization of the signal-to-noise ratio of an engram from the perspective of a single neuron. The closed-form solution to this problem allowed us to compactly describe the structure of optimally noise-robust hetero- and auto-associative memory networks. Importantly, we demonstrated that this class of networks generalizes the famous memory models that we previously referred to as first- and third-generation models, or $1/\sqrt{N}$-models, such as the different variants of the Hopfield network and the Kanerva network. This approach is fundamentally different from the classical energy-based method, and offers a new, simple perspective on optimal memory encoding, based on the idea of max-margin classification. It also naturally incorporates a generalization of dendritic processing in the form of a kernel, thus providing a direct link between storage capacity and neuron complexity.

In chapter 3, we provided an answer to the second question, by framing synaptic pruning as a regularization, which is implicitly applied to the consolidation process by partitioning each synapse and representing it as a product of sub-cellular components. This, in turn, allowed us tackle our third question: we modified the problem of engram consolidation and derived a learning rule that encodes memories in both a noise-robust and energy-efficient manner, using a only a small fraction of all available synapses. More importantly, the learning rule offers a consolidation-based explanation to the function of memory replay, homeostatic scaling,

and weight-dependent synaptic plasticity, thus serving as a first step towards answering the fourth and final question of the thesis.

A common thread throughout the results in this thesis has been the significance of neural and synaptic complexity in learning. While the results in chapter 2 suggest that separate dendritic compartments with active properties play an important role in enhancing the storage capacity of cortical circuits, the findings in chapter 3 suggest that the intricate internal machinery of synaptic connections can serve as a regularization mechanism, which forces cortical connectivity to be sparse and energy efficient. Although the first conclusion already has been proposed in past theoretical (Poirazi & Mel, 2001) and experimental studies (Gidon et al., 2020), the second conclusion is, to the best of our knowledge, novel to the field of neuroscience, despite being a well-established idea in statistics and machine learning (Hoff, 2017; Amid & Warmuth, 2020; Schwarz et al., 2021).

Our approach to modeling plasticity and consolidation has, at its core, relied on the normative assumption that cortical anatomy and dynamics can, at least on an abstract level, be understood in terms of an optimization process that seeks to solve specific cognitive tasks (Richards et al., 2019). While this methodology requires a mathematical description of cortical circuits that is heavily simplified, with one- or two-compartment neurons evolving synchronously in discretized time, the aim of our work has been to disentangle and identify the necessary mechanisms for efficient consolidation, and form a compact, general understanding of this process that rests on as few axioms as possible.

There are mainly three directions in which our work can be directly continued. First, the predictions formulated at the end of chapter 3, regarding sleep, spine dynamics, and developmental pruning, offer an entry-point to experimentally test the validity of our theory.

Second, our model of structural plasticity makes no statement about the exact nature of LTP- or LTD-induction in existing synapses, and can therefore be considered to be complementary to models of functional plasticity. It should therefore be straightforward to implement our plasticity rule as an added feature in existing large-scale computational models with anatomically and biophysically detailed neurons, in order to evaluate our conclusions in more realistic simulations of cortical plasticity (see, e.g., Chindemi et al., 2022).

Third, the generality of our method allows it to be directly applied to other cognitive tasks, neuron types, or network architectures where one wishes to study the relationship between function, connectivity, and energy-efficiency from an optimization-based perspective. In fact, any neural network model, whether it is deterministic or probabilistic, is amenable to this analysis, as long as the task at hand can be defined in terms of an objective function paired with a regularization. Situations like this naturally arise, in the deterministic case, when dealing with conventional constrained maximization or minimization problems, or, in the case of probabilistic models, when performing variational Bayesian inference. This approach can be particularly useful in settings that lack sufficient data or prior knowledge to build a bottom-up model, and where top-down results can narrow down the model search space and generate hypotheses that initiate the cycle of theorization and experimentation.

The importance of understanding the link between sparse, parsimonious neural network models on one hand, and cognitive ability or task performance on the other, has recently been emphasized both in the neuroscience and the machine learning community. For machine learning practitioners, the impressive achievements of modern deep learning applications has come at the price of dramatically increased network sizes. Today, cutting-edge models comprise billions of parameters, which require substantial amounts of data and computing resources to be trained. This has generated an entire subfield of research into methods for sparsifying deep networks and identifying more economical models with comparable levels of performance (Hoefler et al., 2021).

For neuroscientists, understanding how the sparse, structured connectivity of neocortex emerges, in terms of a small number of organizing principles or mathematical learning rules, is considered a key goal of the field. An accurate description of the intrinsic and experience-dependent dynamics of connections would not only offer valuable insight into cortical function and the formation of cortical representations, but it would also establish a useful link between the algorithmic and mechanistic components on the (microscopic) single-synapse level, and the statistics of connectivity on the (mesoscopic) circuit level. This would be a crucial first step towards understanding, for example, how numerous neuropsychiatric disorders, such as schizophrenia, Alzheimer's disease, autism spectrum disorder, bipolar disorder, and depression cause abnormal structural plasticity patterns in development and adulthood, and how this, ultimately, produces cognitive dysfunction (Cochran et al., 2014; Forrest et al., 2018).

Bibliography

Abbott, L. F. & Arian, Y. (1987). "Storage capacity of generalized networks". *Physical Review A* 36, pp. 5091–5094.

Aghajanian, G. K. & Bloom, F. E. (1967). "The formation of synaptic junctions in developing rat brain: A quantitative electron microscopic study". *Brain Research* 6, pp. 716–727.

Aime, M., Calcini, N., Borsa, M., Campelo, T., Rusterholz, T., Sattin, A., Fellin, T. & Adamantidis, A. (2022). "Paradoxical somatodendritic decoupling supports cortical plasticity during REM sleep". *Science* 376, pp. 724–730.

Alemi, A., Baldassi, C., Brunel, N. & Zecchina, R. (2015). "A three-threshold learning rule approaches the maximal capacity of recurrent neural networks". *PLOS Computational Biology* 11, e1004439.

Amari, S.-I. (1972). "Learning patterns and pattern sequences by self-organizing nets of threshold elements". *IEEE Transactions on Computers* C-21, pp. 1197–1206.

Amari, S.-I. (1989). "Characteristics of sparsely encoded associative memory". *Neural Networks* 2, pp. 451–457.

Amid, E. & Warmuth, M. K. (2020). "Winnowing with gradient descent". *Proceedings of the 33rd Conference on Learning Theory*, pp. 163–182.

Amit, D. J., Campbell, C. & Wong, K. Y. M. (1989). "The interaction space of neural networks with sign-constrained synapses". *Journal of Physics A: Mathematical and General* 22, pp. 4687–4693.

Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985). "Storing infinite numbers of patterns in a spin-glass model of neural networks". *Physical Review Letters* 55, pp. 1530–1533.

Anlauf, J. K. & Biehl, M. (1989). "The AdaTron: an adaptive perceptron algorithm". *Europhysics Letters* 10, pp. 687–692.

Ardeshir, N., Sanford, C. & Hsu, D. (2021). "Support vector machines and linear regression coincide with very high-dimensional features". *Advances in Neural Information Processing Systems 35*.

Arellano, J. I., Benavides-Piccione, R., DeFelipe, J. & Yuste, R. (2007). "Ultrastructure of dendritic spines: correlation between synaptic and spine morphologies". *Frontiers in Neuroscience* 1.

Ashton, J. E. & Cairney, S. A. (2021). "Future-relevant memories are not selectively strengthened during sleep". *PLOS One* 16, e0258110.

Baldassi, C., Gerace, F., Lucibello, C., Saglietti, L. & Zecchina, R. (2016). "Learning may need only a few bits of synaptic precision". *Physical Review E* 93, p. 052313.

Bansal, Y., Advani, M., Cox, D. D. & Saxe, A. M. (2018). "Minnorm training: an algorithm for training over-parameterized deep neural networks". Preprint. arXiv: 1806.00730.

Barry, D. N. & Maguire, E. A. (2019). "Remote memory and the hippocampus: a constructive critique". *Trends in Cognitive Sciences* 23, pp. 128–142.

Benna, M. K. & Fusi, S. (2016). "Computational principles of synaptic memory consolidation". *Nature Neuroscience* 19, pp. 1697–1706.

Bennett, E. L., Diamond, M. C., Krech, D. & Rosenzweig, M. R. (1964). "Chemical and anatomical plasticity of brain". *Science* 146, pp. 610–619.

Berning, S., Willig, K. I., Steffens, H., Dibaj, P. & Hell, S. W. (2012). "Nanoscopy in a living mouse brain". *Science* 335, pp. 551–551.

Bi, G.-q. & Poo, M.-m. (1998). "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type". *Journal of Neuroscience* 18, pp. 10464–10472.

Bliss, T. V. P. & Lømo, T. (1973). "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path". *The Journal of Physiology* 232, pp. 331–356.

Boros, G. & Moll, V. (2004). *Irresistible integrals: symbolics, analysis and experiments in the evaluation of integrals*. Cambridge University Press.

Bouten, M., Engel, A., Komoda, A. & Serneels, R. (1990). "Quenched versus annealed dilution in neural networks". *Journal of Physics A: Mathematical and General* 23, p. 4643.

Bricken, T. & Pehlevan, C. (2021). "Attention approximates sparse distributed memory". *Advances in Neural Information Processing Systems 34*.

Brunel, N. (2016). "Is cortical connectivity optimized for storing information?" *Nature Neuroscience* 19, pp. 749–755.

Brunel, N., Hakim, V., Isope, P., Nadal, J.-P. & Barbour, B. (2004). "Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell". *Neuron* 43, pp. 745–757.

Butz, M. & van Ooyen, A. (2013). "A simple rule for dendritic spine and axonal bouton formation can account for cortical reorganization after focal retinal lesions". *PLOS Computational Biology* 9, e1003259.

Casali, D., Costantini, G., Perfetti, R. & Ricci, E. (2006). "Associative memory design using support vector machines". *IEEE Transactions on Neural Networks* 17, pp. 1165–1174.

Chapeton, J., Fares, T., LaSota, D. & Stepanyants, A. (2012). "Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons". *Proceedings of the National Academy of Sciences* 109, E3614–E3622.

Chapeton, J., Gala, R. & Stepanyants, A. (2015). "Effects of homeostatic constraints on associative memory storage and synaptic connectivity of cortical circuits". *Frontiers in Computational Neuroscience* 9, p. 74.

Chechik, G., Meilijson, I. & Ruppin, E. (1998). "Synaptic pruning in development: a computational account". *Neural Computation* 10, pp. 1759–1777.

Chen, J.-R., Wang, Y.-J. & Tseng, G.-F. (2003). "The effect of epidural compression on cerebral cortex: a rat model". *Journal of Neurotrauma* 20, pp. 767–780.

Chen, L. & Xu, S. (2020). "Deep neural tangent kernel and Laplace kernel have the same RKHS". *Proceedings of the 2021 International Conference on Learning Representations*.

Chen, S. X., Kim, A. N., Peters, A. J. & Komiyama, T. (2015). "Subtype-specific plasticity of inhibitory circuits in motor cortex during motor learning". *Nature Neuroscience* 18, pp. 1109–1115.

Chindemi, G., Abdellah, M., Amsalem, O., Benavides-Piccione, R., Delattre, V., Doron, M., Ecker, A., Jaquier, A. T., King, J., Kumbhar, P., Monney, C., Perin, R., Rössert, C., Tuncel, A. M., Van Geit, W., DeFelipe, J., Graupner, M., Segev, I., Markram, H. & Muller, E. B. (2022). "A calcium-based plasticity model for predicting long-term potentiation and depression in the neocortex". *Nature Communications* 13, p. 3038.

Cho, Y. & Saul, L. (2009). "Kernel methods for deep learning". *Advances in Neural Information Processing Systems 22*.

Chou, P. A. (1989). "The capacity of the Kanerva associative memory". *IEEE Transactions on Information Theory* 35, pp. 281–298.

Clopath, C., Büsing, L., Vasilaki, E. & Gerstner, W. (2010). "Connectivity reflects coding: a model of voltage-based STDP with homeostasis". *Nature Neuroscience* 13, pp. 344–352.

Cochran, J. N., Hall, A. M. & Roberson, E. D. (2014). "The dendritic hypothesis for Alzheimer's disease pathophysiology". *Brain Research Bulletin* 103, pp. 18–28.

Cortes, C. & Vapnik, V. (1995). "Support-vector networks". *Machine Learning* 20, pp. 273–297.

Cossart, R., Aronov, D. & Yuste, R. (2003). "Attractor dynamics of network UP states in the neocortex". *Nature* 423, pp. 283–288.

Cotter, A., Shalev-Shwartz, S. & Srebro, N. (2012). "The kernelized stochastic batch perceptron". *Proceedings of the 29th International Conference on Machine Learning*, pp. 739–746.

Cover, T. M. (1965). "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition". *IEEE Transactions on Electronic Computers* EC-14, pp. 326–334.

Cragg, B. G. (1975). "The development of synapses in kitten visual cortex during visual deprivation". *Experimental Neurology* 46, pp. 445–451.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. & Singer, Y. (2006). "Online passive-aggressive algorithms". *Journal of Machine Learning Research* 7, pp. 551–585.

Crick, F. & Mitchison, G. (1983). "The function of dream sleep". *Nature* 304, pp. 111–114.

De Paola, V., Holtmaat, A., Knott, G., Song, S., Wilbrecht, L., Caroni, P. & Svoboda, K. (2006). "Cell type-specific structural plasticity of axonal branches and boutons in the adult neocortex". *Neuron* 49, pp. 861–875.

De Roo, M., Klauser, P., Mendez, P., Poglia, L. & Muller, D. (2008). "Activity-dependent PSD formation and stabilization of newly formed spines in hippocampal slice cultures". *Cerebral Cortex* 18, pp. 151–161.

Deger, M., Seeholzer, A. & Gerstner, W. (2018). "Multicontact co-operativity in spike-timing–dependent structural plasticity stabilizes networks". *Cerebral Cortex* 28, pp. 1396–1415.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S. & Vermet, F. (2017). "On a model of associative memory with huge storage capacity". *Journal of Statistical Physics* 168, pp. 288–299.

Denis, D., Mylonas, D., Poskanzer, C., Bursal, V., Payne, J. D. & Stickgold, R. (2021). "Sleep spindles preferentially consolidate weakly encoded memories". *Journal of Neuroscience* 41, pp. 4088–4099.

Denis, D., Schapiro, A. C., Poskanzer, C., Bursal, V., Charon, L., Morgan, A. & Stickgold, R. (2020). "The roles of item exposure and visualization success in the consolidation of memories across wake and sleep". *Learning & Memory* 27, pp. 451–456.

Diamond, M. C., Lindner, B. & Raymond, A. (1967). "Extensive cortical depth measurements and neuron size increases in the cortex of environmentally enriched rats". *Journal of Comparative Neurology* 131, pp. 357–364.

Dorkenwald, S., Turner, N. L., Macrina, T., Lee, K., Lu, R., Wu, J., Bodor, A. L., Bleckert, A. A., Brittain, D., Kemnitz, N., Silversmith, W. M., Ih, D., Zung, J., Zlateski, A., Tartavull, I., Yu, S.-C., Popovych, S., Wong, W., Castro, M., Jordan, C. S., Wilson, A. M., Froudarakis, E., Buchanan, J., Takeno, M. M., Torres, R., Mahalingam, G., Collman, F., Schneider-Mizell, C. M.,

Bumbarger, D. J., Li, Y., Becker, L., Suckow, S., Reimer, J., Tolias, A. S., Macarico da Costa, N., Reid, R. C. & Seung, H. S. (2022). "Binary and analog variation of synapses between cortical pyramidal neurons". *eLife* 11, e76120.

Fauth, M., Wörgötter, F. & Tetzlaff, C. (2015). "Formation and maintenance of robust long-term information storage in the presence of synaptic turnover". *PLOS Computational Biology* 11, e1004684.

Feldman, M. L. & Dowd, C. (1975). "Loss of dendritic spines in aging cerebral cortex". *Anatomy and Embryology* 148, pp. 279–301.

Fenn, K. M. & Hambrick, D. Z. (2012). "Individual differences in working memory capacity predict sleep-dependent memory consolidation." *Journal of Experimental Psychology: General* 141, p. 404.

Fenn, K. M. & Hambrick, D. Z. (2015). "General intelligence predicts memory change across sleep". *Psychonomic Bulletin & Review* 22, pp. 791–799.

Fifková, E. (1968). "Changes in the visual cortex of rats after unilateral deprivation". *Nature* 220, pp. 379–381.

Fisher-Lavie, A. & Ziv, N. E. (2013). "Matching dynamics of presynaptic and postsynaptic scaffolds". *Journal of Neuroscience* 33, pp. 13094–13100.

Forrest, B. M. (1988). "Content-addressability and learning in neural networks". *Journal of Physics A: Mathematical and General* 21, pp. 245–255.

Forrest, M. P., Parnell, E. & Penzes, P. (2018). "Dendritic structural plasticity and neuropsychiatric disease". *Nature Reviews Neuroscience* 19, pp. 215–234.

Frieß, T.-T., Cristianini, N. & Campbell, C. (1998). "The Kernel-Adatron algorithm: a fast and simple learning procedure for support vector machines". *Proceedings of the 15th International Conference on Machine Learning*, pp. 188–196.

Gais, S. & Born, J. (2004). "Low acetylcholine during slow-wave sleep is critical for declarative memory consolidation". *Proceedings of the National Academy of Sciences* 101, pp. 2140–2144.

Gala, R., Lebrecht, D., Sahlender, D. A., Jorstad, A., Knott, G., Holtmaat, A. & Stepanyants, A. (2017). "Computer assisted detection of axonal bouton structural plasticity in in vivo time-lapse images". *eLife* 6, e29315.

Gallinaro, J. V., Gašparović, N. & Rotter, S. (2022). "Homeostatic control of synaptic rewiring in recurrent networks induces the formation of stable memory engrams". *PLOS Computational Biology* 18, e1009836.

Gardner, E. (1987a). "Maximum storage capacity in neural networks". *Europhysics Letters* 4, pp. 481–485.

Gardner, E. (1987b). "Multiconnected neural network models". *Journal of Physics A: Mathematical and General* 20, pp. 3453–3464.

Gardner, E. (1988). "The space of interactions in neural network models". *Journal of Physics A: Mathematical and General* 21, pp. 257–270.

Gardner, E. (1989). "Optimal basins of attraction in randomly sparse neural network models". *Journal of Physics A: Mathematical and General* 22, p. 1969.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D. & Brea, J. (2018). "Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules". *Frontiers in Neural Circuits* 12.

Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., Holtkamp, M., Vida, I. & Larkum, M. E. (2020). "Dendritic action potentials and computation in human layer 2/3 cortical neurons". *Science* 367, pp. 83–87.

Globus, A., Rosenzweig, M. R., Bennett, E. L. & Diamond, M. C. (1973). "Effects of differential experience on dendritic spine counts in rat cerebral cortex". *Journal of Comparative and Physiological Psychology* 82, pp. 175–181.

Graupner, M. & Brunel, N. (2012). "Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location". *Proceedings of the National Academy of Sciences* 109, pp. 3991–3996.

Graves, A., Wayne, G. & Danihelka, I. (2014). "Neural Turing machines". Preprint. arXiv:1410.5401.

Grutzendler, J., Kasthuri, N. & Gan, W.-B. (2002). "Long-term dendritic spine stability in the adult cortex". *Nature* 420, pp. 812–816.

Gutfreund, H. & Stein, Y. (1990). "Capacity of neural networks with discrete synaptic couplings". *Journal of Physics A: Mathematical and General* 23, pp. 2613–2630.

Harris, J. J., Jolivet, R. & Attwell, D. (2012). "Synaptic energy use and supply". *Neuron* 75, pp. 762–777.

Harris, K. D. & Shepherd, G. M. G. (2015). "The neocortical circuit: themes and variations". *Nature Neuroscience* 18, pp. 170–181.

Hayama, T., Noguchi, J., Watanabe, S., Takahashi, N., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Matsuzaki, M. & Kasai, H. (2013). "GABA promotes the competitive selection of dendritic spines by controlling local Ca2+ signaling". *Nature Neuroscience* 16, pp. 1409–1416.

Hayashi-Takagi, A., Yagishita, S., Nakamura, M., Shirai, F., Wu, Y. I., Loshbaugh, A. L., Kuhlman, B., Hahn, K. M. & Kasai, H. (2015). "Labelling and optical erasure of synaptic memory traces in the motor cortex". *Nature* 525, pp. 333–338.

Hazan, L. & Ziv, N. E. (2020). "Activity dependent and independent determinants of synaptic size diversity". *Journal of Neuroscience* 40, pp. 2828–2848.

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. John Wiley & Sons.

Hennequin, G., Agnes, E. J. & Vogels, T. P. (2017). "Inhibitory plasticity: balance, control, and codependence". *Annual Review of Neuroscience* 40, pp. 557–579.

Hintzman, D. L. (1984). "MINERVA 2: A simulation model of human memory". *Behavior Research Methods, Instruments, & Computers* 16, pp. 96–101.

Hiratani, N. & Fukai, T. (2018). "Redundancy in synaptic connections enables neurons to learn optimally". *Proceedings of the National Academy of Sciences* 115, E6871–E6879.

Hoefler, T., Alistarh, D., Ben-Nun, T. & Dryden, N. (2021). "Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks". *Journal of Machine Learning Research* 23.

Hoel, E. (2021). "The overfitted brain: dreams evolved to assist generalization". *Patterns* 2, p. 100244.

Hoff, P. D. (2017). "Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization". *Computational Statistics & Data Analysis* 115, pp. 186–198.

Holler, S., Köstinger, G., Martin, K. A. C., Schuhknecht, G. F. P. & Stratford, K. J. (2021). "Structure and function of a neocortical synapse". *Nature* 591, pp. 111–116.

Holtmaat, A., Bonhoeffer, T., Chow, D. K., Chuckowree, J., De Paola, V., Hofer, S. B., Hübener, M., Keck, T., Knott, G., Lee, W.-C. A., Mostany, R., Mrsic-Flogel, T. D., Nedivi, E., Portera-Cailliau, C., Svoboda, K., Trachtenberg, J. T. & Wilbrecht, L. (2009). "Long-term, high-resolution

imaging in the mouse neocortex through a chronic cranial window". *Nature Protocols* 4, pp. 1128–1144.

Holtmaat, A. & Svoboda, K. (2009). "Experience-dependent structural synaptic plasticity in the mammalian brain". *Nature Reviews Neuroscience* 10, pp. 647–658.

Holtmaat, A., Wilbrecht, L., Knott, G. W., Welker, E. & Svoboda, K. (2006). "Experience-dependent and cell-type-specific spine growth in the neocortex". *Nature* 441, pp. 979–983.

Holtmaat, A. J. G. D., Trachtenberg, J. T., Wilbrecht, L., Shepherd, G. M., Zhang, X., Knott, G. W. & Svoboda, K. (2005). "Transient and persistent dendritic spines in the neocortex in vivo". *Neuron* 45, pp. 279–291.

Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities". *Proceedings of the National Academy of Sciences* 79, pp. 2554–2558.

Hopfield, J. J. (1984). "Neurons with graded response have collective computational properties like those of two-state neurons". *Proceedings of the National Academy of Sciences* 81, pp. 3088–3092.

Hopfield, J. J., Feinstein, D. I. & Palmer, R. G. (1983). "'Unlearning' has a stabilizing effect in collective memories". *Nature* 304, pp. 158–159.

Huttenlocher, P. R. (1979). "Synaptic density in human frontal cortex — developmental changes and effects of aging". *Brain Research* 163, pp. 195–205.

Huttenlocher, P. R., de Courten, C., Garey, L. J. & Van der Loos, H. (1982). "Synaptogenesis in human visual cortex — evidence for synapse elimination during normal development". *Neuroscience Letters* 33, pp. 247–252.

Ishii, K., Nagaoka, A., Kishida, Y., Okazaki, H., Yagishita, S., Ucar, H., Takahashi, N., Saito, N. & Kasai, H. (2018). "In vivo volume dynamics of dendritic spines in the neocortex of wild-type and Fmr1 KO mice". *eNeuro* 5, e0282–18.2018.

Jacot, A., Gabriel, F. & Hongler, C. (2018). "Neural tangent kernel: convergence and generalization in neural networks". *Advances in Neural Information Processing Systems 31*.

Ji, D. & Wilson, M. A. (2007). "Coordinated memory replay in the visual cortex and hippocampus during sleep". *Nature Neuroscience* 10, pp. 100–107.

Ji, Z., Srebro, N. & Telgarsky, M. (2021). "Fast margin maximization via dual acceleration". *Proceedings of the 38th International Conference on Machine Learning*, pp. 4860–4869.

Josselyn, S. A., Köhler, S. & Frankland, P. W. (2017). "Heroes of the engram". *Journal of Neuroscience* 37, pp. 4647–4657.

Kalisman, N., Silberberg, G. & Markram, H. (2005). "The neocortical microcircuit as a tabula rasa". *Proceedings of the National Academy of Sciences* 102, pp. 880–885.

Kanerva, P. (1988). *Sparse distributed memory*. MIT Press.

Kanter, I. & Eisenstein, E. (1990). "On the capacity per synapse". *Journal of Physics A: Mathematical and General* 23, pp. L935–L938.

Kappel, D., Habenschuss, S., Legenstein, R. & Maass, W. (2015). "Network plasticity as Bayesian inference". *PLOS Computational Biology* 11, e1004485.

Kasai, H., Ziv, N. E., Okazaki, H., Yagishita, S. & Toyoizumi, T. (2021). "Spine dynamics in the brain, mental disorders and artificial neural networks". *Nature Reviews Neuroscience* 22, pp. 407–422.

Kaufman, M., Corner, M. A. & Ziv, N. E. (2012). "Long-term relationships between cholinergic tone, synchronous bursting and synaptic remodeling". *PLOS One* 7, e40980.

Keeler, J. D. (1988). "Comparison between Kanerva's SDM and Hopfield-type neural networks". *Cognitive Science* 12, pp. 299–329.

Kepler, T. B. & Abbott, L. F. (1988). "Domains of attraction in neural networks". *Journal de Physique* 49, pp. 1657–1662.

Khona, M. & Fiete, I. R. (2022). "Attractor and integrator networks in the brain". *Nature Reviews Neuroscience* 23, pp. 744–766.

Klinzing, J. G., Niethard, N. & Born, J. (2019). "Mechanisms of systems memory consolidation during sleep". *Nature Neuroscience* 22, pp. 1598–1610.

Knoblauch, A., Körner, E., Körner, U. & Sommer, F. T. (2014). "Structural synaptic plasticity has high memory capacity and can explain graded amnesia, catastrophic forgetting, and the spacing effect". *PLOS One* 9, e96485.

Knott, G. W., Holtmaat, A., Wilbrecht, L., Welker, E. & Svoboda, K. (2006). "Spine growth precedes synapse formation in the adult neocortex in vivo". *Nature Neuroscience* 9, pp. 1117–1124.

Köhler, H. M. & Widmaier, D. (1991). "Sign-constrained linear learning and diluting in neural networks". *Journal of Physics A: Mathematical and General* 24, pp. L495–L502.

Kohonen, T. (1972). "Correlation matrix memories". *IEEE Transactions on Computers* C-21, pp. 353–359.

Koiran, P. (1994). "Dynamics of discrete time, continuous state Hopfield networks". *Neural Computation* 6, pp. 459–468.

Konorski, J. (1948). *Conditioned reflexes and neuron organization*. Cambridge University Press.

Krauth, W. & Mezard, M. (1987). "Learning algorithms with optimal stability in neural networks". *Journal of Physics A: Mathematical and General* 20, pp. L745–L752.

Krauth, W. & Mézard, M. (1989). "Storage capacity of memory networks with binary couplings". *Journal de Physique* 50, pp. 3057–3066.

Krauth, W., Mézard, M. & Nadal, J.-P. (1988). "Basins of attraction in a perceptron-like neural network". *Complex Systems* 2, pp. 387–408.

Krieg, D. & Triesch, J. (2014). "A unifying theory of synaptic long-term plasticity based on a sparse distribution of synaptic strength". *Frontiers in Synaptic Neuroscience* 6.

Krotov, D. & Hopfield, J. J. (2016). "Dense associative memory for pattern recognition". *Advances in Neural Information Processing Systems 29*, pp. 1172–1180.

Krotov, D. & Hopfield, J. J. (2020). "Large associative memory problem in neurobiology and machine learning". *Proceedings of the 9th International Conference on Learning Representations*.

Larkum, M. (2013). "A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex". *Trends in Neurosciences* 36, pp. 141–151.

Le Bé, J.-V. & Markram, H. (2006). "Spontaneous and evoked synaptic rewiring in the neonatal neocortex". *Proceedings of the National Academy of Sciences* 103, pp. 13214–13219.

Lee, Y. C., Doolen, G., Chen, H. H., Sun, G. Z., Maxwell, T., Lee, H. Y. & Giles, C. L. (1986). "Machine learning using a higher order correlation network". *Physica D: Nonlinear Phenomena* 22, pp. 276–306.

Lee, Y. & Kim, W. C. (2014). *Concise formulas for the surface area of the intersection of two hyperspherical caps*. Technical Report. Department of Industrial and Systems Engineering, KAIST.

Lefort, S., Tomm, C., Floyd Sarria, J.-C. & Petersen, C. C. H. (2009). "The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex". *Neuron* 61, pp. 301–316.

Lendvai, B., Stern, E. A., Chen, B. & Svoboda, K. (2000). "Experience-dependent plasticity of dendritic spines in the developing rat barrel cortex in vivo". *Nature* 404, pp. 876–881.

Levy, W. B. (2004). "Contrasting rules for synaptogenesis, modification of existing synapses, and synaptic removal as a function of neuronal computation". *Neurocomputing* 58-60, pp. 343–350.

Li, H. L. & van Rossum, M. C. W. (2020). "Energy efficient synaptic plasticity". *eLife* 9, e50804.

Li, S. (2011). "Concise formulas for the area and volume of a hyperspherical cap". *Asian Journal of Mathematics and Statistics* 4, pp. 66–70.

Li, W., Ma, L., Yang, G. & Gan, W.-B. (2017). "REM sleep selectively prunes and maintains new synapses in development and learning". *Nature Neuroscience* 20, pp. 427–437.

Lisman, J. (2017). "Glutamatergic synapses are structurally and biochemically complex because of multiple plasticity processes: long-term potentiation, long-term depression, short-term potentiation and scaling". *Philosophical Transactions of the Royal Society B: Biological Sciences* 372, p. 20160260.

Little, W. A. (1974). "The existence of persistent states in the brain". *Mathematical Biosciences* 19, pp. 101–120.

Loebel, A., Bé, J.-V. L., Richardson, M. J. E., Markram, H. & Herz, A. V. M. (2013). "Matched pre- and post-synaptic changes underlie synaptic plasticity over long time scales". *Journal of Neuroscience* 33, pp. 6257–6266.

Loewenstein, Y., Kuras, A. & Rumpel, S. (2011). "Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo". *Journal of Neuroscience* 31, pp. 9481–9488.

Loewenstein, Y., Yanover, U. & Rumpel, S. (2015). "Predicting the dynamics of network connectivity in the neocortex". *Journal of Neuroscience* 35, pp. 12535–12544.

Louie, K. & Wilson, M. A. (2001). "Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep". *Neuron* 29, pp. 145–156.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Majewska, A. K., Newton, J. R. & Sur, M. (2006). "Remodeling of synaptic structure in sensory cortical areas in vivo". *Journal of Neuroscience* 26, pp. 3021–3029.

Major, G., Larkum, M. E. & Schiller, J. (2013). "Active properties of neocortical pyramidal neuron dendrites". *Annual Review of Neuroscience* 36, pp. 1–24.

Maletic-Savatic, M., Malinow, R. & Svoboda, K. (1999). "Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity". *Science* 283, pp. 1923–1927.

Mangasarian, O. L. (1999). "Arbitrary-norm separating plane". *Operations Research Letters* 24, pp. 15–23.

Markram, H., Gerstner, W. & Sjöström, P. J. (2011). "A history of spike-timing-dependent plasticity". *Frontiers in Synaptic Neuroscience* 3.

McEliece, R., Posner, E., Rodemich, E. & Venkatesh, S. (1987). "The capacity of the Hopfield associative memory". *IEEE Transactions on Information Theory* 33, pp. 461–482.

Mel, B. W. (1991). "The clusteron: toward a simple abstraction for a complex neuron". *Advances in Neural Information Processing Systems 4*, pp. 35–42.

Micchelli, C. A. (1986). "Interpolation of scattered data: distance matrices and conditionally positive definite functions". *Constructive Approximation* 2, pp. 11–22.

Mickes, L., Wais, P. E. & Wixted, J. T. (2009). "Recollection is a continuous process: implications for dual-process theories of recognition memory". *Psychological Science* 20, pp. 509–515.

Millidge, B., Salvatori, T., Song, Y., Lukasiewicz, T. & Bogacz, R. (2022). "Universal Hopfield networks: a general framework for single-shot associative memory models". *Proceedings of the 39th International Conference on Machine Learning*, pp. 15561–15583.

Minerbi, A., Kahana, R., Goldfeld, L., Kaufman, M., Marom, S. & Ziv, N. E. (2009). "Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity". *PLOS Biology* 7, e1000136.

Miyamoto, D., Marshall, W., Tononi, G. & Cirelli, C. (2021). "Net decrease in spine-surface GluA1-containing AMPA receptors after post-learning sleep in the adult mouse cortex". *Nature Communications* 12, p. 2881.

Mongillo, G., Rumpel, S. & Loewenstein, Y. (2017). "Intrinsic volatility of synaptic connections — a challenge to the synaptic trace theory of memory". *Current Opinion in Neurobiology* 46, pp. 7–13.

Morrison, A., Aertsen, A. & Diesmann, M. (2007). "Spike-timing-dependent plasticity in balanced random networks". *Neural Computation* 19, pp. 1437–1467.

Morrison, A., Diesmann, M. & Gerstner, W. (2008). "Phenomenological models of synaptic plasticity based on spike timing". *Biological Cybernetics* 98, pp. 459–478.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N. & Soudry, D. (2019). "Convergence of gradient descent on separable data". *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428.

Nadal, J.-P. (1990). "On the storage capacity with sign-constrained synaptic couplings". *Network: Computation in Neural Systems* 1, pp. 463–466.

Nägerl, U. V., Eberhorn, N., Cambridge, S. B. & Bonhoeffer, T. (2004). "Bidirectional activity-dependent morphological plasticity in hippocampal neurons". *Neuron* 44, pp. 759–767.

Nägerl, U. V., Köstinger, G., Anderson, J. C., Martin, K. A. C. & Bonhoeffer, T. (2007). "Protracted synaptogenesis after activity-dependent spinogenesis in hippocampal neurons". *Journal of Neuroscience* 27, pp. 8149–8156.

Nakano, K. (1972). "Associatron — a model of associative memory". *IEEE Transactions on Systems, Man, and Cybernetics* SMC-2, pp. 380–388.

Neal, R. M. (1996). "Priors for infinite networks". *Bayesian Learning for Neural Networks*. Ed. by Neal, R. M. Vol. 118. Lecture Notes in Statistics. Springer, pp. 29–53.

Nishiyama, J. & Yasuda, R. (2015). "Biochemical computation for spine structural plasticity". *Neuron* 87, pp. 63–75.

Nowicki, D. & Siegelmann, H. (2010). "Flexible kernel memory". *PLOS One* 5, e10955.

O'Connor, D. H., Wittenberg, G. M. & Wang, S. S.-H. (2005). "Dissection of bidirectional synaptic plasticity into saturable unidirectional processes". *Journal of Neurophysiology* 94, pp. 1565–1573.

O'Kusky, J. & Colonnier, M. (1982). "Postnatal changes in the number of neurons and synapses in the visual cortex (area 17) of the macaque monkey: a stereological analysis in normal and monocularly deprived animals". *Journal of Comparative Neurology* 210, pp. 291–306.

O'Kusky, J. R. (1985). "Synapse elimination in the developing visual cortex: a morphometric analysis in normal and dark-reared cats". *Developmental Brain Research* 22, pp. 81–91.

Oh, W. C., Hill, T. C. & Zito, K. (2013). "Synapse-specific and size-dependent mechanisms of spine structural plasticity accompanying synaptic weakening". *Proceedings of the National Academy of Sciences* 110, E305–E312.

Parnavelas, J. G., Globus, A. & Kaups, P. (1973). "Continuous illumination from birth affects spine density of neurons in the visual cortex of the rat". *Experimental Neurology* 40, pp. 742–747.

Peretto, P. & Niez, J. J. (1986). "Long term memory storage capacity of multiconnected neural networks". *Biological Cybernetics* 54, pp. 53–63.

Perin, R., Berger, T. K. & Markram, H. (2011). "A synaptic organizing principle for cortical neuronal groups". *Proceedings of the National Academy of Sciences* 108, pp. 5419–5424.

Personnaz, L., Guyon, I. & Dreyfus, G. (1986). "Collective computational properties of neural networks: new learning mechanisms". *Physical Review A* 34, pp. 4217–4228.

Petanjek, Z., Judaš, M., Šimić, G., Rašin, M. R., Uylings, H. B. M., Rakic, P. & Kostović, I. (2011). "Extraordinary neoteny of synaptic spines in the human prefrontal cortex". *Proceedings of the National Academy of Sciences* 108, pp. 13281–13286.

Plato (1990). *Theaetetus*. Ed. by Burnyeat, M. Trans. by M. J. Levett. Hackett Publishing Company.

Poggio, T. & Girosi, F. (1990a). "Networks for approximation and learning". *Proceedings of the IEEE* 78, pp. 1481–1497.

Poggio, T. & Girosi, F. (1990b). "Regularization algorithms for learning that are equivalent to multilayer networks". *Science* 247, pp. 978–982.

Poirazi, P., Brannon, T. & Mel, B. W. (2003). "Pyramidal neuron as two-layer neural network". *Neuron* 37, pp. 989–999.

Poirazi, P. & Mel, B. W. (2001). "Impact of active dendrites and structural plasticity on the memory capacity of neural tissue". *Neuron* 29, pp. 779–796.

Polsky, A., Mel, B. W. & Schiller, J. (2004). "Computational subunits in thin dendrites of pyramidal cells". *Nature Neuroscience* 7, pp. 621–627.

Qiao, Q., Ma, L., Li, W., Tsai, J.-W., Yang, G. & Gan, W.-B. (2016). "Long-term stability of axonal boutons in the mouse barrel cortex". *Developmental Neurobiology* 76, pp. 252–261.

Radhakrishnan, A., Belkin, M. & Uhler, C. (2020). "Overparameterized neural networks implement associative memory". *Proceedings of the National Academy of Sciences* 117, pp. 27162–27170.

Rakic, P., Bourgeois, J.-P., Eckenhoff, M. F., Zecevic, N. & Goldman-Rakic, P. S. (1986). "Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex". *Science* 232, pp. 232–235.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., Klambauer, G., Brandstetter, J. & Hochreiter, S. (2021). "Hopfield networks is all you need". *Proceedings of the 9th International Conference on Learning Representations*.

Redondo, R. L. & Morris, R. G. M. (2011). "Making memories last: the synaptic tagging and capture hypothesis". *Nature Reviews Neuroscience* 12, pp. 17–30.

Renart, A., Song, P. & Wang, X.-J. (2003). "Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks". *Neuron* 38, pp. 473–485.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., Berker, A. d., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D.,

Zenke, F., Zylberberg, J., Therien, D. & Kording, K. P. (2019). "A deep learning framework for neuroscience". *Nature Neuroscience* 22, pp. 1761–1770.

Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Spartan Books.

Rosenzweig, M. R., Krech, D., Bennett, E. L. & Diamond, M. C. (1962). "Effects of environmental complexity and training on brain chemistry and anatomy: a replication and extension". *Journal of Comparative and Physiological Psychology* 55, pp. 429–437.

Rosset, S., Zhu, J. & Hastie, T. (2003). "Margin maximizing loss functions". *Advances in Neural Information Processing Systems 16*, pp. 1237–1244.

Roy, D. S., Park, Y.-G., Kim, M. E., Zhang, Y., Ogawa, S. K., DiNapoli, N., Gu, X., Cho, J. H., Choi, H., Kamentsky, L., Martin, J., Mosto, O., Aida, T., Chung, K. & Tonegawa, S. (2022). "Brain-wide mapping reveals that engrams for a single memory are distributed across multiple brain regions". *Nature Communications* 13, p. 1799.

Rubin, R., Abbott, L. F. & Sompolinsky, H. (2017). "Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity". *Proceedings of the National Academy of Sciences* 114, E9366–E9375.

Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. Vol. 1. MIT Press.

Sacramento, J., Wichert, A. & van Rossum, M. C. W. (2015). "Energy efficient sparse connectivity from imbalanced synaptic plasticity rules". *PLOS Computational Biology* 11, e1004265.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C. & Norman, K. A. (2018). "Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance". *Nature Communications* 9, p. 3920.

Scholl, C., Rule, M. E. & Hennig, M. H. (2021). "The information theory of developmental pruning: Optimizing global network architectures using local synaptic rules". *PLOS Computational Biology* 17, e1009458.

Schreiner, T., Petzka, M., Staudigl, T. & Staresina, B. P. (2021). "Endogenous memory reactivation during sleep in humans is clocked by slow oscillation-spindle complexes". *Nature Communications* 12, p. 3112.

Schwarz, J., Jayakumar, S. M., Pascanu, R., Latham, P. E. & Teh, Y. W. (2021). "Powerpropagation: a sparsity inducing weight reparameterisation". *Advances in Neural Information Processing Systems 34*.

Shouval, H. Z. (2005). "Clusters of interacting receptors can stabilize synaptic efficacies". *Proceedings of the National Academy of Sciences* 102, pp. 14440–14445.

Shouval, H. Z., Bear, M. F. & Cooper, L. N. (2002). "A unified model of NMDA receptor-dependent bidirectional synaptic plasticity". *Proceedings of the National Academy of Sciences* 99, pp. 10831–10836.

Sohal, V. S. & Hasselmo, M. E. (2000). "A model for experience-dependent changes in the responses of inferotemporal neurons". *Network: Computation in Neural Systems* 11, pp. 169–190.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S. & Chklovskii, D. B. (2005). "Highly nonrandom features of synaptic connectivity in local cortical circuits". *PLOS Biology* 3, e68.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. & Srebro, N. (2018). "The implicit bias of gradient descent on separable data". *The Journal of Machine Learning Research* 19, pp. 2822–2878.

Squire, L. R., Genzel, L., Wixted, J. T. & Morris, R. G. (2015). "Memory consolidation". *Cold Spring Harbor Perspectives in Biology* 7, a021766.

Squire, L. R., Stark, C. E. & Clark, R. E. (2004). "The medial temporal lobe". *Annual Review of Neuroscience* 27, pp. 279–306.

Statman, A., Kaufman, M., Minerbi, A., Ziv, N. E. & Brenner, N. (2014). "Synaptic size dynamics as an effectively stochastic process". *PLOS Computational Biology* 10, e1003846.

Steffens, H., Mott, A. C., Li, S., Wegner, W., Švehla, P., Kan, V. W. Y., Wolf, F., Liebscher, S. & Willig, K. I. (2021). "Stable but not rigid: chronic in vivo STED nanoscopy reveals extensive remodeling of spines, indicating multiple drivers of plasticity". *Science Advances* 7, eabf2806.

Sukhbaatar, S., Szlam, A., Weston, J. & Fergus, R. (2015). "End-to-end memory networks". *Advances in Neural Information Processing Systems 28*.

Suykens, J. & Vandewalle, J. (1999). "Least squares support vector machine classifiers". *Neural Processing Letters* 9, pp. 293–300.

Teyler, T. J. & Rudy, J. W. (2007). "The hippocampal indexing theory and episodic memory: updating the index". *Hippocampus* 17, pp. 1158–1169.

Thomson, A. & Lamy, C. (2007). "Functional maps of neocortical local circuitry". *Frontiers in Neuroscience* 1.

Tonegawa, S., Liu, X., Ramirez, S. & Redondo, R. (2015). "Memory engram cells have come of age". *Neuron* 87, pp. 918–931.

Tony Cai, T. & Jiang, T. (2012). "Phase transition in limiting distributions of coherence of high-dimensional random matrices". *Journal of Multivariate Analysis* 107, pp. 24–39.

Toyoizumi, T., Kaneko, M., Stryker, M. P. & Miller, K. D. (2014). "Modeling the dynamic interaction of Hebbian and homeostatic plasticity". *Neuron* 84, pp. 497–510.

Trachtenberg, J. T., Chen, B. E., Knott, G. W., Feng, G., Sanes, J. R., Welker, E. & Svoboda, K. (2002). "Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex". *Nature* 420, pp. 788–794.

Triesch, J., Vo, A. D. & Hafner, A.-S. (2018). "Competition for synaptic building blocks shapes synaptic plasticity". *eLife* 7, e37836.

Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P. & Salakhutdinov, R. (2019). "Transformer dissection: an unified understanding for transformer's attention via the lens of kernel". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4344–4353.

Tsodyks, M. V. & Feigel'man, M. V. (1988). "The enhanced storage capacity in neural networks with low activity level". *Europhysics Letters* 6, pp. 101–105.

Turner, A. M. & Greenough, W. T. (1985). "Differential rearing effects on rat visual cortex synapses. I. Synaptic and neuronal density and synapses per neuron". *Brain Research* 329, pp. 195–203.

Turrigiano, G. G. (2008). "The self-tuning neuron: synaptic scaling of excitatory synapses". *Cell* 135, pp. 422–435.

Tyulmankov, D., Fang, C., Vadaparty, A. & Yang, G. R. (2021). "Biological learning in key-value memory networks". *Advances in Neural Information Processing Systems 34*.

Valverde, F. (1967). "Apical dendritic spines of the visual cortex and light deprivation in the mouse". *Experimental Brain Research* 3, pp. 337–352.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). "Attention is all you need". *Advances in Neural Information Processing Systems 30*.

Venkatesh, S. S. (1986). "Epsilon capacity of neural networks". *AIP Conference Proceedings* 151, pp. 440–445.

Viswanathan, R. R. (1993). "Sign-constrained synapses and biased patterns in neural networks". *Journal of Physics A: Mathematical and General* 26, pp. 6195–6203.

Wegner, W., Steffens, H., Gregor, C., Wolf, F. & Willig, K. I. (2022). "Environmental enrichment enhances patterning and remodeling of synaptic nanoarchitecture as revealed by STED nanoscopy". *eLife* 11, e73603.

Weston, J., Chopra, S. & Bordes, A. (2015). "Memory networks". *Proceedings of the 2015 International Conference on Learning Representations*.

Wiegert, J. S. & Oertner, T. G. (2013). "Long-term depression triggers the selective elimination of weakly integrated synapses". *Proceedings of the National Academy of Sciences* 110, E4510–E4519.

Williams, C. (1996). "Computing with infinite networks". *Advances in Neural Information Processing Systems 9*.

Winfield, D. A. (1981). "The postnatal development of synapses in the visual cortex of the cat and the effects of eyelid closure". *Brain Research* 206, pp. 166–171.

Wright, M. A. & Gonzalez, J. E. (2021). "Transformers are deep infinite-dimensional non-Mercer binary kernel machines". Preprint. arXiv: 2106.01506.

Wu, Y., Wayne, G., Graves, A. & Lillicrap, T. (2018). "The Kanerva machine: a generative distributed memory". *Proceedings of the 2018 International Conference on Learning Representations*.

Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., Jones, T. & Zuo, Y. (2009). "Rapid formation and selective stabilization of synapses for enduring motor memories". *Nature* 462, pp. 915–919.

Yang, G., Lai, C. S. W., Cichon, J., Ma, L., Li, W. & Gan, W.-B. (2014). "Sleep promotes branch-specific formation of dendritic spines after learning". *Science* 344, pp. 1173–1178.

Yang, G., Pan, F. & Gan, W.-B. (2009). "Stably maintained dendritic spines are associated with lifelong memories". *Nature* 462, pp. 920–924.

Yasumatsu, N., Matsuzaki, M., Miyazaki, T., Noguchi, J. & Kasai, H. (2008). "Principles of long-term dynamics of dendritic spines". *Journal of Neuroscience* 28, pp. 13592–13608.

Yau, H. W. (1992). *Phase space techniques in neural network models*. PhD thesis. University of Edinburgh.

Yuste, R. (2015). "The discovery of dendritic spines by Cajal". *Frontiers in Neuroanatomy* 9.

Zhang, D., Zhang, C. & Stepanyants, A. (2019). "Robust associative learning is sufficient to explain the structural and dynamical properties of local cortical circuits". *Journal of Neuroscience* 39, pp. 6888–6904.

Zheng, P., Dimitrakakis, C. & Triesch, J. (2013). "Network self-organization explains the statistics and dynamics of synaptic connection strengths in cortex". *PLOS Computational Biology* 9, e1002848.

Zhou, Y., Lai, C. S. W., Bai, Y., Li, W., Zhao, R., Yang, G., Frank, M. G. & Gan, W.-B. (2020). "REM sleep promotes experience-dependent dendritic spine elimination in the mouse cortex". *Nature Communications* 11, p. 4819.

Zieliński, K. (2006). "Jerzy Konorski on brain associations". *Acta Neurobiologiae Experimentalis* 66, pp. 75–90.

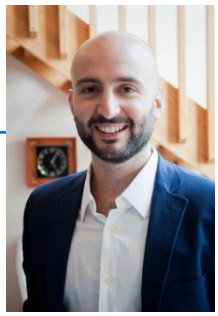
Ziv, N. E. & Brenner, N. (2018). "Synaptic tenacity or lack thereof: spontaneous remodeling of synapses". *Trends in Neurosciences* 41, pp. 89–99.

Zuo, Y., Lin, A., Chang, P. & Gan, W.-B. (2005a). "Development of long-term dendritic spine stability in diverse regions of cerebral cortex". *Neuron* 46, pp. 181–189.

Zuo, Y., Yang, G., Kwon, E. & Gan, W.-B. (2005b). "Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex". *Nature* 436, pp. 261–265.

Curriculum Vitae

GEORGIOS IATROPOULOS



Date of birth: 1990-12-31
Citizenship: Swedish
ORCID: 0000-0001-6978-311X

+41 78 745 3772
georgios.iatropoulos@gmail.com

EXPERIENCE

Doctoral researcher **Aug 2018 - Sep 2023**
Doctoral Program in Neuroscience
EPFL, Lausanne, Switzerland
Thesis advisors: Prof. Henry Markram, Prof. Wulfram Gerstner
My PhD project is a collaboration between the Blue Brain Project and the Laboratory of Computational Neuroscience. The aim is to develop a mathematical model of structural synaptic plasticity, describing how synapses are strengthened, weakened, and pruned during consolidation. The long-term goal is to simulate how neural circuits in neocortex rewire during long-term memory formation.

Research assistant **Jan 2016 - June 2018**
Department of Psychology (Division of Perception and Psychophysics)
Stockholm University, Stockholm, Sweden
The aim of this research project was to develop a model of olfactory memory based on psychophysical and neurobiological data. The project was carried out in collaboration with the Computational Brain Science Lab at the KTH Royal Institute of Technology.

Research intern **Nov 2013 - Feb 2014**
Computational Neuroscience and Neuroinformatics group
KTH Royal Institute of Technology, Stockholm, Sweden
The research conducted by the Computational Neuroscience and Neuroinformatics group is focused on modeling and simulating neural activity and functionality in brain regions such as cerebral cortex, basal ganglia and hippocampus. Over the past years, the group has been developing a model of prefrontal cortex and its ability to maintain working memory. The goals of my project was to implement this model in the simulation software NEST, validate it by reproducing experimental results, and investigate its scaling properties.

Research summer intern **July 2013 - Aug 2013**
Laboratory for Biomolecular Modeling (LBM)
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
The LBM research group at EPFL conducts computational research in molecular and structural biology, using primarily molecular dynamics simulations. The main goal of my project was to use such simulations to investigate the structural stability of a certain type of amyloid-beta fibril. Amyloid-beta is a protein that is believed to be one of the causes of Alzheimer's disease.

Industry summer intern **June 2012 - July 2012**
Scania, Södertälje, Sweden

EDUCATION

PhD Neuroscience **2018 - 2023**
EPFL, Lausanne, Switzerland

MSc Engineering Physics, Biomedical Physics track **2012 - 2015**
KTH Royal Institute of Technology, Stockholm, Sweden
Part of the Swedish civilingenjör program in Engineering Physics.
GPA: 5.0/5.0 (Scale: A=5.0, B=4.5, C=4.0, D=3.5, E=3.0, F=0)
Thesis: *The Effects of Scaling of Columnar Cell Population Sizes in a Spiking Attractor Network Model of Working Memory*

Erasmus exchange studies **Jan 2013 - June 2013**
EPFL, Lausanne, Switzerland

BSc Engineering Physics **2009 - 2012**
KTH Royal Institute of Technology, Stockholm, Sweden
Part of the Swedish civilingenjör program in Engineering Physics.
GPA: 4.6/5.0 (Scale: A=5.0, B=4.5, C=4.0, D=3.5, E=3.0, F=0)

Upper secondary school **2006 - 2009**
Södra Latin's Gymnasium, Stockholm, Sweden
Natural science program (**GPA: 20.0/20.0**)

Publications

Iatropoulos G, Brea J, Gerstner W. (2022) Kernel memory networks: A unifying framework for memory modeling. *Advances in Neural Information Processing Systems* 36.

Iatropoulos G, Herman P, Lansner A, Karlgren J, Larsson M, Olofsson JK. (2018) The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. *Cognition* 178, 37-49.

Iatropoulos G, Olofsson JK, Herman P, Lansner A, Larsson M. (2017) Analysis of statistics of semantic relations of odor-describing words written in olfactory versus non-olfactory contexts. *Chem. Senses* 42, E34–E35

Teaching

BIO465 Biological Modeling of Neural Networks Spring 2020
EPFL, Lausanne, Switzerland

BIO322 Introduction to Machine Learning for Bioengineers Fall 2019, 2020
EPFL, Lausanne, Switzerland

CS116 Projects in C++ for Life Science Students Fall 2018
EPFL, Lausanne, Switzerland

DD2432 Artificial Neural Networks and Other Learning Systems Spring 2015, 2016
KTH Royal Institute of Technology, Stockholm, Sweden

Scholarships

KTH General Student Scholarship (Håkansson Foundation) 2015
KTH General Student Scholarship (Rundqvist Foundation) 2013
The Henrik Göransson Sandviken Scholarship Foundation 2013
The Harriet and Sten Gustavsson Scholarship in Mathematics 2009

Attended Workshops, Conferences, and Summer Schools

Bernstein Conference on Computational Neuroscience Sep 2022
Berlin, Germany

Neuroscience School of Advanced Studies: Learning & Memory Aug 2022
Venice, Italy

EITN Workshop on Synaptic Plasticity Jan 2020
Paris, France

GRC Dendrites: Molecules, Structure, and Function Mar 2019
Ventura, CA, USA

26th Annual Meeting of the European Chemoreception Research Organization Sep 2016
Athens, Greece

4th Baltic-Nordic Summer School on Neuroinformatics June 2016
Nencki Institute of Experimental Biology, University of Warsaw, Warsaw, Poland

Workshop on Mechanisms of Reward and Associative Learning Mar 2015
Karolinska Institute, Stockholm, Sweden

Workshop on Progress in Brain-Like Computing Feb 2014
KTH Royal Institute of Technology, Stockholm, Sweden

Programming Skills

Scientific Programming	Python, Matlab, Julia, R
Other Programming	C++, Java, Bash, LaTeX
Deep Learning	PyTorch, Keras, Flux
Spiking Neural Networks	NEST, Brian simulator
Parallel Computing	MPI4Py

Language Skills (graded according to the ILR and CEFR scales)

Swedish	*Native proficiency* (ILR 5/CEFR C2)
English	*Full professional proficiency* (ILR 4-5/CEFR C2)
French	*Elementary proficiency* (ILR 1/CEFR A2)
Greek	*Native proficiency* (ILR 5/CEFR C2)

Other Activities

Corporate event organizer Sep 2011 - May 2012
Corporate event group, KTH Students' Union

Corporate event host Apr 2011
Physics chapter career fair (FArm)

Musician Jan 2010 - May 2010
Physics chapter theater orchestra